

# Determination of the effective cointegration rank in high-dimensional time-series predictive regressions

Puyi Fang<sup>1</sup>, Zhaoxing Gao<sup>2</sup>, and Ruey S. Tsay<sup>\*3</sup>

<sup>1</sup>School of Economics, Zhejiang University

<sup>2</sup>Center for Data Science, Zhejiang University

<sup>3</sup>Booth School of Business, University of Chicago

Current Version: April 26, 2023

## Abstract

This paper proposes a new approach to identifying the effective cointegration rank in high-dimensional unit-root (HDUR) time series from a prediction perspective using reduced-rank regression. For a HDUR process  $\mathbf{x}_t \in \mathbb{R}^N$  and a stationary series  $\mathbf{y}_t \in \mathbb{R}^p$  of interest, our goal is to predict future values of  $\mathbf{y}_t$  using  $\mathbf{x}_t$  and lagged values of  $\mathbf{y}_t$ . The proposed framework consists of a two-step estimation procedure. First, the Principal Component Analysis is used to identify all cointegrating vectors of  $\mathbf{x}_t$ . Second, the co-integrated stationary series are used as regressors, together with some lagged variables of  $\mathbf{y}_t$ , to predict  $\mathbf{y}_t$ . The estimated reduced rank is then defined as the effective cointegration rank of  $\mathbf{x}_t$ . Under the scenario that the autoregressive coefficient matrices are sparse (or of low-rank), we apply the Least Absolute Shrinkage and Selection Operator (or the reduced-rank techniques) to estimate the autoregressive coefficients when the dimension involved is high. Theoretical properties of the estimators are established under the assumptions that the dimensions  $p$  and  $N$  and the sample size  $T \rightarrow \infty$ . Both simulated and real examples are used to illustrate the proposed framework, and the empirical application suggests that the proposed procedure fares well in predicting stock returns.

*Keywords:* Cointegration, Factor model, Reduced rank, High dimension, LASSO

# 1 Introduction

The availability of large-scale or vast time-series data in recent years brings new challenges and opportunities to time series modeling. Analysis of high-dimensional (HD) time series has emerged as one of the important and active research areas in statistics, economics, finance, and engineering, among other scientific fields. For example, returns of a large number of assets form a HD time series and play an important role in asset pricing, portfolio allocation, and risk management. Environmental studies often employ HD time series consisting of a large number of pollution indexes collected from many monitoring stations over time. In many applications, data often exhibit characteristics of unit-root nonstationarity. For instance, the series of quarterly gross domestic products, total exports, and total imports of an economy tend to contain unit roots. In theory, the vector autoregressive moving-average (VARMA) models can be used to analyze such data, but they often encounter the difficulties of cointegration testing, overparametrization, and lack of identifiability. See, for example, Johansen (2002), Tiao and Tsay (1989), Lütkepohl (2006) and Tsay (2014), and the references therein. To overcome these difficulties, dimension reduction or structural regularization becomes a necessity, and various methods have been developed in the literature including the regularized estimation method for HD VAR models in Lin and Michailidis (2017) and the factor modeling by Stock and Watson (2005), Bai and Ng (2002), Forni et al. (2005), Peña and Poncela (2006), Lam, Yao and Bathia (2011), Lam and Yao (2012) and Gao and Tsay (2019, 2021b, 2021c, 2022), among others. However, most of the studies mentioned above focus on stationary processes and are not applicable to unit-root nonstationary series. The only exceptions are Bai (2004), Peña and Poncela (2006) and Gao and Tsay (2021c). On the other hand, the unit-root nonstationarity is commonly seen in many empirical applications and the complexity of the dynamical dependence in such data requires further investigation.

It is well known that cointegration is often used to account for common trends and to avoid non-invertibility induced by over-differencing unit-root time series. See Engle and Granger (1987), Johansen (1988, 1991) and Tsay (2014), and the references therein. In practice, the cointegration rank of a given vector time series is unknown, and many approaches have been proposed to estimate the rank; see, for example, Engle and Granger (1987), Johansen (1988, 1991), Saikkonen and Lütkepohl (2000) and Aznar and Salvador (2002). However, these methods are rarely applied to HD time series due to their poor

finite-sample performance, as discussed in Johansen (2002). Yet there are many real applications that involve HD time series. For example, Banerjee, Marcellino and Masten (2014) emphasized the importance of testing for no cross-sectional cointegration in panel cointegration analysis, and the cross-sectional dimension of modern macroeconomic panel can easily be as large as several hundreds. Recently, there are some studies on identifying the cointegration rank of unit-root time series from a factor modeling perspective. See Peña and Poncela (2006) for the case of fixed dimensions and Bai (2004), Zhang, Robinson and Yao (2019) and Gao and Tsay (2021c) for HD time series. However, the situation changes in the case of growing dimension because the estimated cointegration rank usually grows as the dimension increases and the cointegration relationships are often hard to interpret when there are many cointegrating vectors.

This paper marks a further development in estimating the cointegration rank of HDUR time series from a predictive perspective. To avoid employing a large number of cointegrating vectors given by a high-dimensional method, we estimate the *effective cointegration rank* in a predictive framework. Specifically, suppose our goal is to predict the future values of a HD stationary time series  $\mathbf{y} \in \mathbb{R}^p$  using  $\mathbf{x}$  as predictors. It is well known that only the cointegrated series have potential predictive power for the stationary process  $\mathbf{y}$ . If the number of cointegrating vectors is large, the stacked variables obtained by cointegrating vectors form a HD stationary time series, and can be used as potential predictors. But not all cointegrated series have predictive power for  $\mathbf{y}$ , and we define the *effective cointegration rank* as the effective dimension of the stacked variables that have predictive power for  $\mathbf{y}$ . The resulting effective rank can be much smaller than the cointegration rank of  $\mathbf{x}$ .

The proposed method consists of a two-step estimation procedure. First, we postulate that the HDUR time series follows a factor model as that specified in Bai (2004), where the common factors capture the nonstationary common trends of all the components, and the idiosyncratic term is a stationary process. We apply the Principal Component Analysis (PCA) to estimate the common stochastic trends and their associated loading matrix, and the orthogonal complement of the loading matrix consists of the cointegrating vectors. Second, we put together all stationary series obtained by the cointegrating vectors of the first step to form a set of predictors, and perform a reduced-rank regression between the  $\mathbf{y}_t$  series of interest and the predictors. To further explain the variability of the data, we also include some lagged variables of  $\mathbf{y}_t$  in the regression and assume their coefficient

matrices are of low-dimensional structures. We propose two procedures to estimate all the coefficient matrices depending on whether the autoregressive (AR) matrices are sparse or of low-rank. When the AR coefficient matrices are sparse, we apply the nuclear norm penalty to the regression coefficient matrix of the stationary predictors obtained from the first step, and the LASSO penalty to the coefficients of the lagged variables. When both the AR matrices and the coefficient matrix of the predictors are of low-rank, we propose an integrative reduced-rank approach to estimate all unknown parameters. Two iterative, alternating procedures are proposed to estimate all unknown coefficients under the two aforementioned scenarios. Theoretical properties of the estimators are established under the assumption that the dimensions  $p$  and  $N$  and the sample size  $T \rightarrow \infty$ . Both simulated and real examples are used to illustrate the proposed procedure. The empirical application suggests that the 13 macroeconomic variables from Welch and Goyal (2008) provide satisfactory performance as predictors in forecasting the returns of 79 stocks in the S&P 500 index.

The idea of using predictive regression to estimate the cointegrating vector can be found in, for example, Koo et al. (2020). However, the method of Koo et al. (2020) only identifies one cointegrating vector in predicting another univariate time series, whereas the proposed method not only recovers the total cointegration rank, but also identifies the effective cointegration rank in predicting a large panel of time series. In addition, the proposed estimation method is different from theirs as we use PCA, reduced-rank, and LASSO techniques to achieve our goals while they focus mainly on the use of LASSO regularization. Note also that our framework is established for data with time series dependence structure and we use a combination of reduced-rank and sparsity techniques in the estimation procedure, which is different from most of the methods discussed in Reinsel, Velu and Chen (2022+) that focus on the reduced-rank techniques for *i.i.d.* observations. The only exception is the work of Lin and Michailidis (2017) in studying regularized estimation for multi-block stationary VAR models using the tools and techniques developed in Negahban and Wainwright (2011) and Agarwal, Negahban and Wainwright (2012). Furthermore, none of the work mentioned above deals with HDUR time series data.

