Supervised Factor Modeling for High-Dimensional Linear Time Series

Feiqing Huang[a], Kexin Lu[a], Yao Zheng[b] and Guodong Li[a*]

[a] *Department of Statistics and Actuarial Science, University of Hong Kong, China*

[b] *Department of Statistics, University of Connecticut, United States of America*

Abstract

Motivated by Tucker tensor decomposition, this paper imposes low-rank structures to the column and row spaces of coefficient matrices in a multivariate infinite-order vector autoregression (VAR), which leads to a supervised factor model with two factor modelings being conducted to responses and predictors simultaneously. Interestingly, the stationarity condition implies an intrinsic weak group sparsity mechanism of infinite-order VAR, and hence a rank-constrained group Lasso estimation is considered for high-dimensional linear time series. Its non-asymptotic properties are discussed by balancing the estimation, approximation and truncation errors. Moreover, an alternating gradient descent algorithm with hard-thresholding is designed to search for high-dimensional estimates, and its theoretical justifications, including statistical and convergence analysis, are also provided. Theoretical and computational properties of the proposed methodology are verified by simulation experiments, and the advantages over existing methods are demonstrated by analyzing US quarterly macroeconomic variables.

Keywords: Dimension reduction; High-dimensional time series; Infinite-order VAR; Tensor decomposition; Weak group sparsity.

*Correspondence author at: Department of Statistics and Actuarial Science, University of Hong Kong, Pokfulam Road, Hong Kong, China; Email addresses: amieehuang@connect.hku.hk (F. Huang), neithen@connect.hku.hk (K. Lu), yao.zheng@uconn.edu (Y. Zheng), gdli@hku.hk (G. Li)

1 Introduction

The high-speed advance in technology has spurred the rapid growth of high-dimensional data, especially time-dependent data, and examples can be found in many fields such as economics, finance, biology and neuroscience (Dowell and Pinson, 2016; Nicholson et al., 2020; Peña and Tsay, 2021). It is urgent to develop suitable models and methods for these larger and more complex time series data. Consider an N-dimensional time series $\{\mathbf{y}_t\}$. If it is a general linear process (GLP), then it can be written as

$$\mathbf{y}_t = \sum_{j=1}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{\varepsilon}_{t-j} + \boldsymbol{\varepsilon}_t, \tag{1.1}$$

where $\mathbf{y}_t \in \mathbb{R}^N$, $\boldsymbol{\varepsilon}_t \in \mathbb{R}^N$ is the white noise, and $\boldsymbol{\Psi}_j$'s are $N \times N$ coefficient matrices (Tsay, 2014). When N is large, a primary workhorse for the modeling and forecasting of GLP is the vector autoregressive (VAR) model (Basu et al., 2019; Basu and Michailidis, 2015; Zheng and Cheng, 2021; Wang et al., 2022b) due to its easy implementation. In fact, the GLP has a VAR(∞) representation,

$$\mathbf{y}_t = \sum_{j=1}^{\infty} \boldsymbol{A}_j \mathbf{y}_{t-j} + \boldsymbol{\varepsilon}_t, \tag{1.2}$$

where $\boldsymbol{A}_j$'s are $N \times N$ coefficient matrices. Thus, VAR models with a fixed AR order are not applicable to GLPs in general.

While empirical studies of multivariate time series data often use a small AR order, the potential increase in forecasting accuracy from a larger order cannot be realized without a viable high-dimensional estimation method for such processes. To fill this gap, this paper proposes an easy-to-implement VAR(∞)-based method for modeling and forecasting high-dimensional GLPs [1]. Our theoretical analysis encompasses GLPs in the form of (1.1), i.e. the VAR(∞) processes (1.2), while the proposed estimation method is based on a VAR(T_0) sieve approximation with a large running AR order T_0. When N is fixed, the asymptotic theory of VAR sieve approximations for VAR(∞) processes has been well established. It is shown that the truncation error can be adequately

[1] The codes are available on GitHub at: `https://github.com/neithen-Lu/Supervised-Factor-Modeling/tree/main`.

controlled by choosing a running order T_0 which grows with the sample size at an appropriate rate; see Lütkepohl (2005); Li et al. (2014). However, there has been no formal discussion in the literature on estimating high-dimensional GLPs via VAR sieve approximations. We refer to high dimensionality as settings where N may grow with the sample size. However, it is worth noting that the curse of dimensionality in the large N setup is compounded by the large running order T_0, which must also grow with the sample size to achieve a good approximation. Thus, this paper conducts simultaneous dimension reduction across both N variables and T_0 lags, while establishing the approximation theory of the VAR(T_0) sieve for the GLP in the high-dimensional setting.

Another fundamental approach to estimating VAR(∞) processes is through the vector autoregressive moving average (VARMA) model (Tsay, 2014). Empirical studies have shown that the VARMA model can provide more accurate forecasts than the VAR model with a fixed AR order when time series are longer (Chan et al., 2016; Wilms et al., 2023). Compared to the VARMA model, the VAR sieve approximation involves many more parameters. However, its advantages are at least threefold. First, compared with the VARMA model, the VAR(∞) model is more flexible in accommodating diverse temporal patterns. It allows for patterns where the non-zero lags are non-consecutive, including seasonal patterns; see Section 5.2 for simulation studies demonstrating the superior forecasting performance of the VAR(∞) model. Second, the VAR model is computationally easier to estimate than the VARMA model, enabling it to accommodate various dimension reduction schemes, including both sparse and low-rank methods. Third, while the VARMA model requires complicated identification constraints, the VAR sieve approach avoids this issue, as it seeks to approximate a VAR(∞) process which is identifiable. These reasons motivate us to tackle high-dimensional GLP modeling by combining VAR sieve with dimension reduction methods.

For high-dimensional VAR models, there are two major categories of dimension reduction methods. The first category is sparsity-inducing methods (e.g., Basu and Michailidis, 2015; Han et al., 2015; Guo et al., 2016; Zhu et al., 2017), where the sparse nonzero entries of the coefficient matrices can be interpreted as spillover effects from one variable to another. However, these methods have disadvantages when the connectivity among variables is dense and when the AR

order is large. In financial and economic time series, strong cross-sectional dependence is often observed among variables. In this case, rather than assuming that only some of them influence one another and that co-movement is sparse, it is more reasonable to assume that most variables are driven by a small number of common latent factors (Lam and Yao, 2012; Bai and Wang, 2016; Fan et al., 2022). Moreover, when the running AR order T_0 is large, many entries of the coefficient matrices $\boldsymbol{A}_j$'s may be very small but significant especially at high lags. As a result, it may be difficult for sparsity-inducing estimation to capture all of these weak signals on an entrywise basis; this issue also partially motivated the sparse interpretable VAR(∞) model with parametric lags in Zheng (2024).

On the other hand, the second category, which imposes various low-rank structures on finite-order VAR models (e.g., Velu and Reinsel, 2013; Carriero et al., 2016; Basu et al., 2019; Wang et al., 2022a,b; Billio et al., 2023; Samadi and Herath, 2024), is more suitable when strong cross-sectional dependence is prevalent among the variables. The most classical model in this category is the reduced-rank VAR model (Velu and Reinsel, 2013; Koop et al., 2019), where the rank of the coefficient matrices corresponds to the number of common factors in the time series. However, unlike static factor models (Lam and Yao, 2012; Bai and Wang, 2016), VAR models can be directly used for forecasting and provide interpretations regarding predictive relationships. Recently, Wang et al. (2022b) impose low-rank structures based on the Tucker decomposition of the stacked coefficient tensor of the finite-order VAR model. Our approach can be viewed as an adaptation of their approach to VAR(∞) processes. For VAR(∞) processes, we impose different low-rank structures on the column and row spaces of coefficient matrices $\boldsymbol{A}_j$'s. This can be interpreted as projecting responses and predictors into a small number of latent factors, termed *response and predictor factors*, respectively. The response factors are used to summarize all predictable components of the market, while the predictor factors contain all driving forces; see Section 2.1 for details. As demonstrated in the empirical example in Section 6, the corresponding factor loadings provide interpretations of group patterns among response and predictor variables. Thus, we refer to the proposed model as the *supervised factor model* to emphasize both its factor interpretations

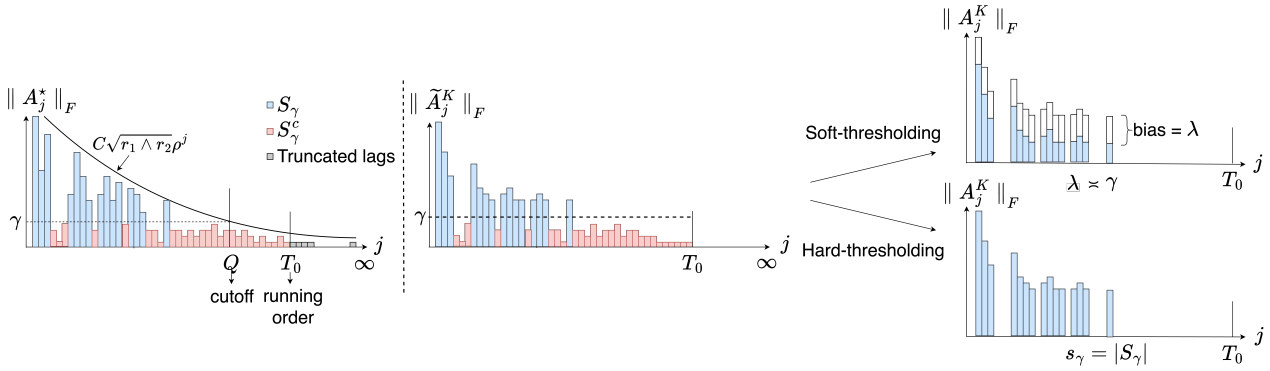

Figure 1: Illustration of exponentially decaying true AR coefficient matrices (left panel), estimated coefficient matrices truncated at running order T_0 before thresholding (middle panel), and estimated coefficient matrices after soft- or hard-thresholding (right panel). In the left panel, for a given threshold $\gamma > 0$, $S_\gamma = \left\{j \in \{1, \ldots, T_0\} \mid \|\boldsymbol{A}_j^*\|_{\mathrm{F}} > \gamma\right\}$ denotes the active set, and $S_\gamma^c = \{1, \ldots, T_0\} \backslash S_\gamma$. The middle and right panels correspond to Lines 6 and 7 of Algorithm 1, respectively. See Remarks 6 and 10 for details.

and its supervised, forecasting-oriented nature.

As discussed above, our low-rank assumption enables factor extraction for response and predictor variables. Additionally, it allows us to simultaneously address the curse of dimensionality arising from the large running order T_0 by leveraging the weak sparsity of lags in VAR(∞) processes. Our method is related to those of Nicholson et al. (2017), where several group-wise exact sparsity mechanisms are considered for the lag dimension of finite-order VAR models, such as the hierarchical group Lasso (HLag); see also Wilms et al. (2017). However, their methods are motivated by user-preferred special group-wise features across the lags. In contrast, our motivation stems from the interesting fact that the stationarity of the process intrinsically induces a constraint on the AR coefficient matrices, i.e. they lie within a generalized ℓ_1-ball, $\{\sum_{j=1}^{\infty} \|\boldsymbol{A}_j\|_{\mathrm{F}} \leqslant R\}$ with radius $0 < R < \infty$. This constraint exactly matches the weak group sparsity scenario with each coefficient matrix being a group of parameters (Raskutti et al., 2011; Wainwright, 2019); see Figure 1 for an illustration. The main contributions of this paper are summarized below:

- For modeling and forecasting high-dimensional GLPs, this paper introduces the supervised

factor model, a high-dimensional VAR sieve approximation method that imposes low-rank constraints on the column and row spaces of the coefficient matrices, along with weak group sparsity across the lags.

- To estimate the proposed model for high-dimensional GLPs, we first consider the rank-constrained group Lasso method. Our non-asymptotic analysis of the estimator shows that the truncation error can be adequately controlled, as long as the running order T_0 grows at least at a logarithmic rate relative to the effective sample size, while the estimation and approximation errors benefit from the weak group sparsity.

- To enforce the low-rank structures in our algorithmic implementation, we employ an alternating gradient descent approach (Chi et al., 2019), which enjoys low computational cost and storage complexity. While combining this approach with soft-thresholding for group Lasso regularization yields an algorithm for the proposed estimator, its performance tends to be sensitive to the choice of T_0. To enhance stability, we further develop an alternating gradient descent algorithm with hard-thresholding (Wainwright, 2019). We also provide the non-asymptotic convergence analysis of the statistical and optimization errors of this alternative algorithm.

In addition, we summarize the main differences between this paper and related literature as follows. Compared to existing high-dimensional VAR and VARMA models, our approach offers three main advantages. First, we allow for an infinite AR order, which provides a basis to explore potential improvements in forecast accuracy, whereas existing VAR models (e.g., Basu and Michailidis, 2015; Wang et al., 2022a) are typically constrained to small AR orders in a large-N setup due to the curse of dimensionality. Second, our method combines low-rank structures in the cross-sectional dimension with group sparsity across the lag dimension. Unlike methods reliant solely on sparsity (e.g., Zheng, 2024), our low-rank approach is better suited for cases where variables exhibit co-movement and hence the coefficient matrices may not be entrywise sparse. Third, the VAR(∞) framework circumvents the complicated identification problem of VARMA models

(e.g., Chan et al., 2016; Wilms et al., 2023), which also eases computation and interpretation. Methodologically, our low-rank approach is also related to the tensor decomposition method for coefficient tensors of the matrix AR model introduced in Chen et al. (2021) and the tensor AR model of Wang et al. (2023). However, we focus on vector-valued time series rather than matrix or tensor data, and the tensor decomposition in our context pertains only to the AR coefficient matrices across lags. Moreover, there is a growing interest in matrix and tensor factor models for high-dimensional time series (e.g., Chen and Fan, 2023; Wang et al., 2019; Chen et al., 2022). The aim of these models is to understand latent factor structures in matrix- and tensor-valued time series rather than forecasting. In contrast, our VAR(∞) model can be regarded as a supervised framework which extracts factor structures to achieve optimal forecasting performance. Lastly, it is worth noting that the use of infinite AR order can be integrated into other econometric methods such as the factor augmented regression (Stock and Watson, 2002a,b); this serves as one of the benchmarks in our numerical studies.

The remainder of the paper is organized as follows. Section 2.1 formally introduces the supervised factor model, and Section 2.2 develops high-dimensional estimation methods for the model. The corresponding alternating gradient descent algorithm with soft- or hard-thresholding is provided in Sections 3.1 and 3.2, respectively, with the latter being the primary focus in our simulation and empirical studies. Section 4 summarizes the theoretical results, with Section 4.1 focusing on the non-asymptotic properties of the rank-constrained group Lasso estimator, and Section 4.2 providing convergence analysis for the hard-thresholding-based algorithm. Section 5 conducts simulation experiments to evaluate the finite-sample estimation and prediction performance of the proposed methodology, and its usefulness is further demonstrated by a macroeconomic application in Section 6. Section 7 gives a short conclusion and discussion, and all technical proofs are relegated to a separate online supplementary file.

Throughout the paper, tensors are denoted by calligraphic capital letters; see, e.g. $\boldsymbol{\mathcal{A}}$, $\boldsymbol{\mathcal{B}}$, etc., and a brief introduction to tensor notations and Tucker decomposition is provided in the supplementary file. For two scalars a and b, we denote $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$.

For vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, denote by $\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \sum_j a_j b_j$ and $\|\boldsymbol{a}\|_2 = \sqrt{\langle \boldsymbol{a}, \boldsymbol{a} \rangle}$ the inner product and ℓ_2-norm, respectively. For a matrix $\boldsymbol{A} \in \mathbb{R}^{d_1 \times d_2}$, let $\boldsymbol{A}'$, $\mathrm{rank}(\boldsymbol{A})$, $\sigma_{\max}(\boldsymbol{A})$ (or $\sigma_{\min}(\boldsymbol{A})$), $\lambda_{\max}(\boldsymbol{A})$ (or $\lambda_{\min}(\boldsymbol{A})$), $\|\boldsymbol{A}\|_{\mathrm{op}} = \sigma_{\max}(\boldsymbol{A})$ and $\|\boldsymbol{A}\|_{\mathrm{F}} = \sqrt{\sum_{i,j} \boldsymbol{A}_{ij}^2}$ be its transpose, rank, largest (or smallest non-zero) singular value, largest (or smallest) eigenvalue, operator norm and Frobenius norm, respectively. Moreover, for any $d_1 \geqslant d_2$, the set of orthonormal matrices is denoted by $\mathcal{O}^{d_1 \times d_2} := \{\boldsymbol{A} \in \mathbb{R}^{d_1 \times d_2} \mid \boldsymbol{A}'\boldsymbol{A} = \boldsymbol{I}_{d_2}\}$, where $\boldsymbol{I}_{d_2}$ is a $d_2 \times d_2$ identity matrix. On the other hand, for any two sequences x_n and y_n, we 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 $x_n \gtrsim y_n$, $x_n = O(y_n)$ if $x_n \lesssim y_n$, and $x_n = o(y_n)$ if $\lim_{n\to\infty} x_n / y_n = 0$.

2 High-dimensional linear time series modeling

2.1 Supervised factor model for high-dimensional time series

Let $\mathbb{M}_1 = \mathrm{colspace}\{\boldsymbol{\Psi}_j, j \geqslant 1\}$ be the space spanned by the columns of all coefficient matrices $\boldsymbol{\Psi}_j$'s. Similarly we can define the row space as $\mathbb{M}_2 = \mathrm{rowspace}\{\boldsymbol{\Psi}_j, j \geqslant 1\}$, and their dimensions are denoted by $r_i = \dim(\mathbb{M}_i) \leqslant N$ with $i = 1$ and 2. In addition, the matrix polynomial for the GLP is defined as $\boldsymbol{\Psi}(z) = \boldsymbol{I} - \sum_{j=1}^{\infty} \boldsymbol{\Psi}_j z^j$, where $z \in \mathbb{C}$, and $\mathbb{C}$ is the complex space. Note that r_1 and r_2 are not equal in general. Moreover, the low-rank constraint is imposed on the column and row spaces of all $\boldsymbol{\Psi}_j$'s, and hence each matrix may have different ranks. In fact, it can be verified that $\mathrm{rank}(\boldsymbol{\Psi}_j) \leqslant \min(r_1, r_2)$ for all $j \geqslant 1$.

Assumption 1 (Invertibility condition)**.** *The determinant of $\boldsymbol{\Psi}(z)$ is not equal to zero for all $|z| < 1$, and $\sum_{j=1}^{\infty} \|\boldsymbol{\Psi}_j\|_{\mathrm{op}} < \infty$.*

Proposition 1. *If Assumption 1 holds, then the VAR(∞) form at* (1.2) *can be uniquely identified with $\sum_{j=1}^{\infty} \|\boldsymbol{A}_j\|_{\mathrm{op}} < \infty$. Moreover, $\mathbb{M}_1$ and $\mathbb{M}_2$ are also the column and row spaces of coefficient matrices $\boldsymbol{A}_j$'s, respectively.*

From the above proposition, the GLP has an equivalent VAR(∞) form, whose coefficient ma-

trices satisfy $\boldsymbol{A}_j = \boldsymbol{\Psi}_j - \sum_{k=1}^{j-1} \boldsymbol{\Psi}_{j-k}\boldsymbol{A}_k$ for all $j \geqslant 1$. It can be verified that the determinant of $\boldsymbol{A}(z)$ is not equal to zero for all $|z| < 1$, where $\boldsymbol{A}(z) = \boldsymbol{I} - \sum_{j=1}^{\infty} \boldsymbol{A}_j z^j$ is the VAR(∞) matrix polynomial. Moreover, coefficient matrices of the GLP and its VAR(∞) representation share the same column and row spaces.

Consider two matrices $\boldsymbol{U}_1 \in \mathcal{O}^{N\times r_1}$ and $\boldsymbol{U}_2 \in \mathcal{O}^{N\times r_2}$, which contain the bases of subspaces $\mathbb{M}_1$ and $\mathbb{M}_2$, respectively. Then, by some tensor algebra, there exist two sequences of $r_1 \times r_2$ matrices $\{\boldsymbol{H}_j, \boldsymbol{G}_j, j \geqslant 1\}$ such that $\boldsymbol{\Psi}_j = \boldsymbol{U}_1\boldsymbol{H}_j\boldsymbol{U}_2'$ and $\boldsymbol{A}_j = \boldsymbol{U}_1\boldsymbol{G}_j\boldsymbol{U}_2'$ for all $j \geqslant 1$. As a result, under Assumption 1 and the low-rank constraint, models (1.1) and (1.2) can be rewritten into

$$\mathbf{y}_t = \boldsymbol{U}_1 \sum_{j=1}^{\infty} \boldsymbol{H}_j\boldsymbol{U}_2'\boldsymbol{\varepsilon}_{t-j} + \boldsymbol{\varepsilon}_t \quad \text{and} \quad \mathbf{y}_t = \boldsymbol{U}_1 \sum_{j=1}^{\infty} \boldsymbol{G}_j\boldsymbol{U}_2'\mathbf{y}_{t-j} + \boldsymbol{\varepsilon}_t. \tag{2.1}$$

Note that $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ are not unique, while the projection matrices of $\mathbb{M}_1$ and $\mathbb{M}_2$ can be uniquely defined by $\boldsymbol{P}_1 = \boldsymbol{U}_1\boldsymbol{U}_1'$ and $\boldsymbol{P}_2 = \boldsymbol{U}_2\boldsymbol{U}_2'$, respectively. Moreover, for $i = 1$ and 2, let $\boldsymbol{U}_i^{\perp} \in \mathcal{O}^{N\times(N-r_i)}$ such that $(\boldsymbol{U}_i, \boldsymbol{U}_i^{\perp})$ is an $N \times N$ orthonormal matrix, and then $\boldsymbol{P}_i^{\perp} = \boldsymbol{U}_i^{\perp}\boldsymbol{U}_i^{\perp\prime}$ is the projection matrix of $\mathbb{M}_i^{\perp}$, i.e. orthogonal complement of subspace $\mathbb{M}_i$.

Note that model (2.1) involves two types of dimension reduction. We first consider the projection of $\mathbf{y}_t$ onto subspace $\mathbb{M}_1$ and its orthogonal complement, i.e. $\mathbf{y}_t = \boldsymbol{P}_1\mathbf{y}_t + \boldsymbol{P}_1^{\perp}\mathbf{y}_t$, and these two parts can be verified to have completely different dynamic structures,

$$\boldsymbol{P}_1\mathbf{y}_t = \boldsymbol{U}_1 \sum_{j=1}^{\infty} \boldsymbol{H}_j\boldsymbol{U}_2'\boldsymbol{\varepsilon}_{t-j} + \boldsymbol{P}_1\boldsymbol{\varepsilon}_t \quad \text{and} \quad \boldsymbol{P}_1^{\perp}\mathbf{y}_t = \boldsymbol{P}_1^{\perp}\boldsymbol{\varepsilon}_t,$$

where all information of $\mathbf{y}_t$ related to temporally dependent structures is contained in $\mathbb{M}_1$, whereas $\mathbb{M}_1^{\perp}$ includes only purely idiosyncratic and serially independent components. In fact, model (2.1) has a form of static factor models,

$$\mathbf{y}_t = \boldsymbol{U}_1\boldsymbol{f}_t + \boldsymbol{\varepsilon}_t \quad \text{with} \quad \boldsymbol{f}_t = \sum_{j=1}^{\infty} \boldsymbol{H}_j\boldsymbol{U}_2'\boldsymbol{\varepsilon}_{t-j} = \sum_{j=1}^{\infty} \boldsymbol{G}_j\boldsymbol{U}_2'\mathbf{y}_{t-j}, \tag{2.2}$$

where $\boldsymbol{f}_t \in \mathbb{R}^{r_1}$ contains r_1 latent factors, and $\boldsymbol{U}_1$ is the corresponding loading matrix; see Lam and Yao (2012); Bai and Wang (2016). Consequently, we call $\boldsymbol{f}_t = \boldsymbol{U}_1'\mathbf{y}_t - \boldsymbol{U}_1'\boldsymbol{\varepsilon}_t$ or $\boldsymbol{U}_1'\mathbf{y}_t$ the *response factor* since it summarizes all predictable components in the response, and accordingly $\mathbb{M}_1$ can be referred to as the *response factor space*.

On the other hand, for the dimension reduction on predictors, we project $\mathbf{y}_{t-j}$ onto $\mathbb{M}_2$ and $\mathbb{M}_2^{\perp}$, and it holds that $\mathbf{y}_{t-j} = \boldsymbol{P}_2\mathbf{y}_{t-j} + \boldsymbol{P}_2^{\perp}\mathbf{y}_{t-j}$. Moreover, for N-dimensional random vectors $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{z}_j$'s with $j \geqslant 1$, the partial covariance function is usually used to measure the relationship between $\mathbf{x}_1$ and $\mathbf{x}_2$ after removing the effects of $\mathbf{z}_j$'s, and it has the form of $\mathrm{pcov}(\mathbf{x}_1, \mathbf{x}_2|\mathbf{z}_1, \mathbf{z}_2, \ldots) = \mathrm{cov}(\mathbf{x}_1 - \hat{\mathbf{x}}_1, \mathbf{x}_2 - \hat{\mathbf{x}}_2)$, where, for $i = 1$ and 2, $\hat{\mathbf{x}}_i = \sum_{j=1}^{\infty} \hat{\boldsymbol{B}}_j^{(i)}\mathbf{z}_j$ and $(\hat{\boldsymbol{B}}_1^{(i)}, \hat{\boldsymbol{B}}_2^{(i)}, \ldots) = \arg\min E\|\mathbf{x}_i - \sum_{j=1}^{\infty} \boldsymbol{B}_j\mathbf{z}_j\|_2^2$; see Fan and Yao (2003); Tsay (2014). As a result, if $E\|\mathbf{y}_t\|_2^2 < \infty$, then

$$\mathrm{pcov}(\mathbf{y}_t, \mathbf{y}_{t-j}|\boldsymbol{P}_2\mathbf{y}_{t-1}, \boldsymbol{P}_2\mathbf{y}_{t-2}, \ldots) = \mathrm{pcov}(\mathbf{y}_t, \boldsymbol{P}_2^{\perp}\mathbf{y}_{t-j}|\boldsymbol{P}_2\mathbf{y}_{t-1}, \boldsymbol{P}_2\mathbf{y}_{t-2}, \ldots) = 0 \text{ for all } j \geqslant 1,$$

i.e. the space $\mathbb{M}_2$ can summarize all information of $\mathbf{y}_{t-j}$ that contributes to predicting $\mathbf{y}_t$, or $\boldsymbol{U}_2'\mathbf{y}_{t-j}$ contains all driving forces of the market. Thus, we call $\boldsymbol{U}_2'\mathbf{y}_{t-j}$ the *predictor factor* for simplicity, and $\mathbb{M}_2$ is referred to as the *predictor factor space*. Since model (2.1) is a supervised problem in nature, and we call it the *supervised factor model* to emphasize the above interpretations from unsupervised factor modeling perspectives (Lam and Yao, 2012; Bai and Wang, 2016).

Example 1. *In the macroeconomic application in Section 5, the estimated ranks for the large dataset are* $(\hat{r}_1, \hat{r}_2) = (1, 1)$. *Under this specification, the model at* (2.1) *has the fitted factor form of* $\widehat{\boldsymbol{u}}_1'\mathbf{y}_t = \sum_{j=1}^{\infty} \hat{g}_j\widehat{\boldsymbol{u}}_1'\mathbf{y}_{t-j}\widehat{\boldsymbol{u}}_2' + \widehat{\boldsymbol{u}}_1'\boldsymbol{\varepsilon}_t$, *where* $\widehat{\boldsymbol{u}}_1'\mathbf{y}_t$ *and* $\widehat{\boldsymbol{u}}_2'\mathbf{y}_{t-j}$ *can be viewed as two different macroeconomic indices. The response factor* $\widehat{\boldsymbol{u}}_1'\mathbf{y}_t$ *captures how the present economy responds to changes in the market condition, which essentially can be viewed as principal component analysis on the response variables. On the other hand, the predictor factor* $\widehat{\boldsymbol{u}}_2'\mathbf{y}_{t-j}$ *summarizes important past information that is predictive of the present market condition.*

Model (2.1) is partially motivated from tensor techniques, and the two types of dimension reduction can be imposed naturally from viewpoints of tensor decomposition; see the supplementary file for more details on tensor notations and decomposition. Specifically, for the VAR(∞) form at (1.2), we first rearrange the coefficient matrices into a tensor $\boldsymbol{\mathcal{A}}_{\infty} \in \mathbb{R}^{N\times N\times\infty}$ such that its mode-1 matricization is $(\boldsymbol{\mathcal{A}}_{\infty})_{(1)} = (\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots)$, and then its mode-2 matricization assumes the form of $(\boldsymbol{\mathcal{A}}_{\infty})_{(2)} = (\boldsymbol{A}_1', \boldsymbol{A}_2', \ldots)$. Note that the column spaces of $(\boldsymbol{\mathcal{A}}_{\infty})_{(1)}$ and $(\boldsymbol{\mathcal{A}}_{\infty})_{(2)}$ are $\mathbb{M}_1$ and $\mathbb{M}_2$, respectively, and $r_1 = \mathrm{rank}\{(\boldsymbol{\mathcal{A}}_{\infty})_{(1)}\}$ and $r_2 = \mathrm{rank}\{(\boldsymbol{\mathcal{A}}_{\infty})_{(2)}\}$ are the first two Tucker ranks.

Accordingly, we have the Tucker decomposition (Tucker, 1966; De Lathauwer et al., 2000) below,

$$\boldsymbol{\mathcal{A}}_\infty = \boldsymbol{\mathcal{G}}_\infty \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2, \tag{2.3}$$

where $\boldsymbol{\mathcal{G}}_\infty \in \mathbb{R}^{r_1 \times r_2 \times \infty}$, $\boldsymbol{U}_1 \in \mathbb{R}^{N \times r_1}$ and $\boldsymbol{U}_2 \in \mathbb{R}^{N \times r_2}$. In particular, we can choose $\boldsymbol{U}_1$, $\boldsymbol{U}_2$ and $\boldsymbol{G}_j$'s in (2.1) and let $\boldsymbol{G}_j$ be the j-th frontal slice of $\boldsymbol{\mathcal{G}}_\infty$ for $j \geqslant 1$, i.e. $(\boldsymbol{\mathcal{G}}_\infty)_{(1)} = (\boldsymbol{G}_1, \boldsymbol{G}_2, \ldots)$. As a result, the VAR($\infty$) form at (1.2), together with the low-Tucker-rank constraint at (2.3), is equivalent to the supervised factor model at (2.1).

Remark 1 (Varying ranks across lags). While $\text{rank}(\boldsymbol{\Psi}_j)$ for $j \geqslant 1$ are allowed to vary, we simplify the problem by focusing on their upper bound as determined by r_1 and r_2. The upper bound $\min(r_1, r_2)$ is sharper when $\boldsymbol{\Psi}_j$'s have similar ranks. However, it is possible for some $\text{rank}(\boldsymbol{\Psi}_j)$'s to be much larger than the others. In particular, since the first few lags are often considered more important, users may prefer to relax the low-rank constraints on them. In such cases, we may alternatively define r_1 and r_2 for $\boldsymbol{\Psi}_j$ with $j \geqslant q_0$, where q_0 is a pre-determined small order. Additionally, we may further impose entrywise sparsity on $\boldsymbol{\Psi}_j$ for $1 \leqslant j < q_0$ for dimension reduction. We leave such extensions as an interesting direction for future research.

Remark 2 (Connections with static factor models). Based on the static factor model form at (2.2), if the response factor space $\mathbb{M}_1$ is the only focus, then it suffices to conduct unsupervised factor modeling (Lam and Yao, 2012; Gao and Tsay, 2023). However, our approach to standardizing $\boldsymbol{U}_1$ differs from the usual standardization of the loading matrix in factor models. Specifically, $\boldsymbol{U}_1$ at (2.2) is standardized to be orthonormal, which can lead to the factors $\boldsymbol{f}_t$ having a diverging variance as $N \to \infty$. For example, consider a stationary VAR(1) model, $\mathbf{y}_t = \boldsymbol{A}\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t$ with $\boldsymbol{A} = 0.9 \cdot \mathbf{1}_N(1, 0, \ldots, 0)'$, where all elements of the N-dimensional vector $\mathbf{1}_N$ are one. Then it can be verified that $\text{var}(\boldsymbol{f}_t) = O(N)$.

Remark 3 (Connections with generalized dynamic factor models in Forni et al. (2000)). In the special case with $r_1 = N$, we can choose $\boldsymbol{U}_1 = \boldsymbol{I}_N$. Let $\boldsymbol{H}(L) = \sum_{j=1}^{\infty} \boldsymbol{H}_j L^j \in \mathbb{R}^{N \times r_2}$ with L being the lag operator, and $\boldsymbol{u}_t = \boldsymbol{U}_2' \boldsymbol{\varepsilon}_t \in \mathbb{R}^{r_2}$, which can be standardized to have an identity variance matrix. As a result, the VMA(∞) process at (2.1) can be rewritten into $\mathbf{y}_t = \boldsymbol{H}(L)\boldsymbol{u}_t + \boldsymbol{\varepsilon}_t$, i.e. it

admits a generalized dynamic factor model in Forni et al. (2000). Note that the proposed model is for a supervised problem, while the generalized dynamic factor model is fundamentally for an unsupervised one. Moreover, the idiosyncratic component $\boldsymbol{\varepsilon}_t$ is assumed to be independent of $\boldsymbol{u}_t$ at all leads and lags in Forni et al. (2000, 2005), while this cannot be satisfied for model (2.1).

Remark 4 (Connections with dynamic factor models in Amengual and Watson (2007)). Consider a special case of model (2.1) with $\boldsymbol{U}_1 = \boldsymbol{U}_2$, i.e. the response and predictor factors are identical, and then the VAR(∞) representation can be rewritten into

$$\boldsymbol{f}_t = \sum_{j=1}^{\infty} \boldsymbol{G}_j \boldsymbol{f}_{t-j} + \boldsymbol{U}_1' \boldsymbol{\varepsilon}_t \quad \text{with} \quad \boldsymbol{f}_t = \boldsymbol{U}_1' \mathbf{y}_t,$$

where $\{\boldsymbol{U}_1' \boldsymbol{\varepsilon}_t\}$ is the new white noise sequence. As discussed in Wang et al. (2022b), its finite-order case corresponds to the dynamic factor model in Amengual and Watson (2007) with no measurement error. This relationship leads to an alternative estimation method. When there are infinite number of lags, we can first perform factor modeling on $\{\mathbf{y}_t\}$ and then a low-dimensional VAR(∞) model to the summarized factors $\boldsymbol{f}_t$ with group lasso penalization on the coefficient matrices to select the non-zero lags.