This paper makes multiple contributions. First, the cointegration problem has been a central issue in modeling HDUR time series, but the lack of clear interpretations of a large number of cointegrating vectors renders the existing HD methods less appealing. We define the effective cointegration rank to select the most significant cointegration

relationships from a predictive point of view using reduced-rank method. This method often produces a small number of significant cointegrating vectors which are easier to interpret in general. Second, our predictive regression model consists of both nonstationary and stationary variables as predictors and has a wide range of applications including the prediction of stock returns using macroeconomic series in finance and the prediction of PM<sub>2.5</sub> values using other air pollution and meteorological indexes in environmental studies. Third, the proposed approach combines the advantages of using two regularization methods, reduced-rank and LASSO, to reduce the dimension of a large system, and the asymptotic results derived suggest that properties of both methods continue to hold when they are used simultaneously in a regression model with serially dependent data. This is a theoretical contribution.

This paper is organized as follows. We introduce the proposed model, estimation methodology, and the modeling procedure in Section 2. Section 3 is devoted to theoretical properties of the proposed model and its associated estimates, and Section 4 presents some simulation results to demonstrate the performance of the proposed method in finite samples. In Section 5, we apply the proposed method to the prediction of stock returns using some commonly used macroeconomic predictors. Section 6 provides some discussions and concluding remarks. All technical proofs of the theorems are relegated to an online supplement.

*Notation.* To begin, we summarize here the notation used throughout the paper. The bold upper case, bold lower case, and lower case letters are used to denote matrices, vectors, and scalars, respectively. For a matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$ , we use  $\|\mathbf{A}\|_F$ ,  $\|\mathbf{A}\|_*$ , and  $\|\mathbf{A}\|_2$  to denote its Frobenius, nuclear, and operator norms, that is,  $\sqrt{\text{tr}(\mathbf{A}'\mathbf{A})}$ , the sum of singular values of  $\mathbf{A}$ , and the largest singular value of  $\mathbf{A}$ , respectively.  $\mathbf{I}_p$  denotes the  $p \times p$  identity matrix. The superscript  $'$  denotes the transpose of a vector or a matrix. For a matrix  $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n)$ , we use  $\text{vec}(\mathbf{A})$  to denote its vectorization, which is equal to  $(\mathbf{a}'_1, \mathbf{a}'_2, \dots, \mathbf{a}'_n)'$ , and we further use  $\|\text{vec}(\mathbf{A})\|_1 = \sum_{i,j} |a_{ij}|$  to denote the  $l_1$ -norm of  $\mathbf{A} = [a_{ij}]$ . Finally, for two matrices  $\mathbf{A}$  and  $\mathbf{B}$  with commensurate dimensions, their inner product is defined as  $\langle \mathbf{A}, \mathbf{B} \rangle = \text{tr}(\mathbf{A}'\mathbf{B})$ . We also use the notation  $a \asymp b$  to denote  $a = O(b)$  and  $b = O(a)$ . Finally, we use  $L(\cdot)$  to denote the lag operator, which can shift a scalar, vector or matrix time series back by one time period. For instance, for the matrix  $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_T)$ ,  $L(\mathbf{Y}) = (\mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_{T-1})$ .

## 2 The Model and Methodology

### 2.1 Model Setting

Let  $\mathbf{y}_t = (y_{1t}, y_{2t}, \dots, y_{pt})'$  be an observable  $p$ -dimensional stationary time series, and  $\mathbf{x}_t = (x_{1t}, x_{2t}, \dots, x_{Nt})'$  an observable  $N$ -dimensional  $I(1)$  process. We consider the following predictive regression model:

$$\mathbf{y}_t = \mathbf{W}\mathbf{x}_{t-1} + \Phi_1\mathbf{y}_{t-1} + \Phi_2\mathbf{y}_{t-2} + \dots + \Phi_d\mathbf{y}_{t-d} + \mathbf{e}_t, \quad t = 1, \dots, T, \quad (2.1)$$

where  $\mathbf{W}$  is a  $p \times N$  coefficient matrix associated with the  $I(1)$  process  $\mathbf{x}_t$ , and  $\Phi_i$  is the  $p \times p$  coefficient matrix of  $\mathbf{y}_{t-i}$ , for  $1 \leq i \leq d$ , and  $\mathbf{e}_t \sim \text{WN}(0, \Sigma_e)$  is a white noise error term with mean zero and a nonsingular covariance  $\Sigma_e$ . Our goal is to estimate  $\mathbf{W}$  and  $\Phi_i$  based on a given sample, and to forecast future values of  $\mathbf{y}_t$ . For simplicity, all variables are set to zero if the time index is not positive. Also, Model (2.1) can be extended to multi-step ahead predictions for  $\mathbf{y}_{t+h}$  with  $h > 0$ .

In Model (2.1),  $\mathbf{x}_t$  is nonstationary but all other variables are stationary so that it only makes sense if some variables in  $\mathbf{x}_t$  are cointegrated, otherwise,  $\mathbf{W}$  would essentially be a zero matrix because the correlation between a stationary process and a unit-root nonstationary one is zero in general. If we blindly apply the Least Squares (LS) method to estimate the model, the number of parameters to be estimated is large, and the resulting estimator  $\widehat{\mathbf{W}}$  would be hard to interpret as we do not know whether all or only a few rows in  $\widehat{\mathbf{W}}$  are the estimated cointegrating vectors. In theory, if all the cointegrating vectors of  $\mathbf{x}_t$  are known, then the resulting linear combinations of the  $I(1)$  variables are stationary and can be useful predictors in Model (2.1). However, not all cointegration relationships are helpful in predicting  $\mathbf{y}_{t+h}$  in general, especially when the dimension  $N$  of  $\mathbf{x}_t$  is large.

In view of the above discussion, we modify Model (2.1) as follows. First, similarly to the setting in Bai (2004), we assume that  $\mathbf{x}_t$  admits a latent factor structure:

$$\mathbf{x}_t = \mathbf{B}\mathbf{f}_t + \boldsymbol{\varepsilon}_t, \quad t = 1, 2, \dots, T, \quad (2.2)$$

where  $\mathbf{f}_t = (f_{1t}, f_{2t}, \dots, f_{rt})'$  is an  $r$ -dimensional factor process that constitutes the common stochastic trends of  $\mathbf{x}_t$ , that is,

$$\mathbf{f}_t = \mathbf{f}_{t-1} + \mathbf{u}_t, \quad (2.3)$$

where  $\mathbf{u}_t$  is an  $r$ -dimensional zero-mean stationary process that drives  $\mathbf{f}_t$ . The idiosyncratic term  $\boldsymbol{\varepsilon}_t$  in (2.2) is assumed to be a stationary process independent of the common factors  $\mathbf{f}_t$ . Therefore, the cointegration rank of  $\mathbf{x}_t$  is  $N - r$ . For ease in model identification, we assume that  $\mathbf{B}$  is an orthonormal matrix such that  $\mathbf{B}'\mathbf{B} = \mathbf{I}_r$ ; see also Bai and Ng (2002) and Fan, Liao and Mincheva (2013) for details.

Let  $\mathbf{B}_c \in \mathbb{R}^{N \times (N-r)}$  be an orthogonal complement matrix of  $\mathbf{B}$  such that  $\mathbf{B}'_c\mathbf{B}_c = \mathbf{I}_{N-r}$  and  $\mathbf{B}'_c\mathbf{B} = \mathbf{0}$ . It follows from Model (2.2) that the columns of  $\mathbf{B}_c$  can be treated as a set of cointegrating vectors of  $\mathbf{x}_t$  because  $\mathbf{B}'_c\mathbf{x}_t = \mathbf{B}'_c\boldsymbol{\varepsilon}_t$  is stationary. Letting

$$\mathbf{z}_t = \mathbf{B}'_c\mathbf{x}_t = \mathbf{B}'_c\boldsymbol{\varepsilon}_t, \quad (2.4)$$

we define  $\mathbf{W} = \mathbf{A}\mathbf{B}'_c$  and rewrite Model (2.1) as follows:

$$\mathbf{y}_t = \mathbf{A}\mathbf{z}_{t-1} + \boldsymbol{\Phi}_1\mathbf{y}_{t-1} + \boldsymbol{\Phi}_2\mathbf{y}_{t-2} + \cdots + \boldsymbol{\Phi}_d\mathbf{y}_{t-d} + \mathbf{e}_t, \quad t = 1, \dots, T, \quad (2.5)$$

where  $\mathbf{z}_t$  is now a stationary process defined in (2.4). Similarly to the identifiability issue in factor models,  $\mathbf{A}$  and  $\mathbf{B}_c$  are not uniquely defined. Nonetheless, the product,  $\mathbf{W} = \mathbf{A}\mathbf{B}'_c$ , is uniquely defined. Therefore, we split Model (2.1) into (2.2)–(2.5), and our goal is to estimate the factor loading matrix  $\mathbf{B}$  or equivalently the cointegrating vector matrix  $\mathbf{B}_c$ , the coefficient matrices  $\mathbf{A}$  and  $\boldsymbol{\Phi}_i$ , for  $1 \leq i \leq d$ .

Although  $\mathbf{B}$ ,  $\mathbf{B}_c$  and  $\mathbf{A}$  are not uniquely defined due to the identification issue, the linear spaces spanned by the columns of  $\mathbf{B}$  and  $\mathbf{B}_c$ , denoted as  $\mathcal{M}(\mathbf{B})$  and  $\mathcal{M}(\mathbf{B}_c)$  respectively, are uniquely defined. For any specific choice of  $\mathbf{B}_c$ ,  $\mathbf{A}$  can also be uniquely determined. Therefore, when we mention the estimation or consistency of the loading matrix  $\mathbf{B}$  or  $\mathbf{B}_c$  in the sequel, we always refer to their column spaces to avoid any confusion. The estimation of  $\mathbf{A}$  is also based on a given and fixed  $\mathbf{B}_c$  so that the procedure is valid.

## 2.2 Estimation Methodology

We consider two approaches to estimating the effective cointegration rank, or equivalently, the reduced-rank of the coefficient matrix  $\mathbf{A}$ , and the AR coefficients  $\boldsymbol{\Phi}_i$ 's for high-dimensional cases under different assumptions. The first approach is based on imposing a reduced-rank structure on the matrix  $\mathbf{A}$  and some sparsity assumptions on the AR coefficient matrices. The second approach requires that all predictors, including the lagged

variables, have their own low-rank coefficient matrices.

### 2.2.1 A Reduced-Rank and Sparse Regression Approach

In this section, we introduce a Reduced-Rank and Sparse Regression approach (RRSRA) to estimating the coefficient matrices  $\mathbf{A}$  and  $\Phi_i$  for observed data  $\{\mathbf{x}_1, \dots, \mathbf{x}_T\}$  and  $\{\mathbf{y}_1, \dots, \mathbf{y}_T\}$ . Note that the dimensions of  $\mathbf{A} \in \mathbb{R}^{p \times (N-r)}$  and  $\Phi_i \in \mathbb{R}^{p \times p}$  can be very large under the assumption that the number of common stochastic trends  $r$  is finite as  $p, N \rightarrow \infty$ . Even if  $\{\mathbf{z}_1, \dots, \mathbf{z}_T\}$  were given, the traditional LS method would lead to overfitting because there are many parameters to estimate. Therefore, some structure regularization must be imposed on the coefficient matrices. For simplicity, we assume the matrix  $\mathbf{A}$  is singular and has a reduced-rank form with  $r_{\mathbf{A}} = \text{rank}(\mathbf{A}) \ll \min(p, N - r)$ , and the AR coefficient matrices  $\Phi_i$ 's are sparse in the sense that only a small number of elements in each matrix are nonzero, for  $1 \leq i \leq d$ .

Assume that the number of common stochastic trends  $r$  in Model (2.2) and the order  $d \geq 1$  in Model (2.5) are known. Their selections will be discussed below. Note that  $\mathbf{z}_t$  is unobservable in Model (2.5) and needs to be estimated from the data  $\mathbf{x}_t$ . We briefly introduce the proposed two-step estimation procedure. First, similarly to that in Bai (2004), we estimate the factor loading matrix  $\mathbf{B}$  by solving the following optimization problem:

$$(\hat{\mathbf{B}}, \hat{\mathbf{F}}) = \arg \min_{\mathbf{B}, \mathbf{F}} \|\mathbf{X} - \mathbf{B}\mathbf{F}\|_{\mathbb{F}}^2, \quad \text{subject to } \mathbf{B}'\mathbf{B} = \mathbf{I}_r, \quad (2.6)$$

where  $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T]$  and  $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_T]$  are the stacked matrices across the time horizon. It is not hard to show that the optimization method in (2.6) is equivalent to Principal Component estimation, and the columns of  $\hat{\mathbf{B}}$  are just the  $r$  standardized eigenvectors of  $\mathbf{X}\mathbf{X}'$  associated with the  $r$  largest eigenvalues. Therefore, we choose  $\hat{\mathbf{B}}_c$  such that its columns are the  $N - r$  standardized eigenvectors associated with the  $N - r$  smallest eigenvalues of  $\mathbf{X}\mathbf{X}'$ . Then, we define  $\hat{\mathbf{z}}_t = \hat{\mathbf{B}}_c' \mathbf{x}_t$ , which serves as a proxy of  $\mathbf{z}_t$  and will be used as predictors in the second step of estimation.

Next, we introduce a method to estimate the coefficient matrices  $\mathbf{A}$  and  $\Phi_i$ , for  $1 \leq i \leq d$ . To begin, define  $\Phi = [\Phi_1, \dots, \Phi_d] \in \mathbb{R}^{p \times dp}$  and  $\mathbf{P}_{t-1} = (\mathbf{y}'_{t-1}, \dots, \mathbf{y}'_{t-d})'$ . For any given penalty parameters  $\lambda_{\mathbf{A}} > 0$  and  $\lambda_{\Phi} > 0$ , we solve the following optimization

problem:

$$(\hat{\mathbf{A}}, \hat{\mathbf{\Phi}}) = \arg \min_{\mathbf{A}, \mathbf{\Phi}} \left\{ \frac{1}{2T} \sum_{t=1}^T \|\mathbf{y}_t - \mathbf{A}\hat{\mathbf{z}}_{t-1} - \mathbf{\Phi}\mathbf{P}_{t-1}\|_2^2 + \lambda_{\mathbf{A}} \|\mathbf{A}\|_* + \lambda_{\mathbf{\Phi}} \|\text{vec}(\mathbf{\Phi})\|_1 \right\}, \quad (2.7)$$

where the data are set to  $\mathbf{0}$  if the subscript  $t \leq 0$ . For reduced-rank regression, we refer the readers to the new monograph by Reinsel, Velu and Chen (2022+). In particular, its Chapters 9 to 12 discuss some recent developments in reduced-rank regressions under high-dimensional settings, including the use of nuclear-norm penalty in (2.7). Similar ideas can also be found in Negahban and Wainwright (2011) and Chen, Dong and Chan (2013), among others. However, most of the methods considered in the aforementioned literature only deal with i.i.d. data, while we consider serially dependent data in this paper both theoretically and empirically.

It is generally not easy to obtain the true global solutions to the optimization problem in (2.7) because the objective function in the bracket of (2.7) involves different types of penalties. Therefore, we formulate an iterative procedure to obtain an approximate set of numerical solutions to (2.7) in Algorithm 1. Specifically, for a fixed  $\mathbf{A}$ , we can estimate  $\mathbf{\Phi}$  via a standard LASSO procedure, and there are several methods and software packages available to obtain sparse solutions. See, for example, Hastie, Tibshirani and Wainwright (2015). When  $\mathbf{\Phi}$  is fixed, the estimation of  $\mathbf{A}$  is an instance of a *semidefinite program*. See Vandenberghe and Boyd (1996) and Ji and Ye (2009). Since the objective function is convex, it is also biconvex in both sets of parameters. If the estimates in all iterations lie within a small ball around the true parameters, the convergence of the estimates to a stationary point is guaranteed. Because the function is convex, the estimates also achieve a global minimum. See, for example, Tseng (2001) and Burai (2013). The theoretical results in Section 3 below are developed for the optimal solutions  $\hat{\mathbf{A}}$  and  $\hat{\mathbf{\Phi}}$ . The simulation results in Section 4 suggest that the initial values in Algorithm 1 have little impact on the asymptotic behavior of the estimates.

Next we turn to the interpretation of the low-rank structure of the matrix  $\mathbf{A}$  in Model (2.5). From Negahban and Wainwright (2011), we see that the estimation of the rank of  $\mathbf{A}$  is equivalent to an optimal selection of the penalty parameter  $\lambda_{\mathbf{A}}$ . A similar argument applies to the sparsity of  $\mathbf{\Phi}$  and the choices of  $\lambda_{\mathbf{\Phi}}$ . See also, Hastie, Tibshirani and Wainwright (2015). Suppose the true rank of  $\mathbf{A}$  is equal to  $k_0 \ll \min\{p, N - r\}$ , we may decompose  $\mathbf{A}$  as  $\mathbf{A} = \mathbf{C}\mathbf{R}'$  with  $\mathbf{C} \in \mathbb{R}^{p \times k_0}$  and  $\mathbf{R} \in \mathbb{R}^{(N-r) \times k_0}$ . Therefore,  $\mathbf{R}'\mathbf{z}_t$  is a

---

**Algorithm 1** An Iterative Procedure for Estimating  $\mathbf{A}$  and  $\Phi$ 

---

**Input:** the data matrices  $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_T]$  and  $\widehat{\mathbf{Z}} = [\widehat{\mathbf{z}}_0, \dots, \widehat{\mathbf{z}}_{T-1}]$

**Output:**  $\widehat{\mathbf{A}} \leftarrow \widehat{\mathbf{A}}^{(k)}$ ,  $\widehat{\Phi} \leftarrow \widehat{\Phi}^{(k)}$