Remark 5 (Identification condition). The Tucker decomposition at (2.3) is not unique, since $\boldsymbol{\mathcal{A}}_\infty = \boldsymbol{\mathcal{G}}_\infty \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 = (\boldsymbol{\mathcal{G}}_\infty \times_1 \boldsymbol{O}_1 \times_2 \boldsymbol{O}_2) \times_1 (\boldsymbol{U}_1 \boldsymbol{O}_1^{-1}) \times_2 (\boldsymbol{U}_2 \boldsymbol{O}_2^{-1})$ for any invertible matrices $\boldsymbol{O}_i \in \mathbb{R}^{r_i \times r_i}$ with $i = 1$ and 2. However, it is worth noting that the theoretical properties in this paper are directly established for $\boldsymbol{\mathcal{A}}_\infty$, rather than $\boldsymbol{\mathcal{G}}_\infty$ and $\boldsymbol{U}_i$'s. Thus, the identification issue can be avoided. If estimating the components $\boldsymbol{\mathcal{G}}_\infty$ and $\boldsymbol{U}_i$'s is of interest, then similar to Wang et al. (2022b), the following assumptions can be made to guarantee the identifiability of these components. Specifically, we may consider the higher-order singular value decomposition (HOSVD) of $\boldsymbol{\mathcal{A}}_\infty$, a special Tucker decomposition defined by choosing $\boldsymbol{U}_i$ as the tall matrix consisting of the top r_i left singular vectors of $(\boldsymbol{\mathcal{A}}_\infty)_{(i)}$ and then setting $\boldsymbol{\mathcal{G}}_\infty = \boldsymbol{\mathcal{A}}_\infty \times_1 \boldsymbol{U}_1' \times_2 \boldsymbol{U}_2'$. In addition, the singular values of $\boldsymbol{\mathcal{A}}_{(1)}$ and $\boldsymbol{A}_{(2)}$ are assumed to be all distinct, and the first nonzero element in each column of $\boldsymbol{U}_i$ is assumed to be positive.

2.2 High-dimensional estimation

For an observed time series $\{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$ generated by model (1.2) with the low-rank constraint (2.3), its true coefficient matrices are denoted by $\boldsymbol{A}_j^*$'s, and this paper adopts the VAR sieve approximation method to estimate them, i.e.

$$\mathbf{y}_t = \sum_{j=1}^{T_0} \boldsymbol{A}_j^* \mathbf{y}_{t-j} + \widetilde{\boldsymbol{\varepsilon}}_t, \quad \widetilde{\boldsymbol{\varepsilon}}_t = \boldsymbol{\varepsilon}_t + \boldsymbol{r}_t \text{ and } \operatorname{rank}\{(\boldsymbol{\mathcal{A}}_\infty^*)_{(i)}\} \leqslant r_i \text{ with } i = 1 \text{ and } 2, \tag{2.4}$$

where T_0 is the running order of VAR models, $\boldsymbol{r}_t = \sum_{j=T_0+1}^{\infty} \boldsymbol{A}_j^* \mathbf{y}_{t-j}$ is the truncated term, and $\boldsymbol{\mathcal{A}}_\infty^*$ is the true full coefficient tensor.

The coefficient matrices of model (2.4) are $\boldsymbol{A}_j$'s with $1 \leqslant j \leqslant T_0$, and they can be rearranged into a coefficient tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{N \times N \times T_0}$ such that its mode-1 matricization is $\boldsymbol{\mathcal{A}}_{(1)} = (\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_{T_0}) \in \mathbb{R}^{N \times NT_0}$. Let $\mathbf{x}_t = (\mathbf{y}_{t-1}', \mathbf{y}_{t-2}', \ldots, \mathbf{y}_{t-T_0}')' \in \mathbb{R}^{NT_0}$, $\boldsymbol{X} = (\mathbf{x}_T, \mathbf{x}_{T-1}, \ldots, \mathbf{x}_{T_0+1}) \in \mathbb{R}^{NT_0 \times (T-T_0)}$ and $\boldsymbol{Y} = (\mathbf{y}_T, \mathbf{y}_{T-1}, \ldots, \mathbf{y}_{T_0+1}) \in \mathbb{R}^{N \times (T-T_0)}$. Consider the ordinary least squares estimation, and then the loss function has the form of

$$\mathcal{L}(\boldsymbol{\mathcal{A}}) = \frac{1}{2T_1} \sum_{t=T_0+1}^{T} \|\mathbf{y}_t - \sum_{j=1}^{T_0} \boldsymbol{A}_j \mathbf{y}_{t-j}\|_{\mathrm{F}}^2 = \frac{1}{2T_1} \|\boldsymbol{Y} - \boldsymbol{\mathcal{A}}_{(1)} \boldsymbol{X}\|_{\mathrm{F}}^2,$$

where the effective sample size is $T_1 = T - T_0$. Denote by $\boldsymbol{\mathcal{A}}^* \in \mathbb{R}^{N \times N \times T_0}$ the true coefficient tensor, and it is a truncated tensor obtained by removing all $\boldsymbol{A}_j^*$ with $j > T_0$ from the $N \times N \times \infty$ full coefficient tensor $\boldsymbol{\mathcal{A}}_\infty^*$. Define the parameter space

$$\boldsymbol{\Theta}(r_1, r_2) = \{\boldsymbol{\mathcal{A}} \in \mathbb{R}^{N \times N \times T_0} \mid \operatorname{rank}(\boldsymbol{\mathcal{A}}_{(1)}) \leqslant r_1, \operatorname{rank}(\boldsymbol{\mathcal{A}}_{(2)}) \leqslant r_2\},$$

and the low-Tucker-rank constraint at (2.4) implies that $\boldsymbol{\mathcal{A}}^* \in \boldsymbol{\Theta}(r_1, r_2)$.

To obtain theoretical justifications, the truncated term $\boldsymbol{r}_t$ is required to approach zero quickly, which requests $\boldsymbol{A}_j^*$ or $\boldsymbol{\Psi}_j^*$ to decay at a sufficient rate as $j \to \infty$, where $\boldsymbol{\Psi}_j^*$'s are true coefficient matrices of the corresponding GLP. While Assumption 1 is sufficient to achieve asymptotic properties for the low-dimensional (or multivariate) time series, we need a stronger exponential decay below to establish non-asymptotic properties for the high-dimensional case.

Assumption 2 (Exponential decay)**.** *There exists some* $\rho \in (0, 1)$ *such that* $\|\boldsymbol{\Psi}_j^*\|_{\mathrm{op}} = O(\rho^j)$ *and* $\|\boldsymbol{A}_j^*\|_{\mathrm{op}} = O(\rho^j)$ *as* $j \to \infty$.

Assumption 2 is mild since all VAR and VARMA processes are still included. In addition, as discussed in Remark 6, the exponential decay implies that the true coefficients of model (2.4) will be within a generalized ℓ_1-ball, $\{\boldsymbol{\mathcal{A}} : \sum_{j=1}^{T_0} \|\boldsymbol{A}_j\|_{\mathrm{F}} \leqslant R\}$. This feature exactly matches the scenario of weak sparsity (Raskutti et al., 2011; Wainwright, 2019), and hence an automated lag selection procedure can be inspired. Specifically, if we regard each $\boldsymbol{A}_j^*$ as a group of parameters, since any $\boldsymbol{A}_j^*$ with a relatively large j must be close to (if not exactly) a zero matrix, then the target coefficient tensor $\boldsymbol{\mathcal{A}}^*$ must be weakly group-sparse. As a result, this paper considers the following rank-constrained group Lasso estimator of $\boldsymbol{\mathcal{A}}^*$,

$$\widehat{\boldsymbol{\mathcal{A}}} = \underset{\boldsymbol{\mathcal{A}} \in \boldsymbol{\Theta}(r_1, r_2)}{\arg\min}\ \mathcal{L}(\boldsymbol{\mathcal{A}}) + \lambda \|\boldsymbol{\mathcal{A}}\|_{\ddagger} \quad \text{with} \quad \|\boldsymbol{\mathcal{A}}\|_{\ddagger} = \sum_{j=1}^{T_0} \|\boldsymbol{A}_j\|_{\mathrm{F}}, \tag{2.5}$$

where $\widehat{\boldsymbol{\mathcal{A}}}_{(1)} = (\widehat{\boldsymbol{A}}_1, \widehat{\boldsymbol{A}}_2, \ldots, \widehat{\boldsymbol{A}}_{T_0})$, and $\lambda > 0$ is a tuning parameter of penalization. The true full coefficient tensor $\boldsymbol{\mathcal{A}}_\infty^*$ can be estimated by $\widehat{\boldsymbol{\mathcal{A}}}_\infty \in \mathbb{R}^{N \times N \times \infty}$, which appends infinitely many zero matrices to $\widehat{\boldsymbol{\mathcal{A}}} \in \mathbb{R}^{N \times N \times T_0}$ such that $(\widehat{\boldsymbol{\mathcal{A}}}_\infty)_{(1)} = (\widehat{\boldsymbol{\mathcal{A}}}_{(1)}, \boldsymbol{0}_{N \times N}, \boldsymbol{0}_{N \times N}, \ldots)$, i.e. $\widehat{\boldsymbol{A}}_j = \boldsymbol{0}$ for $j > T_0$. Note that $\widehat{\boldsymbol{\mathcal{A}}}_\infty$ satisfies the low-Tucker-rank constraint at (2.4).

The optimization problem at (2.5) will produce a group-sparse estimate, with all coefficient matrices close or equal to zero being suppressed. The corresponding algorithm is discussed in Section 3.1. However, the estimated nonzero coefficient matrices are biased. In addition, as shown in our simulation studies in Section S3.5 of the supplementary file, this algorithm is sensitive to changes in T and T_0. Motivated by these issues, in Section 3.2, we develop an algorithm for implementing the alternative sparsity-constrained estimation,

$$\widehat{\boldsymbol{\mathcal{A}}}_{\mathbf{SC}} = \underset{\boldsymbol{\mathcal{A}} \in \boldsymbol{\Theta}(r_1, r_2), \|\boldsymbol{\mathcal{A}}\|_0 \leqslant s}{\arg\min}\ \mathcal{L}(\boldsymbol{\mathcal{A}}) \quad \text{with} \quad \|\boldsymbol{\mathcal{A}}\|_0 = \sum_{j=1}^{T_0} I(\|\boldsymbol{A}_j\|_{\mathrm{F}} > 0), \tag{2.6}$$

where $\|\boldsymbol{\mathcal{A}}\|_0$ is the number of active matrices, and the sparsity level $s > 0$ is a tuning parameter.

Remark 6 (Connection between exponential decay and weak group sparsity of $\boldsymbol{A}_j^*$'s). By Assumption 2 and the low-rank condition at (2.3), we have $\|\boldsymbol{A}_j^*\|_{\mathrm{F}} \leqslant C\sqrt{r_1 \wedge r_2}\rho^j$ for $j \geqslant 1$, where C is an absolute constant. As a result, a smaller ρ corresponds to a faster decay rate of $\boldsymbol{A}_j^*$ and hence a greater level of group sparsity of $\boldsymbol{\mathcal{A}}^*$; i.e. there is a smaller cutoff Q such that all $\boldsymbol{A}_j^*$'s with $j \geqslant Q$ are very close to zero matrices. This is illustrated in the left panel of Figure 1.

3 Algorithms for high-dimensional estimation

3.1 Alternating gradient descent algorithm

The rank-constrained group Lasso estimation at (2.5) involves both the low-rank constraint and sparsity. Specifically, the loss function $\mathcal{L}(\boldsymbol{\mathcal{A}})$ is nonconvex on the low-rank space $\boldsymbol{\Theta}(r_1, r_2)$, while the Lasso penalty $\|\boldsymbol{\mathcal{A}}\|_{\ddagger}$ relies on the convexity for optimization since it is not differentiable. This makes the parameter estimation difficult. Following Agarwal et al. (2012), we consider an alternating gradient descent algorithm to search for estimates, and it is based on the second-order approximation of $\mathcal{L}(\boldsymbol{\mathcal{A}})$; see Remark 7 for more details. Moreover, the thresholding operator in this subsection can approximate the Lasso penalty at (2.5). As a result, the resulting optimizer is close to, but different from, the rank-constrained group Lasso estimator $\widehat{\boldsymbol{\mathcal{A}}}$ at (2.5).

We first deal with the low-rank constraint. Note that, for any $\boldsymbol{\mathcal{A}} \in \boldsymbol{\Theta}(r_1, r_2)$, there exists a Tucker decomposition $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{G}} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2$ with $\boldsymbol{\mathcal{G}} \in \mathbb{R}^{r_1 \times r_2 \times T_0}$, $\boldsymbol{U}_1 \in \mathbb{R}^{N \times r_1}$ and $\boldsymbol{U}_2 \in \mathbb{R}^{N \times r_2}$. As a result, the loss function in Section 2.2 can be rewritten as $\mathcal{L}(\boldsymbol{\mathcal{G}}, \boldsymbol{U}_1, \boldsymbol{U}_2) := \mathcal{L}(\boldsymbol{\mathcal{G}} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2)$, and we further adjust it by adding two regularization terms,

$$\mathcal{L}^{\mathrm{GD}}(\boldsymbol{\mathcal{G}}, \boldsymbol{U}_1, \boldsymbol{U}_2) = \mathcal{L}(\boldsymbol{\mathcal{G}}, \boldsymbol{U}_1, \boldsymbol{U}_2) + \frac{a}{2}(\|\boldsymbol{U}_1'\boldsymbol{U}_1 - b^2\boldsymbol{I}_{r_1}\|_{\mathrm{F}}^2 + \|\boldsymbol{U}_2'\boldsymbol{U}_2 - b^2\boldsymbol{I}_{r_2}\|_{\mathrm{F}}^2),$$

where $a, b > 0$ are two tuning parameters. The above method is motivated by Han et al. (2022) for low-rank tensor estimation, and the regularization terms, $\|\boldsymbol{U}_1'\boldsymbol{U}_1 - b^2\boldsymbol{I}_{r_1}\|_{\mathrm{F}}^2$ and $\|\boldsymbol{U}_2'\boldsymbol{U}_2 - b^2\boldsymbol{I}_{r_2}\|_{\mathrm{F}}^2$, are used to keep $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ from being singular, and meanwhile they can also balance the scaling of $\boldsymbol{\mathcal{G}}$, $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$.

Denote by $\widetilde{\boldsymbol{\mathcal{G}}}$, $\widetilde{\boldsymbol{U}}_1$ and $\widetilde{\boldsymbol{U}}_2$ the minimizers of $\mathcal{L}^{\mathrm{GD}}(\boldsymbol{\mathcal{G}}, \boldsymbol{U}_1, \boldsymbol{U}_2)$, and it then holds that $\widetilde{\boldsymbol{U}}_i'\widetilde{\boldsymbol{U}}_i = b^2\boldsymbol{I}_{r_i}$, i.e. $b^{-1}\widetilde{\boldsymbol{U}}_i$ are orthonormal, for $i = 1$ and 2. In fact, if this is not true, then there exist invertible matrices $\boldsymbol{O}_i \in \mathbb{R}^{r_i \times r_i}$ such that $\widetilde{\boldsymbol{U}}_i = \bar{\boldsymbol{U}}_i\boldsymbol{O}_i$ and $\bar{\boldsymbol{U}}_i'\bar{\boldsymbol{U}}_i = b^2\boldsymbol{I}_{r_i}$, where $1 \leqslant i \leqslant 2$. Let $\bar{\boldsymbol{\mathcal{A}}} = (\widetilde{\boldsymbol{\mathcal{G}}} \times_1 \boldsymbol{O}_1 \times_2 \boldsymbol{O}_2) \times_1 \bar{\boldsymbol{U}}_1 \times_2 \bar{\boldsymbol{U}}_2$. Then $\mathcal{L}(\bar{\boldsymbol{\mathcal{A}}}) = \mathcal{L}(\widetilde{\boldsymbol{\mathcal{A}}})$ while the regularization terms for $\bar{\boldsymbol{U}}_1$ and $\bar{\boldsymbol{U}}_2$ are reduced to zero. This leads to a contradiction with the definition of minimizers. Similar regularization methods have been widely applied to non-convex low-rank matrix estimation problems; see Tu et al. (2016), Wang et al. (2017) and references therein.

Let us define the gradient of $\mathcal{L}^{\mathrm{GD}}$, and its partial derivatives can be calculated as

$$\nabla_{\mathcal{G}}\mathcal{L}^{\mathrm{GD}} = \nabla_{\mathcal{G}}\mathcal{L}(\boldsymbol{\mathcal{A}}) \quad \text{and} \quad \nabla_{U_i}\mathcal{L}^{\mathrm{GD}} = \nabla_{U_i}\mathcal{L}(\boldsymbol{\mathcal{A}}) + a\boldsymbol{U}_i(\boldsymbol{U}_i'\boldsymbol{U}_i - b^2\boldsymbol{I}_{r_i}) \quad \text{with } i = 1 \text{ and } 2,$$

where $\nabla\mathcal{L}(\boldsymbol{\mathcal{A}})$, $\nabla_{U_1}\mathcal{L}(\boldsymbol{\mathcal{A}})$, $\nabla_{U_2}\mathcal{L}(\boldsymbol{\mathcal{A}})$ and $\nabla_{\mathcal{G}}\mathcal{L}(\boldsymbol{\mathcal{A}})$ are the first order partial derivatives of $\mathcal{L}(\boldsymbol{\mathcal{A}})$ with respect to $\boldsymbol{\mathcal{A}}, \boldsymbol{U}_1, \boldsymbol{U}_2$ and $\mathcal{G}$, respectively. We next define the soft-thresholding operation. Consider a coefficient tensor $\boldsymbol{\mathcal{B}} \in \mathbb{R}^{d_1 \times d_2 \times T_0}$ with $\boldsymbol{\mathcal{B}}_{(1)} = (\boldsymbol{B}_1, \boldsymbol{B}_2, \ldots, \boldsymbol{B}_{T_0})$, where for each $1 \leqslant j \leqslant T_0$ the coefficient matrix $\boldsymbol{B}_j \in \mathbb{R}^{d_1 \times d_2}$ is the j-th frontal slice. For a tuning parameter $\lambda > 0$, define the soft-thresholding operator as $\widetilde{\boldsymbol{\mathcal{B}}} = \mathrm{ST}(\boldsymbol{\mathcal{B}}, \lambda) \in \mathbb{R}^{d_1 \times d_2 \times T_0}$ with $\widetilde{\boldsymbol{B}}_j = (1 - \lambda/\|\boldsymbol{B}_j\|_{\mathrm{F}})_+\boldsymbol{B}_j$ for $1 \leqslant j \leqslant T_0$ and $\widetilde{\boldsymbol{\mathcal{B}}}_{(1)} = (\widetilde{\boldsymbol{B}}_1, \widetilde{\boldsymbol{B}}_2, \ldots, \widetilde{\boldsymbol{B}}_{T_0})$, where $\boldsymbol{B}_j$ and $\widetilde{\boldsymbol{B}}_j$ are the j-th frontal slices of $\boldsymbol{\mathcal{B}}$ and $\widetilde{\boldsymbol{\mathcal{B}}}$, respectively, and the function $(x)_+ = x$ for $x > 0$ and zero otherwise. The soft-thresholding operator $\mathrm{ST}(\boldsymbol{\mathcal{A}}, \lambda)$ can project a non-group-sparse coefficient tensor $\boldsymbol{\mathcal{A}}$ into a group-sparse one.

To solve (2.5), we apply the alternating gradient descent algorithm as outlined in Algorithm 1, except that the sparsity parameter s at Line 1 and $\mathrm{HT}(\widetilde{\boldsymbol{\mathcal{A}}}^{k+1}, \mathrm{s})$ at Line 7 are replaced by the tuning parameter λ and soft-thresholding operator $\mathrm{ST}(\widetilde{\boldsymbol{\mathcal{A}}}^{k+1}, \lambda)$, respectively. At the k-th iteration, the three components, $\boldsymbol{U}_1^{k+1}, \boldsymbol{U}_2^{k+1}$ and $\widetilde{\mathcal{G}}^{k+1}$, are first updated by gradient descent separately. Then the resulting estimator $\widetilde{\boldsymbol{\mathcal{A}}}^{k+1} = \widetilde{\mathcal{G}}^{k+1} \times_1 \boldsymbol{U}_1^{k+1} \times_2 \boldsymbol{U}_2^{k+1}$ at Line 6 is converted to a group-sparse coefficient tensor via the soft-thresholding $\mathrm{ST}(\widetilde{\boldsymbol{\mathcal{A}}}^{k+1}, \lambda)$. The above two steps are repeated K times, after which we can obtain the estimate $\boldsymbol{\mathcal{A}}^K$.

Remark 7. Following the method in Agarwal et al. (2012), Algorithm 1 exploits the second-order approximation of $\mathcal{L}(\boldsymbol{\mathcal{A}})$ to derive the alternating gradient descent updates. Specifically, at the k-th iteration, the gradient descent update for $\boldsymbol{U}_i$ in Algorithm 1 results from $\boldsymbol{U}_i^{k+1} = \arg\min_{\boldsymbol{U}_i} \mathcal{L}^{\mathrm{GD}}(\mathcal{G}^k, \boldsymbol{U}_1^k, \boldsymbol{U}_2^k) + \langle \nabla_{U_i}\mathcal{L}^{\mathrm{GD}}, \boldsymbol{U}_i - \boldsymbol{U}_i^k \rangle + (2\eta)^{-1}\|\boldsymbol{U}_i - \boldsymbol{U}_i^k\|_{\mathrm{F}}^2$, for $i = 1, 2$. The subsequent update for $\widetilde{\mathcal{G}}$ is derived similarly.

3.2 An algorithm with hard-thresholding

This subsection considers Algorithm 1 with hard-thresholding at Line 7, which solves the optimization problem at (2.6). We prefer hard-thresholding over soft-thresholding for three main reasons. First, the VAR sieve approximation method depends on the running order T_0. As it increases, the

model complexity will increase, while the number of effective samples $T_1 = T - T_0$ will decrease. As a result, the numerically selected tuning parameter λ for the soft-thresholding may vary dramatically under different T_0, while the hard-thresholding is less sensitive to such type of changes. Indeed, our simulation studies in Section 5.3 demonstrate that the hard-thresholding algorithm is not sensitive to the value of T_0 when T_0 is sufficiently large. Second, the soft-thresholding method always leads to a biased estimator although it is sparse. In fact, the soft-thresholding method or Lasso problem is typically preferred for ensuring the convexity of the loss function, while the loss function at (2.5) is nonconvex due to the low-rankness. It is hence not necessary to insist on using the Lasso method to induce the group sparsity of $\boldsymbol{A}_j$'s. Third, it is easier to establish the convergence analysis of hard-thresholding methods (Tropp and Wright, 2010; Shen and Li, 2017).

We define the hard-thresholding operation as follows. Consider a coefficient tensor $\boldsymbol{\mathcal{B}} \in \mathbb{R}^{d_1 \times d_2 \times T_0}$ with $\boldsymbol{\mathcal{B}}_{(1)} = (\boldsymbol{B}_1, \boldsymbol{B}_2, \ldots, \boldsymbol{B}_{T_0})$, where for each $1 \leqslant j \leqslant T_0$ the coefficient matrix $\boldsymbol{B}_j \in \mathbb{R}^{d_1 \times d_2}$ is the j-th frontal slice. Denote by $\mathrm{HT}(\boldsymbol{\mathcal{B}}, \mathrm{s}) \in \mathbb{R}^{d_1 \times d_2 \times T_0}$ the hard-thresholding operator, which keeps the s largest coefficient matrices in terms of $\|\boldsymbol{B}_j\|_{\mathrm{F}}$'s and suppresses the rest to zero. Define a parameter space with both low-rankness and sparsity below,

$$\boldsymbol{\Theta}^{\mathrm{SP}}(r_1, r_2, s) = \{\boldsymbol{\mathcal{A}} \in \mathbb{R}^{N \times N \times T_0} \mid \|\boldsymbol{\mathcal{A}}\|_0 \leqslant s, \ \mathrm{rank}(\boldsymbol{\mathcal{A}}_{(1)}) \leqslant r_1, \ \mathrm{rank}(\boldsymbol{\mathcal{A}}_{(2)}) \leqslant r_2\}.$$

Note that, for any $1 \leqslant s \leqslant T_0$, $\boldsymbol{\Theta}^{\mathrm{SP}}(r_1, r_2, s) \subseteq \boldsymbol{\Theta}(r_1, r_2)$, and the hard-thresholding operator $\mathrm{HT}(\boldsymbol{\mathcal{A}}, \mathrm{s})$ is a projection from parameter spaces $\boldsymbol{\Theta}(r_1, r_2)$ into $\boldsymbol{\Theta}^{\mathrm{SP}}(r_1, r_2, s)$. For a tensor $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{G}} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2$, it can be verified that $\mathrm{HT}(\boldsymbol{\mathcal{A}}, \mathrm{s}) = \widetilde{\boldsymbol{\mathcal{G}}} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2$, and $\widetilde{\boldsymbol{\mathcal{G}}} = \mathrm{HT}(\boldsymbol{\mathcal{G}}, \mathrm{s})$ if $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ are assumed to have orthonormal columns.

To solve (2.6), unlike the method for (2.5), at Line 7 of Algorithm 1, we project $\widetilde{\boldsymbol{\mathcal{A}}}^{k+1}$ into the group-sparse parameter space $\boldsymbol{\Theta}^{\mathrm{SP}}(r_1, r_2, s)$ by $\mathrm{HT}(\widetilde{\boldsymbol{\mathcal{A}}}^{\mathrm{k+1}}, \mathrm{s}) = \boldsymbol{\mathcal{G}}^{k+1} \times_1 \boldsymbol{U}_1^{k+1} \times_2 \boldsymbol{U}_2^{k+1}$, denoted by $\boldsymbol{\mathcal{A}}^{k+1}$. In Algorithm 1, we may alternatively conduct the hard-thresholding on $\widetilde{\boldsymbol{\mathcal{G}}}^{k+1}$ directly, and a similar performance can be observed. However, it is more convenient to establish the corresponding convergence analysis for the hard-thresholding on $\widetilde{\boldsymbol{\mathcal{A}}}^{k+1}$.

Next we discuss the initialization of the algorithm, which applies to either soft- or hard-thresholding. For the running order T_0, theoretically speaking, it affects the truncation error

Algorithm 1: Alternating gradient descent algorithm with hard-thresholding

Input: Running order T_0, ranks (r_1, r_2), sparsity parameter s, initialization $\mathcal{G}^0, \boldsymbol{U}_1^0, \boldsymbol{U}_2^0$, regularization parameters $a, b > 0$ and step size $\eta > 0$.

For $k = 0, 1, 2, \ldots, K-1$:

$\quad \boldsymbol{U}_1^{k+1} \leftarrow \boldsymbol{U}_1^k - \eta \left[\nabla_{U_1}\mathcal{L}(\boldsymbol{\mathcal{A}}^k) + a\boldsymbol{U}_1^k(\boldsymbol{U}_1^{k\prime}\boldsymbol{U}_1^k - b^2\boldsymbol{I}_{r_1})\right]$

$\quad \boldsymbol{U}_2^{k+1} \leftarrow \boldsymbol{U}_2^k - \eta \left[\nabla_{U_2}\mathcal{L}(\boldsymbol{\mathcal{A}}^k) + a\boldsymbol{U}_2^k(\boldsymbol{U}_2^{k\prime}\boldsymbol{U}_2^k - b^2\boldsymbol{I}_{r_2})\right]$

$\quad \widetilde{\mathcal{G}}^{k+1} \leftarrow \mathcal{G}^k - \eta\nabla_{\mathcal{G}}\mathcal{L}(\boldsymbol{\mathcal{A}}^k)$

$\quad \widetilde{\boldsymbol{\mathcal{A}}}^{k+1} = \widetilde{\mathcal{G}}^{k+1} \times_1 \boldsymbol{U}_1^{k+1} \times_2 \boldsymbol{U}_2^{k+1}$

$\quad \boldsymbol{\mathcal{A}}^{k+1} = \mathcal{G}^{k+1} \times_1 \boldsymbol{U}_1^{k+1} \times_2 \boldsymbol{U}_2^{k+1} \leftarrow \mathrm{HT}(\widetilde{\boldsymbol{\mathcal{A}}}^{\mathrm{k+1}}, \mathrm{s})$

end for

return $\boldsymbol{\mathcal{A}}^K = \mathcal{G}^K \times_1 \boldsymbol{U}_1^K \times_2 \boldsymbol{U}_2^K$

only, while the parameter estimation will not change too much as long as it is sufficient large. However, in practice, a larger T_0 will lead to a smaller effective sample size. As a result, we recommend setting $T_0 = \lfloor\sqrt{T}\rfloor$ as for selecting the maximum lag for sample autocorrelation functions in the literature (Tsay, 2014), where $\lfloor\cdot\rfloor$ is the floor function, and further refinement can be easily made by gradually adjusting the choice of T_0 based on the initial result with $T_0 = \lfloor\sqrt{T}\rfloor$. For example, if the estimated coefficient matrices for lags beyond $T_0 - k$ are all zero, then they may reduce T_0 by k, or to be more conservative, by a slightly smaller amount. For the algorithm with soft-thresholding, the sparsity is induced via tuning parameter λ. A larger T_0 will involve more inactive coefficient matrices, and a larger amount of λ is hence needed to suppress these coefficients. This makes the algorithm less stable. On the contrary, for the algorithm with hard-thresholding, the sparsity level is specified directly, and hence the resulting algorithm is more stable.

To select the low ranks r_1 and r_2 and sparsity level s, since the true model is most likely an VAR(∞) process which lies outside the class of finite-order VAR models, we recommend the Akaike

information criterion (AIC),

$$\mathrm{AIC}(r_1, r_2, s) = \log\left\{(2T_1)^{-1}\|\boldsymbol{Y} - \widetilde{\boldsymbol{\mathcal{A}}}_{(1)}\boldsymbol{X}\|_{\mathrm{F}}^2\right\} + 2[(r_1 + r_2)N + \log T_0]s/T_1;$$

see Remark 8 below for more discussions. Lastly, the $r_1 \times r_2 \times T_0$ tensor $\boldsymbol{\mathcal{G}}^0$ can be set to zero, while $\boldsymbol{U}_1^0$ and $\boldsymbol{U}_2^0$ are initialized by some orthonormal matrices of sizes $N \times r_1$ and $N \times r_2$, respectively. Moreover, to provide a warm-start initialization in practice, we may skip the hard-thresholding operation for the first few iterations.

Remark 8. We prefer AIC over BIC for model selection since we assume that the true data generating process is a VAR(∞) process, and consequently, the true model almost always lie outside of the candidate model set (Shibata, 1980; Goldenshluger and Zeevi, 2001; Bühlmann, 1997; Ing and Wei, 2005). In particular, Shibata (1980) showed that when the true order is infinity or relatively large compared to the sample size, AIC is efficient in the sense that its prediction performance is asymptotically equivalent to the best offered by the candidate models, while BIC is not. For the rank selection, we may alternatively use the ridge-type ratio estimator in Wang et al. (2022b). However, this method may suffer from information loss due to the use of a fixed AR order. Moreover, it can be much less efficient since it is unable to leverage the group sparsity.

Remark 9. To guarantee the convergence of Algorithm 1 theoretically, there are requirements on the choice of regularization parameters, a and b, in $\mathcal{L}^{\mathrm{GD}}(\boldsymbol{\mathcal{G}}, \boldsymbol{U}_1, \boldsymbol{U}_2)$; see Theorem 4 in the next section for details. In practice, both soft- and hard-thresholding algorithms are insensitive to a and b, and we set both to one in all our simulation and empirical studies.

4 Theoretical properties

4.1 Non-asymptotic properties for group Lasso estimation

This subsection derives non-asymptotic properties of the proposed rank-constrained group Lasso estimator, $\widehat{\boldsymbol{\mathcal{A}}}_\infty$, at (2.5) in Section 2. Its accuracy is measured in terms of both parameter estimation and prediction as in the literature (Wainwright, 2019),

$$
\begin{aligned}
e_{\mathrm{est}}(\widehat{\boldsymbol{\mathcal{A}}}_\infty) &= \|\boldsymbol{\mathcal{A}}^*_\infty - \widehat{\boldsymbol{\mathcal{A}}}_\infty\|_{\mathrm{F}}^2 = \sum_{j=1}^{\infty} \|\boldsymbol{A}_j^* - \widehat{\boldsymbol{A}}_j\|_{\mathrm{F}}^2 = \|\boldsymbol{\mathcal{A}}^* - \widehat{\boldsymbol{\mathcal{A}}}\|_{\mathrm{F}}^2 + e_{\mathrm{trunc}}, \\
e_{\mathrm{pred}}(\widehat{\boldsymbol{\mathcal{A}}}_\infty) &= \frac{1}{T_1} \sum_{t=T_0+1}^{T} \Big\| \sum_{j=1}^{\infty} (\boldsymbol{A}_j^* - \widehat{\boldsymbol{A}}_j)\mathbf{y}_{t-j} \Big\|_2^2 = T_1^{-1} \|(\boldsymbol{\mathcal{A}}^* - \widehat{\boldsymbol{\mathcal{A}}})_{(1)} \boldsymbol{X}\|_{\mathrm{F}} + \widetilde{e}_{\mathrm{trunc}},
\end{aligned}
\tag{4.1}
$$

where the truncation errors coming from the sieve approximation are given by

$$
e_{\mathrm{trunc}} = \sum_{j=T_0+1}^{\infty} \|\boldsymbol{A}_j^*\|_{\mathrm{F}}^2 \quad \text{and} \quad \widetilde{e}_{\mathrm{trunc}} = \frac{1}{T_1} \sum_{t=T_0+1}^{T} \left\{ 2 \Big\langle \sum_{j=1}^{T_0} (\boldsymbol{A}_j^* - \widehat{\boldsymbol{A}}_j)\mathbf{y}_{t-j}, \boldsymbol{r}_t \Big\rangle + \|\boldsymbol{r}_t\|_2^2 \right\},
$$

respectively, $\boldsymbol{\mathcal{A}}^*_\infty$ is the true full coefficient tensor, and $\boldsymbol{r}_t$ is defined in (2.4). We first handle the VAR sieve estimator $\widehat{\boldsymbol{\mathcal{A}}}$, which corresponds to the first terms on the right hand side of both equations at (4.1), and the key intermediate step is to prove a more general result, which is known as the oracle inequality in the literature (Negahban et al., 2012).