- 1: Initialize with  $k = 0$  and  $\Phi^{(0)} = \mathbf{0}_{p \times p}$
  - 2: **while**  $\widehat{\mathbf{A}}^{(k)}$  or  $\widehat{\Phi}^{(k)}$  is not convergent **do**
  - 3:      $\widehat{\mathbf{A}}^{(k+1)} \leftarrow \arg \min_{\mathbf{A}} \frac{1}{2T} \sum_{t=1}^T \|\mathbf{y}_t - \mathbf{A}\widehat{\mathbf{z}}_{t-1} - \widehat{\Phi}^{(k)}\mathbf{P}_{t-1}\|_2^2 + \lambda_{\mathbf{A}}\|\mathbf{A}\|_*$
  - 4:      $\widehat{\Phi}^{(k+1)} \leftarrow \arg \min_{\Phi} \frac{1}{2T} \sum_{t=1}^T \|\mathbf{y}_t - \widehat{\mathbf{A}}^{(k)}\widehat{\mathbf{z}}_{t-1} - \Phi\mathbf{P}_{t-1}\|_2^2 + \lambda_{\Phi}\|\text{vec}(\Phi)\|_1$
  - 5:      $k \leftarrow k + 1$
  - 6: **end while**
- 

$k_0$ -dimensional stationary random vector and has some predictive power for the future values of  $\mathbf{y}_t$ . Note that  $\mathbf{R}'\mathbf{z}_t = \mathbf{R}'\mathbf{B}'_c\mathbf{x}_t$ , implying that  $\mathbf{R}'\mathbf{B}'_c$  is the reduced-rank matrix consisting of  $k_0$  significant cointegrating vectors that play important roles in predicting the future values of  $\mathbf{y}_t$ . In other words, the cointegration rank  $N - r$  can be reduced to a smaller number  $k_0$  which is useful in prediction and is also easier to interpret. We call  $k_0$  the *effective cointegration rank* of such a prediction application.

### 2.2.2 An Integrative Reduced-Rank Approach

In this section, we introduce an integrative reduced-rank approach (IRRA) to estimating all the coefficient matrices of Model (2.1). The approach is similar to the setting in Chapter 10 of Reinsel, Velu and Chen (2022+) for i.i.d. observations, but we focus on Model (2.5) with time-series dependence. Specifically, under Models (2.2)–(2.5), the IRRA assumes that each set of predictors has its own low-rank coefficient matrix, that is, in addition to the assumption in Section 2.2.1 that  $r_{\mathbf{A}} \ll \min(p, N - r)$ , we also assume that  $0 \leq r_i = \text{rank}(\Phi_i) \ll p$ , for  $1 \leq i \leq d$ , when both  $p$  and  $N$  are large. This approach bridges the reduced-rank and the sparse models in the sense that the coefficient matrix  $\Phi_i$  is fully sparse with all entries being zero if  $r_i = 0$ . On the other hand, the groupwise low-rank structure in IRRA is more flexible and different from a globally low-rank structure for  $\Phi$  defined in Section 2.2.1. The low-rankness of  $\Phi_i$ 's does not necessarily imply that  $\Phi$  is of low rank, while a low-rank matrix  $\Phi$  implies that each  $\Phi_i$  is of low-rank, which cannot exceed that of  $\Phi$ . Under this assumption, we consider the following convex optimization problem:

$$(\widehat{\mathbf{A}}, \widehat{\Phi}_i) = \arg \min_{\mathbf{A}, \Phi_i} \left\{ \frac{1}{2T} \sum_{t=1}^T \|\mathbf{y}_t - \mathbf{A}\widehat{\mathbf{z}}_{t-1} - \Phi\mathbf{P}_{t-1}\|_2^2 + \lambda_{\mathbf{A}}\|\mathbf{A}\|_* + \sum_{i=1}^d \lambda_i \|\Phi_i\|_* \right\}, \quad (2.8)$$

where  $\lambda_{\mathbf{A}}$  and  $\lambda_i$  are the penalty parameters associated with  $\mathbf{A}$  and  $\Phi_i$ , respectively. Similarly to the setting of Reinsel, Velu and Chen (2022+), we may rewrite  $\lambda_i$  as  $\lambda_i = \lambda_{\Phi} w_i$  for a global penalty  $\lambda_{\Phi}$  and some prescribed constant  $w_i$ , for  $1 \leq i \leq d$ . It is clear that  $\lambda_i$  is a tuning parameter controlling the amount of regularization applied to  $\Phi_i$ . If  $w_i = 1$ , and hence  $\lambda_1 = \dots = \lambda_d$ , all penalty parameters of  $\Phi_i$  are the same. A simple choice is to take

$$w_i = \sigma_1(\mathbf{Y}) \{ \sqrt{p} + \sqrt{\text{rank}(\mathbf{Y})} \} / T, \quad i = 1, \dots, d,$$

so that we only have a single parameter  $\lambda_{\Phi}$  to control the regularization of the coefficient  $\Phi_i$ , for  $1 \leq i \leq d$ .

Because the objective function in (2.8) is convex, there are several feasible algorithms available to solve the optimization problem therein. For example, following the recipe in Boyd et al. (2011), Li, Liu and Chen (2019) proposed an Alternating Direction Method of Multipliers (ADMM) algorithm to fit a model similar to that in (2.5) with reduced-rank structures. However, the ADMM algorithm is relatively more involved as it alternates between a primal step and a dual step. In this paper, we propose an easy-to-implement iterative procedure to estimate all the coefficient matrices with reduced-rank structures, which is similar to the block coordinate descent method in Tseng (2001). The detailed procedure is outlined in Algorithm 2 below. Since the objective function is convex, by the argument in Tseng (2001), the convergence of the estimators via Algorithm 2 to a stationary point is guaranteed. On the other hand, from Boyd and Vandenberghe (2004), we know that the conjugate of a conjugate function of a convex one is itself, by Theorem 2 of Burai (2013), the stationary point obtained by Algorithm 2 is a global minimum. Simulation results in Section 4.2 suggest that the estimators obtained by Algorithm 2 are comparable to those obtained by the ADMM method, while the former is much easier to implement than the latter in practice. Similarly to the argument used at the end of Section 2.2.1, the cointegration rank has been reduced to a much smaller and effective one due to the reduced-rank structure of  $\mathbf{A}$ . We omit the details to save space.

### 2.3 Determination of the Number of Factors

The estimation of  $\widehat{\mathbf{B}}$  and its orthogonal complement  $\widehat{\mathbf{B}}_c$  in the prior sections is based on a given  $r$ , which is unknown in practice. There are several methods available in the literature to determine the number of unit-root factors in Equation (2.2). See, for example,

---

**Algorithm 2** Iterative procedure for Estimations of  $\mathbf{A}$  and  $\Phi_i, i = 1, \dots, d$ 


---

**Input:** the data matrices  $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_T]$  and  $\widehat{\mathbf{Z}} = [\widehat{\mathbf{z}}_0, \dots, \widehat{\mathbf{z}}_{T-1}]$

**Output:**  $\widehat{\mathbf{A}} \leftarrow \widehat{\mathbf{A}}^{(k)}, \widehat{\Phi}_i \leftarrow \widehat{\Phi}_i^{(k)}, i = 1, \dots, d$

- 1: Initialize with  $k = 0$  and  $\Phi_i^{(0)} = \mathbf{0}_{p \times p}, i = 1, \dots, d$
  - 2: **while** any of  $\widehat{\mathbf{A}}^{(k)}, \widehat{\Phi}_1^{(k)}, \dots, \widehat{\Phi}_d^{(k)}$  is not convergent **do**
  - 3:      $\widehat{\mathbf{A}}^{(k+1)} \leftarrow \arg \min_{\Phi} \frac{1}{2T} \sum_{t=1}^T \|\mathbf{y}_t - \mathbf{A}\widehat{\mathbf{z}}_{t-1} - \widehat{\Phi}^{(k)}\mathbf{P}_{t-1}\|_2^2 + \lambda_{\mathbf{A}}\|\mathbf{A}\|_*$
  - 4:     **for**  $i = 1$  to  $d$  **do**
  - 5:          $\widehat{\Phi}_i^{(k+1)} \leftarrow \arg \min_{\Phi_i} \frac{1}{2T} \sum_{t=1}^T \|\mathbf{y}_t - \widehat{\mathbf{A}}^{(k)}\widehat{\mathbf{z}}_{t-1} - \widehat{\Phi}^{(k)}\mathbf{P}_{t-1} + (\widehat{\Phi}_i^{(k)} - \Phi_i)\mathbf{y}_{t-i}\|_2^2$   
             $+ \lambda_{\Phi}\|\Phi_i\|_*$
  - 6:     **end for**
  - 7:      $k \leftarrow k + 1$
  - 8: **end while**
- 

the information criterion in Bai (2004), the Canonical Correlation Analysis (CCA) method in Peña and Poncela (2006), the autocorrelation-based method in Zhang, Robinson and Yao (2019) and its modified version in Gao and Tsay (2021c), among others.