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 $\mathrm{var}(\boldsymbol{\xi}_t) = \boldsymbol{I}_N$. In addition, the coordinates $(\boldsymbol{\xi}_{it})_{1\leqslant i\leqslant N}$ within $\boldsymbol{\xi}_t$ are mutually independent and σ^2-sub-Gaussian, where $\sigma^2 > 0$ is an absolute constant.*

Assumption 4 (Running orders)**.** *There exists an absolute constant $C > 0$ such that $T_1\rho^{T_0/2} \leqslant C$, or equivalently, $T_0 \geqslant 2\{\log(1/\rho)\}^{-1}\log(T_1/C)$.*

Assumption 3 is commonly used in the literature of high-dimensional time series (Zheng and Cheng, 2021; Wang et al., 2023), and the homoskedasticity here can even be further relaxed to unconditional heteroskedasticity at the cost of more complex notation. However, it excludes the commonly used assumptions of heavy tails and conditional heteroskedasticity in low-dimensional settings. Wong et al. (2020) considered the sub-Weibull assumption to allow a little bit heavier tails, and the geometric decaying β-mixing condition of the process is required. Adamek et al. (2023) employed the near-epoch dependence assumption to allow non-Gaussian, serially correlated and conditional heteroskedastic errors; see also Medeiros and Mendes (2016) and the references therein. The above proving techniques are both for lasso problems, and it is challenging to adapt them to our estimation. We leave it for future research. In Assumption 4, the running order

T_0 is required to grow at a rate of $T_0 \gtrsim \log(T_1)$ such that the effect of truncated terms $\boldsymbol{r}_t$'s can be dominated; see the technical proofs for details. In fact, to derive the asymptotic normality of low-dimensional VAR sieve estimation (Lewis and Reinsel, 1985), it is usually assumed that $T_1^{1/2}\sum_{j=T_0+1}^{\infty}\|\boldsymbol{A}_j^*\|_{\mathrm{op}} = o(1)$, which exactly corresponds to $T_0 \gtrsim \log(T_1)$ since $\sum_{j=T_0+1}^{\infty}\|\boldsymbol{A}_j^*\|_{\mathrm{op}} \lesssim \rho^{T_0}$ under Assumption 2.

We next state the oracle inequalities of $\widehat{\boldsymbol{\mathcal{A}}}$, which will rely on the temporal and cross-sectional dependence of $\{\mathbf{y}_t\}$ (Basu and Michailidis, 2015). To this end, we first define

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

where $\boldsymbol{\Psi}_*(z) = \sum_{j=0}^{\infty}\boldsymbol{\Psi}_j^* z^j$ for $z \in \mathbb{C}$, and $\boldsymbol{\Psi}_*^{\mathsf{H}}(z)$ is its complex conjugate. Note that, from Assumption 2, $\mu_{\max}(\boldsymbol{\Psi}_*) \leqslant C(1-\rho^2)^{-1}$ with C being an absolute constant. Let $\kappa_{\mathrm{RSC}} = \lambda_{\min}(\boldsymbol{\Sigma}_\varepsilon)\mu_{\min}(\boldsymbol{\Psi}_*)$ and $\kappa_{\mathrm{RSS}} = \lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)\mu_{\max}(\boldsymbol{\Psi}_*)$, where $\lambda_{\min}(\boldsymbol{\Sigma}_\varepsilon)$ and $\lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)$ are the minimum and maximum eigenvalues of $\boldsymbol{\Sigma}_\varepsilon$, respectively, and κ_{RSC} and κ_{RSS} are key quantities related to the restricted strong convexity and smoothness conditions (Raskutti et al., 2011).

Theorem 1 (Group Lasso oracle inequalities)**.** *Let $S \subset \{1,\ldots,T_0\}$ be an arbitrary index set with cardinality $|S| = s$ and denote $S^c = \{1,\ldots,T_0\}\backslash S$. Suppose that κ_{RSC} and κ_{RSS} are bounded away from zero and infinity, and Assumptions 1–4 hold. If $T_1 \gtrsim \{(r_1 \wedge r_2) + s^2\}N + s^2\log T_0$, and $\lambda \gtrsim \sqrt{\{(r_1 \wedge r_2)N + \log T_0\}/T_1}$, then with probability at least $1 - Ce^{-(r_1\wedge r_2)N - \log T_0}$,*

$$\|\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^*\|_{\mathrm{F}}^2 \lesssim \lambda^2 s + \underbrace{\lambda\|\boldsymbol{\mathcal{A}}_{S^c}^*\|_{\ddagger} + \tau^2\|\boldsymbol{\mathcal{A}}_{S^c}^*\|_{\ddagger}^2}_{\textit{approx error}} \quad \textit{and}$$

$$T_1^{-1}\|(\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^*)_{(1)}\boldsymbol{X}\|_{\mathrm{F}}^2 \lesssim \lambda^2 s + \underbrace{\lambda\|\boldsymbol{\mathcal{A}}_{S^c}^*\|_{\ddagger} + \tau^2\|\boldsymbol{\mathcal{A}}_{S^c}^*\|_{\ddagger}^2}_{\textit{approx error}},$$

where $\tau^2 = C\sqrt{(N + \log T_0)/T_1}$, and C is an absolute constant given in the proof.

Following the standard arguments for weak sparsity (Raskutti et al., 2011; Wainwright, 2019), the two upper bounds in the above theorem hold for any subset S with fixed cardinality s, and each of them consists of two terms: the estimation error, i.e. $\lambda^2 s$, is associated with estimating a total of s unknown $\boldsymbol{A}_j^*$'s, and the remaining part of $\boldsymbol{\mathcal{A}}^*$ that is not estimated, i.e. $\boldsymbol{\mathcal{A}}_{S^c}^*$, gives

rise to the approximation error. The optimal S can be chosen by trading off the estimation and approximation errors. Moreover, by Assumptions 2 and 4, it can be shown that the truncation errors, e_{trunc} and $\widetilde{e}_{\text{trunc}}$, are dominated by the estimation error $\lambda^2 s$. Hence, by balancing the three types of errors and further considering the exponential decay of $\boldsymbol{A}_j^*$ as $j \to \infty$, we can establish the following convergence result.

Theorem 2 (GLP)**.** *Suppose that κ_{RSC} and κ_{RSS} are bounded away from zero and infinity, and Assumptions 1–4 hold. If*

$$T_1 \gtrsim \left[(r_1 \wedge r_2) + \left\{ \frac{\log T_1}{\log(1/\rho)} \right\}^2 \right] N + \left\{ \frac{\log T_1}{\log(1/\rho)} \right\}^2 \log T_0, \tag{4.2}$$

and $\lambda \asymp \sqrt{\{(r_1 \wedge r_2)N + \log T_0\}/T_1}$, then with probability at least $1 - Ce^{-(r_1 \wedge r_2)N - \log T_0}$,

$$e_{\text{est}}(\widehat{\boldsymbol{\mathcal{A}}}_\infty) \lesssim \frac{\{(r_1 \wedge r_2)N + \log T_0\} \log T_1}{T_1 \log(1/\rho)} \text{ and } e_{\text{pred}}(\widehat{\boldsymbol{\mathcal{A}}}_\infty) \lesssim \frac{\{(r_1 \wedge r_2)N + \log T_0\} \log T_1}{T_1 \log(1/\rho)},$$

where C is an absolute constant given in the proof.

The optimal choice of S has cardinality $s \lesssim \log(\sqrt{N}/\lambda)/\log(1/\rho)$, which, together with the rate of λ in Theorem 2, implies that $s \lesssim \log(T_1)/\log(1/\rho)$, i.e. the number of active lags decreases as ρ decreases. This is expected given the relationship between the cutoff and ρ as discussed in Remark 6. Second, the term $\log T_0$ appears in the above theorems, and this is due to the Lasso regularization on T_0 groups of coefficients, $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_{T_0}$. Third, the upper bound on T_0, namely $\log T_0 \lesssim T_1/(\log T_1)^2$, is looser than $T_0 = o(T^{1/3})$, which is necessary for the low-dimensional VAR sieve estimation (Lewis and Reinsel, 1985). It is mainly due to the group Lasso penalty, and this makes it possible to consider a larger T_0 in real applications; see Section 5.1 for numerical evidences. Finally, the above two theorems can be easily adjusted when the two quantities, κ_{RSC} and κ_{RSS}, depend on N, T_0 and T_1, while the assumption that they are bounded away from zero and infinity can simplify the discussions on the three types of errors, as well as the running order selection.

Note that Theorem 2 gives error bounds uniformly for all GLPs, while some VAR models may have finite orders and the coefficient matrices at some lags are even exactly zero; see, e.g., the

commonly used seasonal VAR models in real applications (Cryer and Chan, 2008). For this case, truncation errors disappear, and we even do not need to handle approximation errors. Specifically, consider an VAR(T_0) model with its coefficient matrices satisfying the low-Tucker-rank assumption at (2.4), i.e. $\boldsymbol{r}_t = 0$ and $\widetilde{\boldsymbol{\varepsilon}}_t = \boldsymbol{\varepsilon}_t$, and Assumption 4 is no longer needed.

Corollary 1 (Finite-order VAR process)**.** *Consider an VAR(T_0) process with the VAR matrix polynomial $\boldsymbol{A}(z) = \boldsymbol{I} - \sum_{j=1}^{T_0} \boldsymbol{A}_j z^j$. Suppose that the determinant of $\boldsymbol{A}(z)$ is not equal to zero for all $z \in \mathbb{C}$ and $|z| < 1$. If Assumptions 2 and 3 are satisfied, then Theorem 1 still holds.*

For a group-sparse VAR process, i.e. $\boldsymbol{\mathcal{A}}^*_{S^c} = 0$, the approximation error becomes zero, and the error bounds then have a rate of $s\{(r_1 \wedge r_2)N + \log T_0\}/T_1$, which is sharper than that in Theorem 2 when $s \leqslant T_0$ is fixed or has a much slower rate than $\log(T_1)$.

4.2 Theoretical justifications for the algorithm with hard-thresholding

Algorithm 1 in Section 3.2 is used to solve the optimization problem at (2.6), while it is slightly different due to the alternating mechanism. This subsection provides theoretical justifications, including both statistical and convergence analysis, for this algorithm with hard-thresholding.

Consider the true coefficient tensor $\boldsymbol{\mathcal{A}}^*$ with $\boldsymbol{\mathcal{A}}^*_{(1)} = (\boldsymbol{A}^*_1, \boldsymbol{A}^*_2, \ldots, \boldsymbol{A}^*_{T_0}) \in \mathbb{R}^{N \times NT_0}$ in Section 2 and the parameter space $\boldsymbol{\Theta}^{\mathrm{SP}}(r_1, r_2, s)$ with low-rankness and group-sparsity in Section 3.1. For a given threshold $\gamma > 0$, let the active set $S_\gamma = \big\{j \in \{1, \ldots, T_0\} \mid \|\boldsymbol{A}^*_j\|_{\mathrm{F}} > \gamma\big\}$ with cardinality $s_\gamma = |S_\gamma|$, and $S^c_\gamma = \{1, \ldots, T_0\} \backslash S_\gamma$. We then define a random quantity

$$e_\gamma(r_1, r_2, s) := \sup_{\boldsymbol{\mathcal{M}} \in \boldsymbol{\Theta}^{\mathrm{SP}}(r_1, r_2, s),\ \|\boldsymbol{\mathcal{M}}\|_{\mathrm{F}} = 1} \langle \nabla \mathcal{L}(\boldsymbol{\mathcal{A}}^*_{S_\gamma}), \boldsymbol{\mathcal{M}} \rangle,$$

where s is the running sparsity level from Algorithm 1, and it is fixed in the forthcoming theoretical studies. This quantity is directly related to statistical errors of the proposed algorithm, and we first analyze it statistically by providing error bounds as in Section 4.1.

Theorem 3 (Statistical analysis)**.** *Suppose that κ_{RSC} and κ_{RSS} are bounded away from zero and infinity, and Assumptions 1–4 hold. For any $\gamma \gtrsim \sqrt{\{(r_1 \wedge r_2)N + \log T_0\}/T_1}$, if $T_1 \gtrsim \{(r_1 \wedge r_2) +$*

$s^2\}N + s^2 \log T_0$, *then with probability at least* $1 - Ce^{-(r_1 \wedge r_2)N - \log T_0}$,

$$e_\gamma^2(r_1, r_2, s) \lesssim \gamma^2 s + \|\boldsymbol{\mathcal{A}}^*_{S_\gamma^c}\|_{\mathrm{F}}^2 + \tau^2 \|\boldsymbol{\mathcal{A}}^*_{S_\gamma^c}\|_{\ddagger}^2,$$

where $\tau^2 = C\sqrt{(N + \log T_0)/T_1}$, *and* C *is an absolute constant given in the proof.*

The upper bound in the above theorem has a form similar to that in Theorem 1, and it increases as γ increases. As a result, we can obtain the statistical error bound below,

$$T_1^{-1}[(r_1 \wedge r_2)N + \log T_0][s + \log(T_1)/\log(1/\rho)] \tag{4.3}$$

by choosing $\gamma \asymp \sqrt{\{(r_1 \wedge r_2)N + \log T_0\}/T_1}$, and it can also be verified that $s_\gamma \lesssim \log T_1/\log(1/\rho)$.

We next conduct convergence analysis. To this end, for input $\boldsymbol{\mathcal{A}}^k$ at the k-th iteration, denote its active set by $S_k = \{j \in \{1, \ldots, T_0\} \mid \|\boldsymbol{\mathcal{A}}_j^k\|_{\mathrm{F}} \neq 0\}$, and let $\nu_k = |S_k \cup S_{k+1}|/s - 1$. Note that it corresponds to the case with $S_k = S_{k+1}$ if $\nu_k = 0$ and that with $S_k \cap S_{k+1} = \varnothing$ if $\nu_k = 1$, i.e. $\nu_k \in [0, 1]$ can be used to measure the size of overlap between S_k and S_{k+1}. Let $\nu_{\min} = \min_{0 \leqslant k \leqslant K-1} \nu_k$, and $\nu_{\max} = \max_{0 \leqslant k \leqslant K-1} \nu_k$. Moreover, denote $\sigma_L = \min\{\sigma_{\min}[(\boldsymbol{\mathcal{A}}^*_{S_\gamma})_{(1)}], \sigma_{\min}[(\boldsymbol{\mathcal{A}}^*_{S_\gamma})_{(2)}]\}$, $\sigma_U = \max\{\sigma_{\max}[(\boldsymbol{\mathcal{A}}^*_{S_\gamma})_{(1)}], \sigma_{\max}[(\boldsymbol{\mathcal{A}}^*_{S_\gamma})_{(2)}]\}$, and $\kappa = \sigma_U/\sigma_L$, where their dependence on γ is suppressed without confusion.

Theorem 4 (Convergence analysis)**.** *Consider Algorithm 1 with step size* $\eta = \eta_0(\kappa_{\mathrm{RSC}} + \kappa_{\mathrm{RSS}})^{-1}[(1 + \sigma_U)(1 + \sigma_U^{1/2})]^{-1}$, *where* $\eta_0 \leqslant \min\{150^{-1}, 204\sigma_U^{-1}\}$ *is a positive constant, and denote by* $e_{\mathrm{stat}}^2 = e_\gamma^2(r_1, r_2, 3s)$ *the statistical error. For a given* γ, *suppose that* $b \asymp \sigma_U^{1/4}$, $a \asymp (\kappa_{\mathrm{RSC}}^{-1} + \kappa_{\mathrm{RSS}}^{-1})^{-1}(\sigma_U^{1/2} + \sigma_U)$, $\|\boldsymbol{\mathcal{A}}^0 - \boldsymbol{\mathcal{A}}^*_{S_\gamma}\|_{\mathrm{F}}^2 \lesssim \sigma_L^{5/2}\kappa^{-3/2}$, $\nu_{\max} \lesssim \eta_0 \kappa_{\mathrm{RSC}}^2 \kappa_{\mathrm{RSS}}^{-2} \kappa^{-4}$, $s \geqslant \nu_{\min}^{-1} s_\gamma$, *and* $e_{\mathrm{stat}}^2 \lesssim \eta_0^2 \kappa_{\mathrm{RSC}}^4 \kappa_{\mathrm{RSS}}^{-4} \kappa^{-8}$. *If Assumptions 1–4 hold, and* $T_1 \gtrsim s^2(N + \log T_0)$,

$$\|\boldsymbol{\mathcal{A}}^K - \boldsymbol{\mathcal{A}}^*_{S_\gamma}\|_{\mathrm{F}}^2 \lesssim \kappa^{3/2}\sigma_L^{-1/2}\left(1 - \eta_0^2\delta^2\right)^K \|\boldsymbol{\mathcal{A}}^0 - \boldsymbol{\mathcal{A}}^*_{S_\gamma}\|_{\mathrm{F}}^2 + \kappa^{7/2}\sigma_L^{-1/2}\kappa_{\mathrm{RSC}}^{-2}\eta_0^{-2}\delta^{-2}e_{\mathrm{stat}}^2 \tag{4.4}$$

holds, with probability at least $1 - Ce^{-N - \log T_0}$, *where* $\delta = 1088^{-1}\kappa_{\mathrm{RSC}}\kappa_{\mathrm{RSS}}^{-1}\kappa^{-2}$, $\eta_0\delta < 1$, *and* C *is an absolute constant given in the proof.*

The two terms at the right hand side of (4.4) correspond to the optimization and statistical errors, respectively. The statistical error is discussed at Theorem 3 and, since $\eta_0\delta < 1$, the

linear convergence rate can be implied for the optimization error. Second, by triangle inequality, $0.5\|\boldsymbol{\mathcal{A}}^K - \boldsymbol{\mathcal{A}}^*\|_{\mathrm{F}}^2 \leqslant \|\boldsymbol{\mathcal{A}}^K - \boldsymbol{\mathcal{A}}^*_{S_\gamma}\|_{\mathrm{F}}^2 + \|\boldsymbol{\mathcal{A}}^*_{S_\gamma^c}\|_{\mathrm{F}}^2$, and hence the convergence analysis at Theorem 4 can be readily extended to include approximation errors. Third, if we further assume that κ_{RSS}, κ_{RSC}, σ_U and σ_L are bounded away from zero and infinity as in all the other theorems, then the tuning parameters a, b and η in Algorithm 1 will be at a constant level, i.e. they do not depend on N, T_0 or T_1. Finally, it is required by Theorem 4 that $s \geqslant s_\gamma$ and, from (4.3), we then can choose $s \asymp \log T_1/\log(1/\rho)$. This hence leads to the following results for GLPs.

Corollary 2 (GLP with hard-thresholding)**.** *Suppose that the conditions of Theorems 3 and 4 hold, and we choose $s \asymp \log T_1/\log(1/\rho)$ in Algorithm 1. After the K-th iteration with*

$$K \gtrsim \frac{\log(\kappa^{7/2}\sigma_L^{-5/2}\kappa_{\mathrm{RSC}}^{-2}\delta^{-2})}{\log(1-\eta_0^2\delta^2)},$$

and η_0 and δ given in Theorem 4, if $T_1 \gtrsim \{(r_1 \wedge r_2) + s^2\}N + s^2 \log T_0$, it then holds that

$$\|\boldsymbol{\mathcal{A}}^K - \boldsymbol{\mathcal{A}}^*\|_{\mathrm{F}}^2 \lesssim \frac{[(r_1 \wedge r_2)N + \log T_0]s}{T_1}$$

with probability at least $1 - Ce^{-(r_1 \wedge r_2)N - \log T_0}$, where C is an absolute constant given in the proof.

The above corollary gives the same bound as that in Theorem 2 since $s \asymp \log T_1/\log(1/\rho)$. Moreover, when the quantities of $\kappa_{\mathrm{RSC}}, \kappa_{\mathrm{RSS}}, \kappa$ and σ_L are bounded away from zero and infinity, the required number of iterations does not depend on N, T_0 or T_1, and this makes sure that the proposed algorithm can be applied to large datasets without any difficulty. Finally, as in Section 2.3, for group-sparse VAR(T_0) processes, i.e. $\boldsymbol{\mathcal{A}}^*_{S_\gamma^c} = 0$ for some $\gamma > 0$, we can obtain the same bound as that in Corollary 2, while s may be fixed or have a much slower rate than $\log(T_1)$.

Remark 10 (Connection between estimators based on soft- vs. hard-thresholding). The estimators obtained based on soft- and hard-thresholding methods differ in Line 7 of Algorithm 1, while Line 6 applies to both. This is illustrated in the middle and right panels of Figure 1. Specifically, the middle panel of Figure 1 depicts the estimated lag-j coefficient matrices before the thresholding operation, $\widetilde{\boldsymbol{A}}_j^K$'s, which are obtained by Line 6 of Algorithm 1 at the Kth iteration. The top-right panel of Figure 1 demonstrates the bias λ for each nonzero $\|\widetilde{\boldsymbol{A}}_j^K\|_F$ resulting from the soft-thresholding. In the bottom-right panel of Figure 1, γ represents a chosen cutoff threshold such

that all $\widetilde{\boldsymbol{A}}_j^K$'s with $\|\widetilde{\boldsymbol{A}}_j^K\|_{\mathrm{F}} \leqslant \gamma$ will be truncated, and the corresponding active set is denoted by S_γ. It can be shown that the same active set S_γ can be obtained from $\widetilde{\boldsymbol{A}}_j^K$'s by either the hard-thresholding operation with $s = s_\gamma := |S_\gamma|$ or the soft-thresholding with some $\lambda \asymp \gamma$.

5 Simulation studies

5.1 Estimation performance of the proposed methodology

This subsection conducts two simulation experiments to evaluate the finite-sample performance of VAR sieve estimators $\boldsymbol{\mathcal{A}}^K$ from Algorithm 1 in Section 3.

The first experiment is to evaluate how the three types of errors, i.e. the estimation, approximation and truncation errors, can be balanced numerically when the sparsity level s varies. We consider two data generating processes below,

(1) VAR process: $\mathbf{y}_t = \boldsymbol{\Phi}_1\mathbf{y}_{t-1} + \boldsymbol{\Phi}_4\mathbf{y}_{t-4} + \boldsymbol{\Phi}_5\mathbf{y}_{t-5} + \boldsymbol{\Phi}_8\mathbf{y}_{t-8} + \boldsymbol{\Phi}_9\mathbf{y}_{t-9} + \boldsymbol{\varepsilon}_t$ with $\boldsymbol{\Phi}_j = \rho^j \boldsymbol{B}_j\boldsymbol{C}_j'$ and $0 < \rho < 1$. For any given r_1, r_2, $\boldsymbol{B}_j, \boldsymbol{C}_j$ are $N \times r^{(j)}$ matrices with orthonormal columns such that $\mathrm{rank}[(\boldsymbol{B}_1, \boldsymbol{B}_4, \boldsymbol{B}_5, \boldsymbol{B}_8, \boldsymbol{B}_9)] = r_1$ and $\mathrm{rank}[(\boldsymbol{C}_1, \boldsymbol{C}_4, \boldsymbol{C}_5, \boldsymbol{C}_8, \boldsymbol{C}_9)] = r_2$ and $r^{(j)} \leqslant \min(r_1, r_2)$ for $j \in \{1, 4, 5, 8, 9\}$.

(2) VARMA process: $\mathbf{y}_t = \boldsymbol{\Phi}\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t - \boldsymbol{\Theta}\boldsymbol{\varepsilon}_{t-1}$ with $\boldsymbol{\Phi} = -0.5\boldsymbol{B}\boldsymbol{J}\boldsymbol{B}'$, $\boldsymbol{\Theta} = \rho\boldsymbol{B}\boldsymbol{J}\boldsymbol{B}'$ and $0 < \rho < 1$, where $\boldsymbol{B}$ is an $N \times N$ orthonormal matrix and for any given r, $\boldsymbol{J} = \mathrm{diag}\{\mathbf{1}_r, \mathbf{0}_{N-r}\}$ with $\mathbf{1}_r = (1, \ldots, 1)'$ is an N-dimensional diagonal matrix with rank r.

The VAR process is a group-sparse VAR(9) with nonzero coefficients at five lags. The dimension of the row and column spaces of $\{\boldsymbol{\Phi}_j, j \geqslant 1\}$ are given by $\mathrm{rank}[(\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_4, \boldsymbol{\Phi}_5, \boldsymbol{\Phi}_8, \boldsymbol{\Phi}_9)] = r_1$ and $\mathrm{rank}[(\boldsymbol{\Phi}_1', \boldsymbol{\Phi}_4', \boldsymbol{\Phi}_5', \boldsymbol{\Phi}_8', \boldsymbol{\Phi}_9')] = r_2$, respectively. We set $r^{(4)} = \max(r_1, r_2) - \min(r_1, r_2)$ and $r^{(j)} = \min(r_1, r_2)$ for $j \in \{1, 5, 8, 9\}$, and matrices $\boldsymbol{B}_j$'s and $\boldsymbol{C}_j$'s are generated randomly; see the supplementary file for details. At each generation, we ensure that the stationarity of VAR processes holds. The VARMA process has a weakly group-sparse VAR(∞) form at (1.2) with $\boldsymbol{A}_j^* = \boldsymbol{\Theta}^{j-1}(\boldsymbol{\Phi} - \boldsymbol{\Theta}) = (-0.5 - \rho)\rho^{j-1}\boldsymbol{B}\boldsymbol{J}\boldsymbol{B}'$ for all $j \geqslant 1$. The row and column spaces of all

coefficient matrices $\{\boldsymbol{A}_j^*, j \geqslant 1\}$ are spanned by first r columns of $\boldsymbol{B}$, and subsequently, the dimension of the row and column spaces of $\{\boldsymbol{A}_j^*, j \geqslant 1\}$ are given by $\text{rank}[\{\boldsymbol{A}_j^*, j \geqslant 1\}] = r_1$ and $\text{rank}[\{\boldsymbol{A}_j^{*\prime}, j \geqslant 1\}] = r_2$ with $r_1 = r_2 = r$. Since the spectral radius of $\boldsymbol{\Phi}$ is 0.5, the VARMA model is stationary.

We fix the settings at $(\rho, N, T) = (0.7, 100, 4000)$ and set $(r_1, r_2) = (4, 2)$ for the VAR process and $r = 4$ for the VARMA process. There are 500 replications for each data generation setting, and we independently generate $\boldsymbol{B}_j$'s and $\boldsymbol{C}_j$'s for VAR or $\boldsymbol{B}$ for VARMA at each replication. Algorithm 1 is applied to each generated sequence with $T_0 = \lfloor \sqrt{T} \rfloor = 63$ and s varying from 3 to 35, and we can obtain the output $\boldsymbol{\mathcal{A}}^K$ until the algorithm converges. The estimation, approximation and truncation errors refer to $\|\boldsymbol{\mathcal{A}}^K - \boldsymbol{\mathcal{A}}_S^*\|_{\text{F}}^2$, $\|\boldsymbol{\mathcal{A}}_{S^c}^*\|_{\text{F}}^2$ and $\sum_{j=T_0+1}^{\infty} \|\boldsymbol{A}_j^*\|_{\text{F}}^2$, respectively, and the parameter estimation error is defined as $\|\boldsymbol{\mathcal{A}}^K - \boldsymbol{\mathcal{A}}^*\|_{\text{F}}^2 = \|\boldsymbol{\mathcal{A}}^K - \boldsymbol{\mathcal{A}}_S^*\|_{\text{F}}^2 + \|\boldsymbol{\mathcal{A}}_{S^c}^*\|_{\text{F}}^2$, where S contains the indices of all estimated active coefficient matrices. The truncation error is zero for the VAR process and 2.33×10^{-21} for the VARMA process, and hence they can be ignored numerically comparing with the other two types of errors. Figure 2 gives the estimation and approximation errors, averaged over 500 replications, and we have three findings below. First, as the sparsity level s increases, linear growth in the estimation errors can be roughly observed for both the VARMA and VAR processes. The approximation error decreases quickly for both processes and, when $s > 5$, it becomes almost zero for the VAR process since the active set can be correctly selected for most replications. Second, for the VARMA process, the approximation error is dominating for the cases with $s < 10$, while the estimation error has much larger values when $s > 10$. As a result, as s increases, the parameter estimation error decreases first and then increases when the sparsity level $s > 10$. This phenomenon can also be observed for the VAR process, and the parameter estimation error reaches the minimum at $s = 5$, which is the true sparsity level. Finally, for large sparsity levels s, the parameter estimation error exhibits linearity for both processes, which is consistent with the theoretical findings at Corollary 2.

The second experiment is to further verify the theoretical bound of parameter estimation errors at Corollary 2, and the two data generating processes in the first experiment are employed again. For the VAR process, the parameter estimation error is expected to have a rate of $\beta =$

$s((r_1 \wedge r_2)N + \log T_0)/(T - T_0)$, while for the VARMA process, given that $r_1 = r_2 = r$, the rate degenerates to $\beta = s(rN + \log T_0)/(T - T_0)$. Moreover, since the linearity with respect to s has already been confirmed in the first experiment, we fix the sparsity level $s = 5$ for the VAR process and $s = 10$ for the VARMA process in this experiment. Finally, we consider three different rates for running orders, i.e. $T_0 = \lfloor cT^\alpha \rfloor$ with $\alpha = 1/4$, $1/3$ or $1/2$, and the true order with $T_0 = 9$ is also considered for VAR models. The value of c is set to 1.5 for the case with $\alpha = 1/2$, while $c = 3$ for those with $\alpha = 1/4$ and $1/3$ such that the resulting T_0 is not too small under the smallest sample size setting.

In order to verify the linearity of parameter estimation errors with respect to the rate β, rank r and dimension N, we consider three groups of settings to generate high-dimensional time series. First, by fixing $(N, r_1, r_2) = (100, 4, 2)$ for the VAR process and $(N, r) = (100, 4)$ for the VARMA process, one can vary the sample size T such that the values of β are equally spaced between 0.4 to 1.0. Second, we fix $(N, T, r_2) = (100, 1200, 4)$ and let r_1 vary in $\{2, 3, 4, 5\}$ for the VAR process, and fix $(N, T) = (100, 2000)$ and let r vary in $\{2, 3, 4, 5\}$ for the VARMA process. Finally, the dimension N varies among $\{25, 50, 75, 100\}$, while (r_1, r_2, T) is fixed at $(4, 2, 1200)$ for the VAR process and (r, T) is fixed at $(4, 2000)$ for the VARMA process. All the other settings are the same as those in the first experiment, and Figure 3 presents the plots of parameter estimation errors, averaged over 500 replications, against the rate β, rank r_1 or r, and dimension N, respectively. The parameter estimation errors change linearly with respect to β and N for both data generating processes, and with respect to r for the VARMA process. For the VAR process, the errors grow linearly with respect to r_1 when $2 \leqslant r_1 \leqslant 4$ and remain flat when r_1 increases to 5. Given that $r_2 = 4$, this trend verifies that the error rate is dependent on $r_1 \wedge r_2$. The above findings confirm the theoretical results at Corollary 2. Moreover, we can observe that the parameter estimation errors are relatively insensitive to different settings of T_0 holding N, r or r_1 and T fixed. When T_0 is larger, due to the decrease in effective sample size, the parameter estimation errors may become slightly worse but the difference is not obvious.

5.2 Comparison of predictive performance

This subsection evaluates the predictive performance of the proposed method against existing ones. Specifically, we provide a comprehensive set of benchmarking methods below:

(i) Default benchmarks: The common benchmarking models used in economics are the random walk, univariate AR(1), AR(2), and unregularized VAR(1), VAR(2).

(ii) VAR(p) models: The VAR(p) models comprise Lasso-regularized VAR (Basu and Michailidis, 2015) with ℓ_1-penalty on coefficient matrices; the multilinear low-rank (MLR) VAR (Wang et al., 2022b) which assumes low-rank structure on the coefficient matrices and along the lag dimension of the stacked coefficient matrices; the sparse higher-order reduced-rank (SHORR) model which further imposes sparsity on the decomposed loading matrices in the MLR model; and Bayesian VAR (BVAR) with zero-mean natural conjugate prior (Chan et al., 2016) that shrinks coefficient matrices to zero.

(iii) VARMA-based models: The VARMA models include those with ℓ_1- or HLag-penalties on both the AR and MA coefficient matrices respectively (Wilms et al., 2023). A parametric VAR(∞) model is concurrently introduced in Zheng (2024) which models the temporal lags using a parametric form derived from a reparametrization of the VARMA model. We refer to it as the "Approx VARMA" model in the rest of the paper.

(iv) An adaption of factor-augmented regression: As another benchmark, we adapt the factor-augmented regression (Stock and Watson, 2002a,b) to multivariate time series forecasting. Their original method aims to forecast a univariate time series based on factors $\boldsymbol{f}_t \in \mathbb{R}^{r\times 1}$ extracted from a multivariate time series $\mathbf{x}_t \in \mathbb{R}^{N\times 1}$. For example, for forecasting y_{it}, the model can be written as $y_{it} = \boldsymbol{\beta}_i'(L)\boldsymbol{f}_t + \gamma_i(L)y_{it} + \varepsilon_{it}$, with $\mathbf{x}_t = \boldsymbol{\Lambda}\boldsymbol{f}_t + \boldsymbol{e}_t$, where L is the lag operator, and $\boldsymbol{\beta}_i(L) = \sum_{k=1}^{p}\boldsymbol{\beta}_{ik}L^k \in \mathbb{R}^{r\times 1}$ and $\gamma_i(L) = \sum_{k=1}^{q}\gamma_{ik}L^k \in \mathbb{R}$ are polynomials in L. In our ad-hoc adaptation, we let $\mathbf{x}_t = \mathbf{y}_t$ and $q = 1$ in the model for each $1 \leqslant i \leqslant N$, and simply apply the PCA to $\mathbf{y}_t$ to extract the common factors. Then univariate time series

models are fitted to each y_{it} with p lags of the extracted factors and $y_{i,t-1}$. We refer to it as "FactorAug Reg" in the rest of the paper.