In this paper, we adopt the auto-correlation based method of Gao and Tsay (2021c). Specifically, let  $\widehat{\Xi} = (\widehat{\xi}_1, \dots, \widehat{\xi}_N) := [\widehat{\mathbf{B}}, \widehat{\mathbf{B}}_c]$  be the matrix containing all the eigenvectors of  $\mathbf{X}\mathbf{X}'$  and  $\widehat{f}_{j,t} = \widehat{\xi}_j' \mathbf{x}_t$  be the  $j$ -th principal component, for  $1 \leq j \leq N$ . For some prescribed integer  $\bar{k} > 0$ , define

$$S_j(\bar{k}) = \sum_{k=1}^{\bar{k}} |\widehat{\rho}_j(k)|, \quad (2.9)$$

where  $\widehat{\rho}_j(k)$  is the lag- $k$  sample autocorrelation function (ACF) of the principal component  $\widehat{f}_{j,t}$ , for  $1 \leq j \leq N$ . If  $\widehat{f}_{j,t}$  is stationary, then under some mild conditions,  $\widehat{\rho}_j(k)$  decays to zero exponentially as  $k$  increases, and  $\lim_{\bar{k} \rightarrow \infty} S_j(\bar{k}) < \infty$  as  $T \rightarrow \infty$ . If  $\widehat{f}_{j,t}$  is unit-root nonstationary, then  $\widehat{\rho}_j(k) \rightarrow 1$ , and  $\lim_{\bar{k} \rightarrow \infty} S_j(\bar{k}) = \infty$  as  $T \rightarrow \infty$ . Therefore, we start with  $j = 1$ . If the average of the absolute sample ACFs  $S_j(\bar{k})/\bar{k} \geq \delta_0$  for some constant  $0 < \delta_0 < 1$ , then  $\widehat{f}_{j,t}$  has a unit root and we increase  $j$  by 1 to repeat the detecting process. This detecting process is continued until  $S_j(\bar{k})/\bar{k} < \delta_0$  or  $j = N$ . If  $S_j(\bar{k})/\bar{k} \geq \delta_0$  for all  $j$ , then  $\widehat{r} = N$ ; otherwise, we denote  $\widehat{r} = j - 1$ .

## 2.4 Selection of the Tuning Parameters

In this section, we briefly introduce a way to choose the tuning parameters  $\lambda_{\mathbf{A}}$  and  $\lambda_{\Phi}$ , and the order  $d$  in (2.7). We only consider the procedure introduced in Section 2.2.1 since the one in Section 2.2.2 is similar. We first fix the order  $d$  and consider the subsamples  $\{\mathbf{y}_1, \dots, \mathbf{y}_{T_1+j}\}$  and  $\{\mathbf{x}_1, \dots, \mathbf{x}_{T_1+j-1}\}$ , for  $0 \leq j \leq T - T_1 - 1$  and  $T_1 < T$ . We then

adopt a rolling-window-based method to select  $\lambda_{\mathbf{A}}$  and  $\lambda_{\Phi}$  from a forecasting perspective. Specifically, we prescribe two candidate intervals  $[a_1, a_2]$  and  $[b_1, b_2]$  with  $a_2 > a_1 > 0$  and  $b_2 > b_1 > 0$ , and choose  $(\lambda_{\mathbf{A}}, \lambda_{\Phi})$  from  $[a_1, a_2] \times [b_1, b_2]$  via a grid-search approach. For any pair  $(\lambda_{\mathbf{A}}, \lambda_{\Phi}) \in [a_1, a_2] \times [b_1, b_2]$  and each  $0 \leq j \leq T - T_1 - 1$ , we first estimate the loading matrix and obtain the stationary process  $\{\widehat{\mathbf{z}}_1, \dots, \widehat{\mathbf{z}}_{T_1+j-1}\}$  based on the sample  $\{\mathbf{x}_1, \dots, \mathbf{x}_{T_1+j-1}\}$ , and apply the iterative procedure in Algorithm 1 to obtain the estimators for all the coefficients based on the subsample  $\{\mathbf{y}_1, \dots, \mathbf{y}_{T_1+j}\}$ . We can then obtain the predicted value  $\widehat{\mathbf{y}}_{T_1+j+1}$  for  $\mathbf{y}_{T_1+j+1}$ . We repeat the above procedure for  $0 \leq j \leq T - T_1 - 1$  and obtain all the forecasts  $\{\widehat{\mathbf{y}}_{T_1+j+1}, \dots, \widehat{\mathbf{y}}_T\}$ . Define the average of forecast errors as

$$\text{FE}_d(\lambda_{\mathbf{A}}, \lambda_{\Phi}) = \frac{1}{p(T - T_1)} \sum_{j=0}^{T-T_1-1} \|\widehat{\mathbf{y}}_{T_1+j+1} - \mathbf{y}_{T_1+j+1}\|_2^2. \quad (2.10)$$

Note that the forecast errors defined in (2.10) also depend on the value of  $d$ , which itself is unknown in practice. We may prescribe an integer  $\bar{d} > 0$  and search the optimal one over  $0 \leq \widehat{d} \leq \bar{d}$  such that the forecast error is minimized. Consequently, the optimal tuning parameters are chosen as

$$(\widehat{\lambda}_{\mathbf{A}}, \widehat{\lambda}_{\Phi}, \widehat{d}) = \arg \min_{\substack{(\lambda_{\mathbf{A}}, \lambda_{\Phi}) \in [a_1, a_2] \times [b_1, b_2] \\ 0 \leq d \leq \bar{d}}} \text{FE}_d(\lambda_{\mathbf{A}}, \lambda_{\Phi}). \quad (2.11)$$

In practice, for simplicity,  $\bar{d}$  is often chosen as a small integer provided that the series under study is not seasonal. This choice can also be justified theoretically, because the marginal model of a  $p$ -dimensional VAR( $d$ ) process is ARMA( $pd, p(d-1)$ ) the order of which can be sufficiently high when  $p$  is large; see, for instance, Chapter 2 of Tsay (2014). In this paper, we choose  $\bar{d} = 3$  and the proposed model and procedure work sufficiently well in the real data analysis.

### 3 Theoretical Properties

In this section, we investigate some theoretical properties of the coefficient estimates  $\widehat{\mathbf{B}}$ ,  $\widehat{\mathbf{A}}$ , and  $\widehat{\Phi}$  under the condition that  $p, N, T \rightarrow \infty$ . We start with some assumptions and postpone proofs of all theorems to an online supplement.

**Assumption 1.** *The process  $\{\mathbf{u}_t, \boldsymbol{\varepsilon}_t\}$  is  $\alpha$ -mixing with the mixing coefficient satisfying*

the condition  $\alpha(k) \leq \exp(-ck^\gamma)$  for some constants  $c > 0$  and  $\gamma > 0$ , where

$$\alpha(k) = \sup_i \sup_{\substack{A \in \mathcal{F}_{-\infty}^i \\ B \in \mathcal{F}_{i+k}^\infty}} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|,$$

and  $\mathcal{F}_i^j$  is the  $\sigma$ -algebra generated by  $\{(\mathbf{u}_t, \boldsymbol{\varepsilon}_t) : i \leq t \leq j\}$ .

**Assumption 2.**  $\mathbf{u}_t$ ,  $\boldsymbol{\varepsilon}_t$  and  $\mathbf{e}_t$  are sub-exponentially distributed in the sense that there are two constants  $C_1, C_2 > 0$  such that  $\mathbb{P}(|\mathbf{v}'(\boldsymbol{\eta}_t - \mathbb{E}(\boldsymbol{\eta}_t))| > x) \leq C_1 \exp(-C_2 x)$  holds for any  $x > 0$  and  $\|\mathbf{v}\|_2 = 1$ , where  $\boldsymbol{\eta}_t$  can be any process of  $\mathbf{u}_t$ ,  $\boldsymbol{\varepsilon}_t$  or  $\mathbf{e}_t$ .

With the identification condition  $\mathbf{B}'\mathbf{B} = \mathbf{I}_r$ , the processes  $\mathbf{f}_t$  and  $\mathbf{u}_t$  have an additional strength of  $\sqrt{N}$ . For the stationary process  $\mathbf{u}_t$  in (2.3), define a normalized process

$$\mathbf{S}_T^r(\mathbf{t}) = (S_T^1(t_1), \dots, S_T^r(t_r))' = \left( \frac{1}{\sqrt{NT}} \sum_{s=1}^{\lfloor Tt_1 \rfloor} u_{1s}, \dots, \frac{1}{\sqrt{NT}} \sum_{s=1}^{\lfloor Tt_r \rfloor} u_{rs} \right)',$$

where  $\mathbf{t} = (t_1, t_2, \dots, t_r)'$  is a constant vector with  $0 \leq t_1 \leq \dots \leq t_r \leq 1$ .

**Assumption 3.** For any vector  $\mathbf{t} = (t_1, t_2, \dots, t_r)'$  with  $0 \leq t_1 \leq \dots \leq t_r \leq 1$ , there exists a Gaussian process  $\mathbf{W}(\mathbf{t}) = (W_1(t_1), \dots, W_r(t_r))'$  such that  $\mathbf{S}_T^r(\mathbf{t}) \xrightarrow{J_1} \mathbf{W}(\mathbf{t})$  on  $D_r[0, 1]$  as  $T \rightarrow \infty$ , where  $\xrightarrow{J_1}$  denotes weak convergence under the Skorokhod  $J_1$  topology (see Billingsley (1999, Chapter 3)), and  $\mathbf{W}(\mathbf{1})$  has a positive definite covariance matrix.

**Assumption 4.** For any  $i \leq r$ ,  $j \leq N$ , it holds that

$$\frac{1}{T} \sum_{t=1}^T f_{it} \varepsilon_{jt} = O_p(1),$$

uniformly in  $i$  and  $j$ .

**Assumption 5.** For the  $p \times p$  matrix polynomial  $\Phi(L) = \mathbf{I}_p - \sum_{i=1}^d \Phi_i L^i$ , all solutions of the determinant equation  $|\Phi(L)| = 0$  are outside the unit circle.

Assumption 1 is standard for dependent random processes. For a theoretical justification of the mixing conditions for VAR models, see Gao et al. (2019). Assumption 2 implies that all moment conditions for the idiosyncratic terms in Bai (2004) are satisfied. Assumption 3 is used to characterize the limiting behavior of the unit-root factors. Similar assumptions are used in Bai (2004), Zhang, Robinson and Yao (2019), and Gao and Tsay

(2021c), among others. Assumptions 1-3 imply that all conditions for the common factors and the idiosyncratic terms in Bai (2004) hold. Assumption 4 is used to control the sample covariance between the common factors and the idiosyncratic terms. The rate in Assumption 4 is not strong and can be established under the setting of Stock (1987), where we can assume the factors and idiosyncratic terms have similar structure as those in (2.4) therein. Assumption 5 is the standard stationarity condition for a VAR process.

Turn to the convergence of the estimated loading matrix and its orthogonal complement. Note that the loading matrix  $\mathbf{B}$  is not uniquely defined due to the identification issue, only the linear space spanned by its columns, denoted by  $\mathcal{M}(\mathbf{B})$ , or the matrix product  $\mathbf{B}\mathbf{B}'$  is uniquely defined. We state the convergence of the estimated loading matrix and its orthogonal complements in the following theorem.

**Theorem 1.** *Suppose Assumptions 1-4 hold. Assume  $r$  is finite and known. Then, as  $N, T \rightarrow \infty$ ,*

$$\|\widehat{\mathbf{B}}\widehat{\mathbf{B}}' - \mathbf{B}\mathbf{B}'\|_2 = O_p(T^{-1}) \quad \text{and} \quad \|\widehat{\mathbf{B}}_c\widehat{\mathbf{B}}_c' - \mathbf{B}_c\mathbf{B}_c'\|_2 = O_p(T^{-1}). \quad (3.1)$$

Consequently,

$$N^{-1/2}\|\widehat{\mathbf{B}}\widehat{\mathbf{f}}_t - \mathbf{B}\mathbf{f}_t\|_2 = O_p(N^{-1/2} + T^{-1/2}).$$

**Remark 1.** *From Theorem 1, the two distances in (3.1) are of the same rate which is reasonable because we used the matrix perturbation theory in the proofs and the two matrices play symmetric roles in Lemma A1 of the Supplement. The discrepancy measure used in Theorem 1 is equivalent to the  $\sin(\Theta)$  distance in the literature concerning the distance between two orthogonal matrices. See (3.2)–(3.4) of Gao and Tsay (2021a) for details. In addition, based on Gao and Tsay (2021a), the first distance  $\|\widehat{\mathbf{B}}\widehat{\mathbf{B}}' - \mathbf{B}\mathbf{B}'\|_2$  in (3.1) is also equivalent to the measure between two linear spaces defined in Pan and Yao (2008):*

$$D(\mathcal{M}(\mathbf{B}), \mathcal{M}(\widehat{\mathbf{B}})) = \sqrt{1 - \text{tr}(\mathbf{B}\mathbf{B}'\widehat{\mathbf{B}}\widehat{\mathbf{B}}')/r},$$

when  $r$  is finite, but the second distance in (3.1) is not because the dimension of  $\mathbf{B}_c$  is diverging.

The following theorem establishes the convergence of the estimated number of common stochastic trends.

**Theorem 2.** *Suppose Assumptions 1–4 hold. If  $N^{1/2} \log(T)T^{-1/2} \rightarrow 0$ , then  $P(\hat{r} = r) \rightarrow 1$  as  $N, T \rightarrow \infty$ , where  $\hat{r}$  is obtained by the autocorrelation-based method in Section 2.3.*

Next, turn to the convergence of the estimated regression coefficients obtained in Section 2. To control the errors between the estimated coefficients and the true ones, we introduce a Restricted Strong Convexity (RSC) condition which is often used in high-dimensional regularized estimation problems. See Agarwal, Negahban and Wainwright (2012) and Wainwright (2019, Chapter 9) for details. For any given  $\lambda_{\mathbf{A}}, \lambda_{\Phi} > 0$ , and a matrix  $\Delta \in \mathbb{R}^{p \times (N-r+dp)} = [\Delta_1, \Delta_2]$  with  $\Delta_1 \in \mathbb{R}^{p \times (N-r)}$  and  $\Delta_2 \in \mathbb{R}^{p \times dp}$ , we use a weighted combination to define an associated norm as follows:

$$\Psi(\Delta) := \lambda_{\mathbf{A}} \|\Delta_1\|_* + \lambda_{\Phi} \|\text{vec}(\Delta_2)\|_1. \quad (3.2)$$

The restricted strong convexity condition under our setting is defined below.

**Definition 1.** *Consider a generic operator  $\mathcal{X} : \mathbb{R}^{p \times (N-r+dp)} \mapsto \mathbb{R}^{p \times T}$ . We say that it satisfies the RSC condition with respect to norm  $\Psi$ , if*

$$\frac{1}{2T} \|\mathcal{X}(\Delta)\|_{\text{F}}^2 \geq \frac{\kappa_1}{2} \|\Delta\|_{\text{F}}^2 - \tau_T \Psi^2(\Delta), \quad \text{for some } \Delta \in \mathbb{R}^{p \times (N-r+dp)},$$

where  $\kappa_1 > 0$  and  $\tau_T > 0$  are the curvature and tolerance constants, respectively.

When  $\tau_T = 0$ , the RSC condition in Definition 1 is called a locally strong convexity condition. See Wainwright (2019, Chapter 9). Denote  $\Delta = [\Delta_{\mathbf{A}}, \Delta_{\Phi}]$  with  $\Delta_{\mathbf{A}} = \hat{\mathbf{A}} - \mathbf{A}$  and  $\Delta_{\Phi} = \hat{\Phi} - \Phi$ . We now establish the convergence rates of the estimated coefficient matrices below.

**Theorem 3.** *Suppose Assumptions 1–5 hold. For the augmented data matrices  $\mathbf{Z} = [\mathbf{z}_0, \dots, \mathbf{z}_{T-1}]$  and  $\mathbf{P} = [\mathbf{P}_0, \dots, \mathbf{P}_{T-1}]$ , where all variables with zero or negative time indexes are set to 0, if the operator*

$$\mathcal{X}([\Delta_{\mathbf{A}}, \Delta_{\Phi}]) := \Delta_{\mathbf{A}} \mathbf{Z} + \Delta_{\Phi} \mathbf{P}$$

*satisfies the RSC condition with the norm in the form of (3.2), curvature  $\kappa_1$  and tolerance  $\tau_T$  such that*

$$\kappa_1 \geq C \tau_T r_{\mathbf{A}} \lambda_{\mathbf{A}}^2, \quad \text{and} \quad \kappa_1 \geq C \tau_T s_{\Phi} \lambda_{\Phi}^2,$$

where  $r_{\mathbf{A}}$  and  $s_{\Phi}$  are the rank of  $\mathbf{A}$  and the cardinality of the support of  $\Phi$ , respectively, then with the regularization parameters  $\lambda_{\mathbf{A}}$  and  $\lambda_{\Phi}$  satisfying

$$\lambda_{\mathbf{A}} \geq \frac{3}{T} \|\mathbf{E}\mathbf{Z}'\|_2 \quad \text{and} \quad \lambda_{\Phi} \geq \frac{2}{T} \|\text{vec}(\mathbf{E}\mathbf{P}')\|_{\infty},$$

where  $\mathbf{E} = [\mathbf{e}_1, \dots, \mathbf{e}_T]$  is the error matrix of (2.5), we have

$$\|\widehat{\mathbf{A}} - \mathbf{A}\|_{\mathbb{F}}^2 + \|\widehat{\Phi} - \Phi\|_{\mathbb{F}}^2 \leq C \frac{\lambda_{\mathbf{A}}^2 r_{\mathbf{A}} + \lambda_{\Phi}^2 s_{\Phi}}{\kappa_1^2}. \quad (3.3)$$