We assess the predictive performance of these models by generating data from the VAR process. To better align the data with the proposed model, we introduce small modifications to the VAR data generating process in the previous subsection. Specifically, when generating matrices $\boldsymbol{B}_j$'s and $\boldsymbol{C}_j$'s, we only keep those with the maximum ℓ_1 norm of the row vectors being smaller than 0.55. As a result, the resulting coefficient matrices are purely low-rank but not sparse. We fix $(N, T, r_1, r_2, \rho) = (100, 1200, 4, 2, 0.9)$ with $T_0 = \lfloor \sqrt{T} \rfloor$. Moreover, we set $\boldsymbol{\Phi}_1$ to zero so that the remaining non-sparse lags are $4, 5, 8$ and 9. To evaluate the forecasting performance, we generate $n = 100$ realized sequences $\{\mathbf{y}_t^{(k)}, 1 \leqslant t \leqslant T\}$ of length T with the last observation $\mathbf{y}_T^{(k)}$ reserved for one-step-ahead forecasting evaluation, for $1 \leqslant k \leqslant n$. The one-step-ahead mean squared forecast error can be calculated as

$$\text{MSFE}_{\text{last-step,model } i} = \frac{1}{n} \sum_{k=1}^{n} \|\widehat{\mathbf{y}}_{T,\text{model } i}^{(k)} - \mathbf{y}_T^{(k)}\|_2^2,$$

where $\widehat{\mathbf{y}}_{T,\text{model } i}^{(k)} = \mathbb{E}_{\text{emp}}(\mathbf{y}_T^{(k)} \mid \mathbf{y}_{T-1}^{(k)}, \mathbf{y}_{T-2}^{(k)}, \ldots, \text{model } i)$ and $\mathbb{E}_{\text{emp}}(\cdot)$ denotes the empirical mean and model i denotes either our proposed model or one of the aforementioned benchmark models. And the subscript "last-step" is added to distinguish its definition with the one-step-ahead rolling forecast error to be defined in the empirical studies. To assess the significance of differences in MSFEs across models, we construct a model confidence set (MCS) and report the corresponding p-values, following the methodology of Hansen et al. (2011). An existing package by Aka and Tschernig (2018) is adopted. Specifically, we use the difference in squared ℓ_2 forecast errors as the relative performance variable and the MCS p-values are calculated using the deviation test statistic, while we also confirm that using range test statistics will lead to similar observations.

The benchmark methods are implemented following details from their respective papers. The supplementary file provides additional details about this. To ensure comparability with the univariate AR and unregularized VAR benchmarks, we set the AR order to $p = 2$ for the finite-order VAR models. The MSFEs and MCS p-values are presented in Table 1. Higher p-values indicate a lower likelihood of the model being excluded from the set of potential 'best' models, suggesting

a greater probability that the model is among the most accurate. Overall, our proposed method outperforms all the benchmarks. Sparsity-based methods generally underperform compared to low-rank or factor-based approaches since the coefficient matrices are low-rank but non-sparse. The worse performance of finite-order VAR models, compared to our method, suggests that truncating the lags may introduce bias, resulting in inferior performance. Similarly, the poorer results from the VARMA model may be due to its inability to capture non-consecutive non-zero lags, leading to a misrepresentation of the temporal structure.

6 Macroeconomic application

In our empirical analysis, we utilize quarterly macroeconomic variables obtained from the FRED-QD dataset (McCracken and Ng, 2020). These variables span a wide range of categories, including prices, earnings and productivity, interest rates, money and credit, exchange rates, stock market, household and non-household balance sheets. These categories are usually considered in the construction of financial condition indices, since they reflect important factors that can affect the stance of monetary policy and aggregate supply and demand conditions (Bulut, 2016; Hatzius et al., 2010). The sequences span from the first quarter of 1959 to the fourth quarter of 2019, covering 244 time points prior to the onset of the COVID-19 pandemic. After excluding variables with incomplete observations, following the same methods as in McCracken and Ng (2020), we transform all sequences to stationarity and standardize them to zero mean and unit variance. Detailed transformations are provided in the supplementary file. We create three datasets of different sizes: "large" with all 112 variables, "medium" with 45 variables from the price category, and "small" with 11 consumer price index related variables from the price category.

We use Algorithm 1 with hard-thresholding to conduct the VAR sieve estimation and initially set the running order to $T_0 = \lfloor \sqrt{T} \rfloor = 15$. It roughly needs two minutes to finish one searching with around 5000 iterations for the large dataset. Based on the initial estimation, if the estimated coefficient matrices for lags beyond $T_0 - k$ are all zero, we then refine T_0 by reducing it to $T_0 - \max(0, k-2)$. Using the AIC given in Section 3.2, the chosen hyperparameters for the large,

medium, and small datasets are $(r_1, r_2, s, T_0) = (1, 1, 1, 3)$, $(1, 1, 1, 3)$, and $(1, 3, 3, 10)$, respectively. For the small dataset, lags 2, 3 and 8 are selected, while the first lag is selected for large and medium datasets. We first compare the predictive performance of the proposed method with the benchmark ones listed in Section 5.2, including default benchmarks, and VAR-based, VARMA-based and factor-based models. We implement each model according to the details provided in the corresponding papers and codes. For the ℓ_1-penalized, MLR, and SHORR VAR models, the AR order is set to one, as selected by BIC, while it is set to four by default in the BVAR implementation. A rolling forecast procedure is adopted for evaluating the predictive performance. Specifically, we first fit the models using the historical data with the ending point iterating from the fourth quarter of 2001 to the third quarter of 2019, and then one-step-ahead prediction $\widehat{\mathbf{y}}_{t+1,\text{model } i}$ as defined in Section 5.2 is produced at every iteration. We use the one-step-ahead rolling mean squared forecast error

$$\text{MSFE}_{\text{model } i} = \frac{1}{72} \sum_{t=173}^{244} \|\widehat{\mathbf{y}}_{t+1,\text{model } i} - \mathbf{y}_{t+1}\|_2^2$$

as our evaluation metric. To assess the significance of differences in MSFEs across all models, we again calculate MCS p-values (Hansen et al., 2011) as in Section 5.2.

Table 2 compares the forecasting performance including MSFE and MCS p-values of various models across small, medium, and large datasets. Our proposed model consistently achieves the lowest MSFEs across all dataset sizes, outperforming the benchmark models. Several key observations can be drawn from the results: (1) The superior performance of the proposed model suggest that the response and predictor subspaces for this macroeconomic dataset are both low-rank and differ from each other. (2) The unregularized VAR and random walk models perform the worst overall, followed by univariate AR methods. (3) The MLR, SHORR, BVAR with shrinkage priors, and factor-augmented regression models perform worse than VAR models with ℓ_1 regularization. (4) VARMA-based methods generally outperform standard VAR-based methods. In addition to the above overall MSFEs, we also plot the cumulative MSFEs in Figure 4 to compare the trajectory of cumulative forecasting errors. Specifically, we calculate the cumulative MSFE ratio of each benchmark method again ours. The four strongest competitors are VAR with with

ℓ_1 penalty, VARMA with ℓ_1 penalty, VARMA with HLag penalty, and the approximate VARMA approach as shown in the bottom panel of Figure 4. The results show that, for the small dataset, our method outperforms all other methods during the periods after 2010. For the medium dataset, our method consistently outperforms all other methods across the entire period from 2003 to end of 2019. Finally, for the large dataset, our method surpasses all others from 2010 to end of 2019.

To further compare the estimated loading matrices of the response and predictor factors, we present the identifiable projection matrices $\widehat{\boldsymbol{U}}_i\widehat{\boldsymbol{U}}_i'$ with $\widehat{\boldsymbol{U}}_i$'s being orthonormal for $i = 1$ or 2 since the loading matrices are not uniquely defined. Figures 5 and 6 show the projection matrices estimated from the large dataset, while those from the medium and small datasets are deferred to the supplementary file. To improve visualization, we reorder the variables as follows: the variable with the largest diagonal value is treated as the first variable. The remaining variables are then ordered according to their values in the first row, from largest to smallest. Next, we select the variables whose diagonal values are at least 0.01 to form a submatrix, and we provide their corresponding names with detailed description in Table **??** at the supplementary file to facilitate interpretation.

As shown by the projection matrix in Figure 5, the predictor factors can be broadly categorized into two groups. The first group predominantly reflects activities related to production, consumption, and investment (e.g., PCED, PPI_FinConsGds_Food,GPDI_Del), along with measures of credit risk within the economy (e.g., Total_Reserves_Repository, Nonborrowed_Reserves_-Repository, BAA_GS10). In contrast, the second group primarily captures changes in short-term interest rates (e.g., TM_3M_FedFunds, FedFunds) and wealth-related metrics (e.g., Real_HHW_-RESA, Real_AHE_MFG). These two groups exert opposing influences on the response variables, reflecting distinct economic driving forces derived from historical data.

Meanwhile, the projection matrix for the response factors, depicted in Figure 6, reveals patterns that differ significantly from those of the predictor factors. Here too, the factors can be grouped into two categories: the first group primarily captures measures of inflation, interest rates, credit, liabilities, and wealth. The second group, represented by long-term interest rates and bonds, responds differently to the predictor variables. This divergence is likely due to the fact that these

indicators are also influenced by market expectations of future economic growth, introducing a level of complexity that cannot be fully captured by historical data alone.

7 Conclusion and discussion

This paper proposes a supervised factor model for high-dimensional time series by introducing low-rank structures to the coefficient matrices of VAR(∞) models. With the help of tensor techniques, the proposed model can be rewritten into a form of two factor models, which allows us to interpret it from unsupervised factor modeling perspectives. For its application on high-dimensional time series, by making use of an interesting fact that the stationarity condition implies the weak group sparsity of coefficient matrices, a rank-constrained group Lasso estimation is considered, and its non-asymptotic properties are carefully investigated by trading-off the estimation, approximation, and truncation errors. Moreover, an alternating gradient descent algorithm with hard-thresholding is suggested to search for the high-dimensional estimate, and its theoretical properties, including both statistical and convergence analysis, are also provided. Finally, as illustrated by empirical analysis, the proposed model exceeds the existing methods in terms of forecasting accuracy while enjoying the nice interpretation of factor models.

The proposed methodology in this paper can be extended along three directions. First, to obtain a reliable estimator, the sample size is required to be $T_1 \gtrsim \{(r_1 \wedge r_2) + s^2\}N + s^2 \log T_0$ in Theorem 1, while the number of variables N may be larger than the sample size T_1, say for time-course gene expression data (Lozano et al., 2009). To handle this case, the group sparsity can be further imposed to the rows of factor matrices $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$, and we then can construct an alternating gradient descent method with three thresholdings, which is similar to Algorithm 1 in Section 3.1. Second, our current proving techniques heavily depend on the exponential decay of coefficient matrices, which actually excludes many important time series models such as the fractionally integrated autoregressive moving average (ARFIMA) model with long memory (Grange and Joyeux, 1980). It is urgent to look for a new proving technique to remove this restriction. Finally, it is of interest to make further inference on estimated coefficient matrices

such as checking the significance of some coefficients (Xia et al., 2022; Cai et al., 2020), and we will leave it for future research.

Acknowledgements

We are deeply grateful to the Co-Editor, an Associate Editor and three anonymous referees for their valuable comments that led to the substantial improvement in the quality of this paper. Zheng was supported by the National Science Foundation Grant DMS-2311178. Li was supported by the Hong Kong Research Grant Council Grants 17306121 and 17313722, and the National Natural Science Foundation of China Grant 72033002.

References

Adamek, R., Smeekes, S., and Wilms, I. (2023). Lasso inference for high-dimensional time series. *Journal of Econometrics*, 235:1114–1143.

Agarwal, A., Negahban, S., and Wainwright, M. J. (2012). Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics*, 40(5):2452 – 2482.

Aka, N. and Tschernig, R. (2018). *modelconf: Estimation of Model Confidence Sets*. R package version 0.1.0.

Amengual, D. and Watson, M. W. (2007). Consistent estimation of the number of dynamic factors in a large N and T panel. *Journal of Business & Economic Statistics*, 25(1):91–96.

Bai, J. and Wang, P. (2016). Econometric analysis of large factor models. *Annual Review of Economics*, 8:53–80.

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.

Basu, S. and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43:1535–1567.

Billio, M., Casarin, R., Iacopini, M., and Kaufmann, S. (2023). Bayesian dynamic tensor regression. *Journal of Business & Economic Statistics*, 41:429–439.

Bühlmann, P. (1997). Sieve bootstrap for time series.

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

Cai, C., Poor, H. V., and Chen, Y. (2020). Uncertainty quantification for nonconvex tensor completion: Confidence intervals, heteroscedasticity and optimality. In *International Conference on Machine Learning*, pages 1271–1282. PMLR.

Carriero, A., Kapetanios, G., and Marcellino, M. (2016). Structural analysis with multivariate autoregressive index models. *Journal of Econometrics*, 192(2):332–348.

Chan, J. C., Eisenstat, E., and Koop, G. (2016). Large Bayesian VARMAs. *Journal of Econometrics*, 192:374–390.

Chen, E. Y. and Fan, J. (2023). Statistical inference for high-dimensional matrix-variate factor models. *Journal of the American Statistical Association*, 118(542):1038–1055.

Chen, R., Xiao, H., and Yang, D. (2021). Autoregressive models for matrix-valued time series. *Journal of Econometrics*, 222(1):539–560.

Chen, R., Yang, D., and Zhang, C.-H. (2022). Factor models for high-dimensional tensor time series. *Journal of the American Statistical Association*, 117(537):94–116.

Chi, Y., Lu, Y. M., and Chen, Y. (2019). Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67:5239–5269.

Cryer, J. D. and Chan, K.-S. (2008). *Time Series Analysis: with Applications in R*. Springer, New York.

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.

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.

Fan, J., Masini, R., and Medeiros, M. C. (2022). Bridging factor and sparse models. *The Annals of Statistics*, To appear.

Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer, New York.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). The generalized dynamic-factor model: Identification and estimation. *Review of Economics and Statistics*, 82:540–554.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2005). The generalized dynamic factor model: one-sided estimation and forecasting. *Journal of the American Statistical Association*, 100:830–840.

Gao, Z. and Tsay, R. S. (2023). A two-way transformed factor model for matrix-variate time series. *Econometrics and Statistics*, 27:83–101.

Goldenshluger, A. and Zeevi, A. (2001). Nonasymptotic bounds for autoregressive time series modeling. *The Annals of Statistics*, 29:417–444.

Grange, C. and Joyeux, R. (1980). An introduction to long-range time series models and fractional differencing. *Journal of Time Series Analysis*, 1:15–29.

Guo, S., Wang, Y., and Yao, Q. (2016). High-dimensional and banded vector autoregressions. *Biometrika*, 103:889–903.

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

Han, R., Willett, R., and Zhang, A. R. (2022). An optimal statistical and computational framework for generalized tensor estimation. *The Annals of Statistics*, 50:1–29.

Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2):453–497.

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 16150, National Bureau of Economic Research.

Ing, C.-K. and Wei, C.-Z. (2005). Order selection for same-realization predictions in autoregressive processes. *The Annals of Statistics*, 33(5):2423 – 2474.

Koop, G., Korobilis, D., and Pettenuzzo, D. (2019). Bayesian compressed vector autoregressions. *Journal of Econometrics*, 210(1):135–154.

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.

Lewis, R. and Reinsel, G. C. (1985). Prediction of multivariate time series by autoregressive model fitting. *Journal of Multivariate Analysis*, 16:393–411.

Li, G., Leng, C., and Tsai, C.-L. (2014). A hybrid bootstrap approach to unit root tests. *Journal of Time Series Analysis*, 35:299–321.

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.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer Science & Business Media, New York.

McCracken, M. and Ng, S. (2020). FRED-QD: A quarterly database for macroeconomic research. Working Paper 26872, National Bureau of Economic Research.

Medeiros, M. C. and Mendes, E. F. (2016). ℓ_1-regularization of high-dimensional time-series models with non-Gaussian and heteroskedastic errors. *Journal of Econometrics*, 191:255–271.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27:538–557.

Nicholson, W. B., Matteson, D. S., and Bien, J. (2017). VARX-L: Structured regularization for large vector autoregressions with exogenous variables. *International Journal of Forecasting*, 33:627–651.

Nicholson, W. B., Wilms, I., Bien, J., and Matteson, D. S. (2020). High dimensional forecasting via interpretable vector autoregression. *Journal of Machine Learning Research*, 21:1–52.

Peña, D. and Tsay, R. S. (2021). *Statistical Learning for Big Dependent Data*. John Wiley & Sons, New Jersey.

Raskutti, G., Wainwright, M. J., and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57:6976–6994.

Samadi, S. Y. and Herath, H. W. B. (2024). Reduced-rank envelope vector autoregressive model. *Journal of Business & Economic Statistics*, 42(3):918–932.

Shen, J. and Li, P. (2017). A tight bound of hard thresholding. *Journal of Machine Learning Research*, 18:7650–7691.

Shibata, R. (1980). Asymptotically Efficient Selection of the Order of the Model for Estimating Parameters of a Linear Process. *The Annals of Statistics*, 8:147 – 164.

Stock, J. H. and Watson, M. W. (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179.

Stock, J. H. and Watson, M. W. (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–162.

Tropp, J. A. and Wright, S. J. (2010). Computational methods for sparse solution of linear inverse problems. *Proceedings of the IEEE*, 98:948–958.

Tsay, R. S. (2014). *Multivariate Time Series Analysis: With R and Financial Applications*. John Wiley & Sons, New Jersey.