**Remark 2.** (i) Under Assumptions 1–5, by the Bernstein-type inequality for weakly dependent data in Merlevède, Peligrad and Rio (2011) and the argument in the proofs of Lemma 3 in Negahban and Wainwright (2011), it is not hard to show that  $\|\mathbf{E}\mathbf{Z}'\|_2 = O_p(\sqrt{(p+N)T})$ . Then, the condition for  $\lambda_{\mathbf{A}}$  becomes  $\lambda_{\mathbf{A}} \geq C\sqrt{(p+N)}/T$ . Similarly, by the Bernstein-type inequality in Merlevède, Peligrad and Rio (2011), we can also show that  $\|\text{vec}(\mathbf{E}\mathbf{P}')\|_{\infty} = O_p(\sqrt{T \log(p)})$ , and therefore, the condition for  $\lambda_{\Phi}$  reduces to  $\lambda_{\Phi} \geq C\sqrt{\log(p)}/T$ , which is the same as that in the LASSO literature. See Wainwright (2019).

(ii) For a properly chosen  $C_* > 0$  such that  $\lambda_{\mathbf{A}} = C_*\sqrt{(p+N)}/T$  and  $\lambda_{\Phi} = C_*\sqrt{\log(p)}/T$  satisfy the conditions in Theorem 3, under the setting that  $p/T \rightarrow 0$  and  $N/T \rightarrow 0$ , we may choose an  $\tau_T > 0$  such that  $\kappa_1 > C \max(\tau_T r_{\mathbf{A}} \lambda_{\mathbf{A}}^2, \tau_T s_{\Phi} \lambda_{\Phi}^2) > 0$  is a positive constant, and then it follows from Theorem 3 that

$$\|\widehat{\mathbf{A}} - \mathbf{A}\|_{\mathbb{F}}^2 + \|\widehat{\Phi} - \Phi\|_{\mathbb{F}}^2 \leq C \left( r_{\mathbf{A}} \frac{p+N}{T} + s_{\Phi} \frac{\log(p)}{T} \right) \rightarrow 0,$$

as  $p, N, T \rightarrow \infty$  for finite  $r_{\mathbf{A}}$  and  $s_{\Phi}$ , implying that the estimated coefficient matrices are consistent.

(iii) Under the settings in Remark 2(ii), we immediately obtain the consistencies for both matrices:

$$\|\widehat{\mathbf{A}} - \mathbf{A}\|_{\mathbb{F}}^2 \rightarrow 0 \quad \text{and} \quad \|\widehat{\Phi} - \Phi\|_{\mathbb{F}}^2 \rightarrow 0, \quad \text{as } p, N, T \rightarrow \infty. \quad (3.4)$$

If there is a positive constant  $C > 0$  such that the minimum nonzero singular value of  $\mathbf{A}$  and the minimum absolute elements in  $\Phi$ , denoted by  $\sigma_{r_{\mathbf{A}}}$  and  $|\Phi|_{\min}$  respectively, satisfy  $\sigma_{r_{\mathbf{A}}} > C > 0$  and  $|\Phi|_{\min} > C > 0$  as  $p, N, T \rightarrow \infty$ , (3.4) implies that  $P(\widehat{r}_{\mathbf{A}} = r_{\mathbf{A}}) \rightarrow 1$  and  $P(\widehat{\mathcal{S}}_{\Phi} = \mathcal{S}_{\Phi}) \rightarrow 1$ , where  $\widehat{r}_{\mathbf{A}} = \text{rank}(\widehat{\mathbf{A}})$ ,  $r_{\mathbf{A}} = \text{rank}(\mathbf{A})$ , and  $\widehat{\mathcal{S}}_{\Phi}$  and  $\mathcal{S}_{\Phi}$  contain all

the indexes of the nonzero elements in  $\widehat{\Phi}$  and  $\Phi$ , respectively. We omit the details to save space.

To establish properties of the estimated coefficients using the IRRA of Section 2.2.2, we first introduce a restricted set that is constructed by a projection of any matrix onto a subspace generated by another one of the same shape. Specifically, for any  $m \times n$  matrix  $\Theta$ , we perform a singular value decomposition (SVD)  $\Theta = \mathbf{U}\mathbf{D}\mathbf{V}'$  with a partition as follows,

$$\Theta = \begin{bmatrix} \mathbf{U}_k & \mathbf{U}_{k,c} \end{bmatrix} \begin{bmatrix} \mathbf{D}_k & \\ & \mathbf{D}_{k,c} \end{bmatrix} \begin{bmatrix} \mathbf{V}'_k \\ \mathbf{V}'_{k,c} \end{bmatrix}, \quad (3.5)$$

where  $\mathbf{U}_k \in \mathbb{R}^{m \times k}$  and  $\mathbf{V}_k \in \mathbb{R}^{n \times k}$  are the sub-matrices consisting of the left and right singular vectors associated with the  $k$  largest singular values of  $\Theta$ , respectively, and  $\mathbf{U}_{k,c} \in \mathbb{R}^{m \times (m-k)}$  and  $\mathbf{V}_{k,c} \in \mathbb{R}^{n \times (n-k)}$  are the remaining ones. Similarly to Negahban and Wainwright (2011), we define two subspaces as follows,

$$\begin{aligned} \mathcal{S}_\Theta(k) &= \{\mathbf{A} \in \mathbb{R}^{m \times n} : \text{range}(\mathbf{A}) \subseteq \text{range}(\mathbf{U}_k), \text{range}(\mathbf{A}') \subseteq \text{range}(\mathbf{V}_k)\}, \text{ and} \\ \mathcal{S}_\Theta^\perp(k) &= \{\mathbf{A} \in \mathbb{R}^{m \times n} : \text{range}(\mathbf{A}) \perp \text{range}(\mathbf{U}_k), \text{range}(\mathbf{A}') \perp \text{range}(\mathbf{V}_k)\}. \end{aligned} \quad (3.6)$$

For any matrix  $\mathbf{M} \in \mathbb{R}^{m \times n}$ , we decompose it as  $\mathbf{M} = \mathbf{M}_1 + \mathbf{M}_2$ , where

$$\mathbf{M}_2 = \mathbf{U}_{k,c} \mathbf{U}'_{k,c} \mathbf{M} \mathbf{V}_{k,c} \mathbf{V}'_{k,c}, \text{ and } \mathbf{M}_1 = \mathbf{M} - \mathbf{M}_2. \quad (3.7)$$

Because  $\mathbf{M}_2 \in \mathcal{S}_\Theta^\perp(k)$ , we use  $\Pi_{\mathcal{S}_\Theta^\perp(k)}(\mathbf{M}) = \mathbf{M}_2$  to denote the projection of matrix  $\mathbf{M}$  onto the subspace  $\mathcal{S}_\Theta^\perp(k)$ .

Turn to the estimated coefficients using the IRRA of Section 2.2.2. By an abuse of notation, we define  $\Delta = [\Delta_{\mathbf{A}}, \Delta_{\Phi}] = [\Delta_{\mathbf{A}}, \Delta_{\Phi_1}, \dots, \Delta_{\Phi_d}]$  with  $\Delta_{\mathbf{A}} = \widehat{\mathbf{A}} - \mathbf{A} \in \mathbb{R}^{p \times (N-r)}$  and  $\Delta_{\Phi_i} = \widehat{\Phi}_i - \Phi_i \in \mathbb{R}^{p \times p}$ , and hence  $\Delta_{\Phi} \in \mathbb{R}^{p \times dp}$ . We decompose  $\Delta_{\mathbf{A}}$  as  $\Delta_{\mathbf{A}} = \Delta_{\mathbf{A},1} + \Delta_{\mathbf{A},2}$  and  $\Delta_{\Phi_i}$  as  $\Delta_{\Phi_i} = \Delta_{\Phi_i,1} + \Delta_{\Phi_i,2}$ , for  $1 \leq i \leq d$ . It follows that  $\Delta_{\mathbf{A},2} = \Pi_{\mathcal{S}_{\widehat{\mathbf{A}}}^\perp(r_{\mathbf{A}})}(\Delta_{\mathbf{A}})$ , and  $\Delta_{\Phi_i,2} = \Pi_{\mathcal{S}_{\widehat{\Phi}_i}^\perp(r_i)}(\Delta_{\Phi_i})$ , for  $1 \leq i \leq d$ . We define a restricted set  $\mathcal{C}$  as

$$\begin{aligned} \mathcal{C}(r_1, \dots, r_d) &= \left\{ \Delta \in \mathbb{R}^{p \times (N-r+dp)} \mid \|\Delta_{\mathbf{A},2}\|_* + \sum_{i=1}^d \|\Delta_{\Phi_i,2}\|_* \leq 3\|\Delta_{\mathbf{A},1}\|_* \right. \\ &\quad \left. + 3 \sum_{i=1}^d \|\Delta_{\Phi_i,1}\|_* \right\}. \end{aligned} \quad (3.8)$$

We make an additional assumption below.