Tu, S., Boczar, R., Simchowitz, M., Soltanolkotabi, M., and Recht, B. (2016). Low-rank solutions of linear matrix equations via procrustes flow. In *International Conference on Machine Learning*, pages 964–973. PMLR.

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31:279–311.

Velu, R. P. and Reinsel, G. C. (2013). *Multivariate Reduced-rank Regression: Theory and Applications*, volume 136. Springer Science & Business Media, New York.

Wainwright, M. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Wang, D., Liu, X., and Chen, R. (2019). Factor models for matrix-valued high-dimensional time series. *Journal of econometrics*, 208(1):231–248.

Wang, D., Zhang, X., Li, G., and Tsay, R. (2022a). High-dimensional vector autoregression with common response and predictor factors. *arXiv preprint arXiv:2203.15170*.

Wang, D., Zheng, Y., and Li, G. (2023). High-dimensional low-rank tensor autoregressive time series modeling. *Journal of Econometrics*, To appear.

Wang, D., Zheng, Y., Lian, H., and Li, G. (2022b). High-dimensional vector autoregressive time series modeling via tensor decomposition. *Journal of the American Statistical Association*, 117:1338–1356.

Wang, L., Zhang, X., and Gu, Q. (2017). A unified computational and statistical framework for nonconvex low-rank matrix estimation. In *Artificial Intelligence and Statistics*, pages 981–990. PMLR.

Wilms, I., Basu, S., Bien, J., and Matteson, D. (2023). Sparse identification and estimation of large-scale vector autoregressive moving averages. *Journal of the American Statistical Association*, 118:571–582.

Wilms, I., Basu, S., Bien, J., and Matteson, D. S. (2017). Interpretable vector autoregressions with exogenous time series. *arXiv preprint arXiv:1711.03623*.

Wong, K. C., Li, Z., and Tewari, A. (2020). Lasso guarantees for β-mixing heavy-tailed time series. *The Annals of Statistics*, 48:1124 – 1142.

Xia, D., Zhang, A. R., and Zhou, Y. (2022). Inference for low-rank tensors-no need to debias. *The Annals of Statistics*, 50:1220–1245.

Zheng, Y. (2024). An interpretable and efficient infinite-order vector autoregressive model for high-dimensional time series. *Journal of the American Statistical Association*, pages 1–14.

Zheng, Y. and Cheng, G. (2021). Finite time analysis of vector autoregressive models under linear

restrictions. *Biometrika*, 108:469–489.

Zhu, X., Pan, R., Li, G., Liu, Y., and Wang, H. (2017). Network vector autoregression. *The Annals of Statistics*, 45:1096–1123.

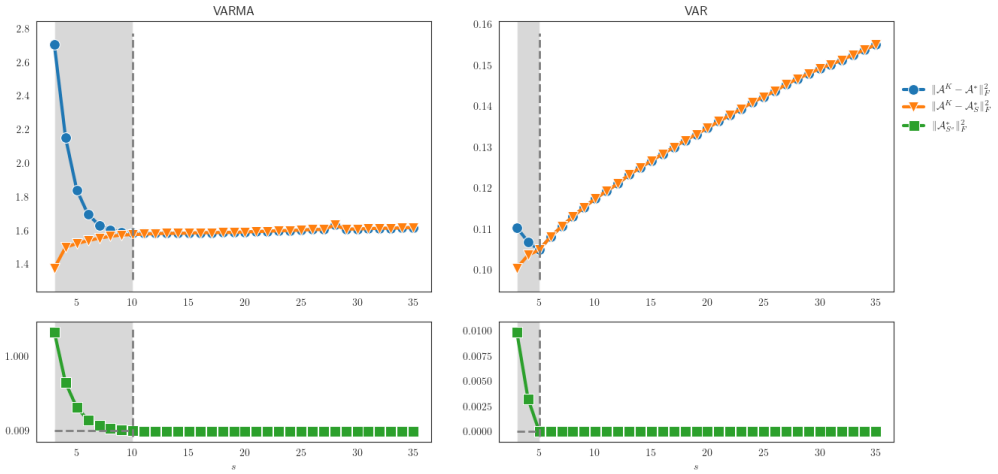

Figure 2: Estimation, approximation and parameter estimation errors, i.e. $\|\boldsymbol{\mathcal{A}}^K - \boldsymbol{\mathcal{A}}^*_S\|_{\mathrm{F}}^2$, $\|\boldsymbol{\mathcal{A}}^*_{S^c}\|_{\mathrm{F}}^2$ and $\|\boldsymbol{\mathcal{A}}^K - \boldsymbol{\mathcal{A}}^*\|_{\mathrm{F}}^2$, at different sparsity levels s for VARMA (left panel) and VAR (right panel) processes. The range of decreasing parameter estimation errors is shaded.

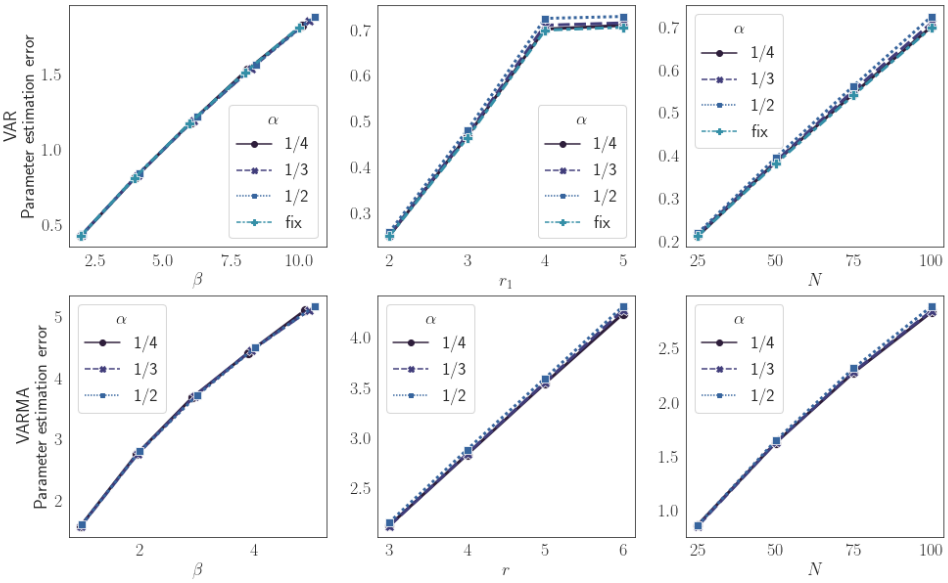

Figure 3: Plots of parameter estimation errors $\|\boldsymbol{\mathcal{A}}^K - \boldsymbol{\mathcal{A}}^*\|_{\mathrm{F}}^2$ against the error rate $\beta = [rN + \log T_0]s/(T - T_0)$ (left panel), rank r (middle panel) and dimension N (right panel), respectively. The data generating processes are VARMA (upper panel) and VAR (lower panel) models, and the running order T_0 is proportional to T^α with "fix" referring to $T_0 = 9$.

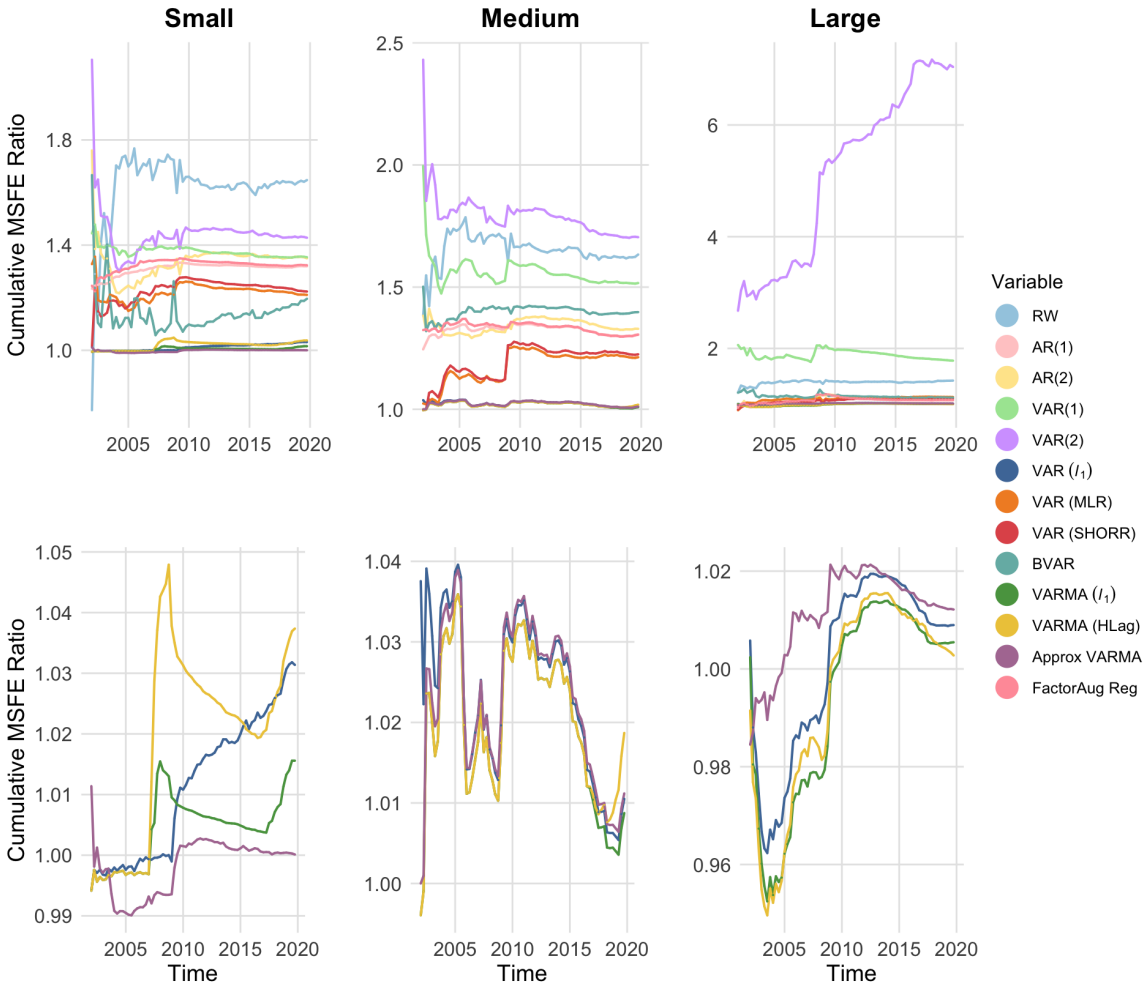

Figure 4: Cumulative MSFE ratios of benchmark methods against ours from 2002 to 2020. In the bottom panel, we plot the cMSFE ratios specifically for the strong benchmark models: VAR (ℓ_1), VARMA (ℓ_1), VARMA (HLag) and Approx VARMA.

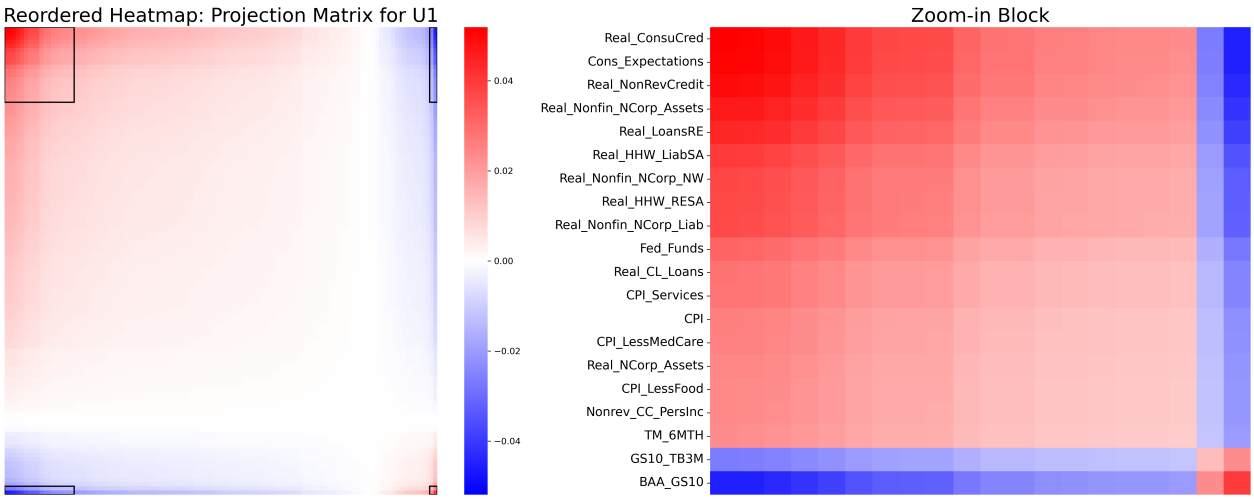

Figure 5: Projection matrices of estimated response loading for the large size dataset, and the selected areas are enlarged at the right panel.

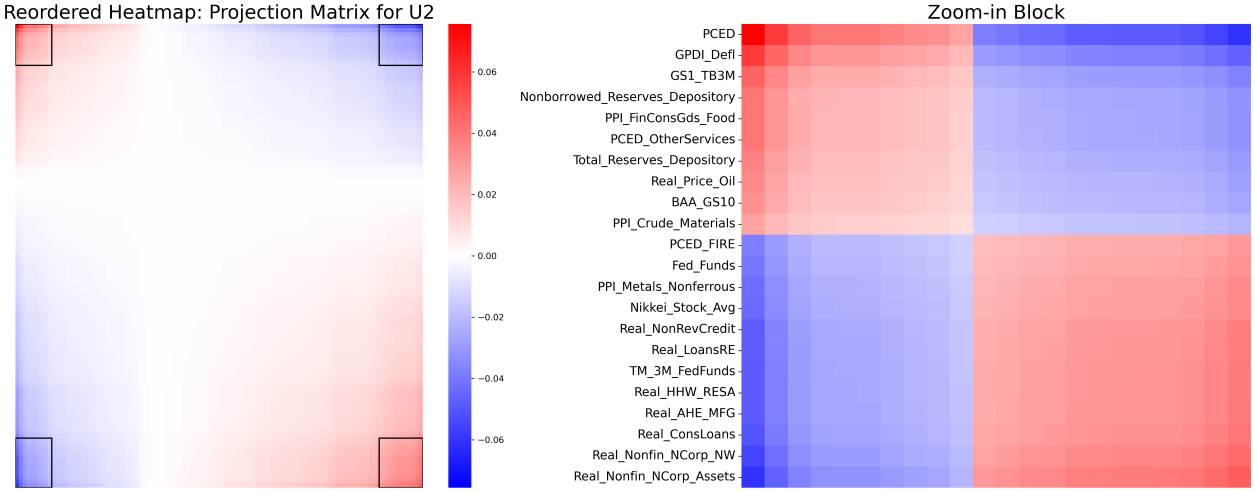

Figure 6: Projection matrices of estimated predictor loading for the large size dataset, and the selected areas are enlarged at the right panel.

Table 1: One-step-ahead mean squared forecast errors ($\text{MSFE}_{\text{last-step}}$) and model confidence set (MCS) p-values of our methods and other ones on the simulated datasets. The best result in each column is highlighted in bold black font.

Models	$\text{MSFE}_{\text{last-step}}$	p_{MCS}
RW	14.00	0.00
AR(1)	10.29	0.23
AR(2)	10.29	0.17
VAR(1)	10.56	0.00
VAR(2)	11.06	0.00
VAR (ℓ_1)	10.42	0.03
VAR (MLR)	10.27	0.64
VAR (SHORR)	10.21	0.64
BVAR	11.40	0.00
VARMA (ℓ_1)	11.55	0.00
VARMA (HLag)	10.95	0.00
Approx VARMA	10.29	0.23
FactorAug Reg	10.24	0.64
Ours	**10.15**	**1.00**

Table 2: One-step-ahead rolling mean squared forecast errors (MSFE) and model confidence set (MCS) p-values of our methods and other ones on small-, medium- and large-size macroeconomic datasets. The best result in each column is highlighted in bold black font.

Datasets\Models	Small		Medium		Large	
	MSFE	p_{MCS}	MSFE	p_{MCS}	MSFE	p_{MCS}
RW	4.83	0.00	10.06	0.00	14.20	0.00
AR(1)	3.87	0.00	8.05	0.00	10.68	0.35
AR(2)	3.96	0.00	8.19	0.00	10.90	0.11
VAR(1)	3.97	0.00	9.34	0.00	17.75	0.00
VAR(2)	4.19	0.00	10.49	0.00	70.18	0.00
VAR (ℓ_1)	3.02	0.10	6.22	0.50	10.06	0.35
VAR (MLR)	3.55	0.01	7.47	0.01	11.27	0.01
VAR (SHORR)	3.59	0.01	7.54	0.01	10.97	0.02
BVAR	3.51	0.01	8.61	0.00	11.17	0.01
VARMA (ℓ_1)	2.98	0.19	6.21	0.62	10.02	0.65
VARMA (HLag)	3.04	0.10	6.27	0.25	10.00	0.84
Approx VARMA	2.93	0.98	6.23	0.43	10.09	0.35
FactorAug Reg	3.88	0.00	8.04	0.00	10.72	0.25
Ours	**2.93**	**1.00**	**6.16**	**1.00**	**9.97**	**1.00**