**Assumption 6.** For the operator  $\mathcal{X}$  defined in Theorem 3, we assume

$$\frac{1}{2T} \|\mathcal{X}(\Delta)\|_{\mathbb{F}}^2 = \frac{1}{2T} \|\Delta_{\mathbf{A}} \mathbf{Z} + \Delta_{\Phi} \mathbf{P}\|_{\mathbb{F}}^2 \geq \kappa_2 \|\Delta\|_{\mathbb{F}}^2, \text{ for all } \Delta \in \mathcal{C}(r_1, \dots, r_d),$$

where  $\kappa_2 > 0$  is a constant and  $\mathcal{C}(r_1, \dots, r_d)$  is defined in (3.8).

Note that Assumption 6 is a locally restricted strong convexity condition by setting  $\tau_T = 0$  in Definition 1. Similar assumptions are also considered in Chapter 10 of Reinsel, Velu and Chen (2022+) for *i.i.d.* data. We next state the convergence of the estimated coefficients based on the IRRA of Section 2.2.2.

**Theorem 4.** Assume Assumptions 1–5 hold. Suppose the predictor matrices  $\mathbf{Z}$  and  $\mathbf{P}$  satisfy the condition in Assumption 6 over the set  $\mathcal{C}$  defined in (3.8). If  $\lambda_{\mathbf{A}}$  and  $\lambda_i$  satisfy

$$\lambda_{\mathbf{A}} \geq \frac{3}{T} \|\mathbf{E}\mathbf{Z}'\|_2 \quad \text{and} \quad \lambda_i \geq \frac{2}{T} \|\mathbf{E}L^i(\mathbf{Y})'\|_2, \quad \text{for } i = 1, 2, \dots, d,$$

then, as  $p, N, T \rightarrow \infty$ , we have

$$\|\widehat{\mathbf{A}} - \mathbf{A}\|_{\mathbb{F}}^2 + \sum_{i=1}^d \|\widehat{\Phi}_i - \Phi_i\|_{\mathbb{F}}^2 \leq C \left( r_{\mathbf{A}} \lambda_{\mathbf{A}}^2 + \sum_{i=1}^d r_i \lambda_i^2 \right) / \kappa_2^2.$$

**Remark 3.** (i) Assumption 6 can be replaced by a weaker RSC condition as that in Theorem 3, and the results in Theorem 4 continue to hold with minor modifications in the proofs given in the online supplement.

(ii) The convergence rates of the estimated coefficients are the same as those in Chapter 10 of Reinsel, Velu and Chen (2022+), even for time-series data with mild serial dependence.

(iii) By the discussions in Remark 2(i)–(ii), we may also choose  $\lambda_{\mathbf{A}} = C_* \sqrt{(p+N)/T}$  and  $\lambda_i = C_* \sqrt{p/T}$  for some constant  $C_* > 0$  satisfying the conditions in Theorem 4, such that the convergence results in Theorem 4 can be rewritten as

$$\|\widehat{\mathbf{A}} - \mathbf{A}\|_{\mathbb{F}}^2 + \sum_{i=1}^d \|\widehat{\Phi}_i - \Phi_i\|_{\mathbb{F}}^2 \leq C \left\{ \frac{(p+N)r_{\mathbf{A}}}{T} + \sum_{i=1}^d \frac{pr_i}{T} \right\},$$

which approaches zero asymptotically under the setting that  $p/T \rightarrow 0$  and  $N/T \rightarrow 0$ , implying that the estimators are consistent. It is straightforward to see that the convergence

rates above are slightly slower than those in Remark 2(ii) if the sparsity parameter therein satisfies  $s_{\Phi}/p \rightarrow 0$ , which is often the case in sparse regression. This is understandable since there are usually more autoregressive coefficients to estimate in a reduced-rank regression in (2.8) than in the sparse counterpart in (2.7).

## 4 Simulation Study

In this section, we evaluate the finite-sample performance of the proposed methodologies under the scenarios when both  $p$  and  $N$  are increasing from small to large. Though the dimensions of  $\mathbf{B}$  and  $\widehat{\mathbf{B}}$  are not necessarily the same, as estimation error in  $r$  may occur, the discrepancy measure adopted in Theorem 1 remains valid. To simplify the presentation and without loss of generality, we set  $d = 1$  in (2.5), and similar results can also be obtained for other choices of finite  $d$ .

### 4.1 Example 1: The Reduced-Rank and Sparse Regression

#### 4.1.1 Data Generating Process

We follow the data generating process in (2.2) and (2.5) and consider a three-factor model, where the factors are  $I(1)$  processes generated by (2.3). We further multiply the factors by  $\sqrt{N}$  because we will use orthonormal loading matrices below and the strength of general loadings is imposed on the factors in line with the assumptions and identification conditions. While the number of factors  $r = 3$  is fixed, we set  $p = 20, 40, 60$ , and  $N = 20, 40, 60$ , respectively, and in each configuration of  $(p, N)$ , we set the sample size  $T = 400, 800, 1200$  to illustrate the proposed method and to exam certain theoretical properties of the estimators. In order to obtain reproducible results, we initialize a random generator in the NumPy package in Python by setting the seed to 1024, and this seed is used throughout the simulation.

To begin, we need to obtain the coefficient matrices of the model. We start with generating the loading matrix  $\mathbf{B}$  and its corresponding orthogonal complement  $\mathbf{B}_c$ . As  $[\widehat{\mathbf{B}}, \widehat{\mathbf{B}}_c]$  is an  $N \times N$  full-rank orthonormal matrix, we first randomly generate an  $N \times N$  orthogonal matrix, and divide its columns in such a way that the submatrix with the first  $r$  columns is chosen as  $\mathbf{B}$  and the remaining columns form naturally the  $\mathbf{B}_c$  matrix. For the low-rank matrix  $\mathbf{A}$ , we first randomly generate two orthonormal matrices  $\mathbf{U} \in \mathbb{R}^{p \times p}$

and  $\mathbf{V} \in \mathbb{R}^{(N-r) \times (N-r)}$ , and a  $p \times (N-r)$  rectangular diagonal matrix  $\mathbf{D}$  with only five positive entries on the upper left of the diagonal while all the other entries are set to zero. The positive diagonal entries in  $\mathbf{D}$  are drawn independently from a uniform distribution on the interval of  $[0.1, 1)$  so that all the five elements are strictly greater than 0. The matrix  $\mathbf{A}$  with rank  $r_{\mathbf{A}} = 5$  is then chosen as  $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}'$ . Next, for the sparse matrix  $\Phi$ , we first create a sparse matrix  $\Phi_1$  with only 20 randomly located non-zero entries each of which is drawn uniformly on the intervals  $(-1, -0.1] \cup [0.1, 1)$ . In order to guarantee the stationarity of  $\mathbf{y}_t$  in (2.5), we use the normalized matrix  $\Phi = 0.9 \times \Phi_1 / \|\Phi_1\|_2$  as the autoregressive coefficient matrix, which implies that Assumption 5 holds.

For each configuration of  $(p, N, T)$ , with the coefficient matrices  $\mathbf{B}, \mathbf{B}_c, \mathbf{A}$  and  $\Phi$  chosen by the aforementioned methods, we generate  $\mathbf{x}_t, \mathbf{z}_t$  and  $\mathbf{y}_t$  according to Models (2.2), (2.4) and (2.5), respectively. To obtain stable results, we use 500 replications for each  $(p, N, T)$  configuration and set  $\boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \mathbf{I}_N)$ ,  $\mathbf{u}_t \sim N(\mathbf{0}, \mathbf{I}_r)$ , and  $\mathbf{e}_t \sim N(\mathbf{0}, \mathbf{I}_p)$  in each realization.

#### 4.1.2 Performance Evaluation

We first study the performance of (2.9) in estimating the number of factors. Because the data generating process  $\mathbf{x}_t$  of the previous section is independent of the dimension  $p$ , we only illustrate the proposed method for the case of  $p = 20$ , and similar results can also be obtained for other cases. Table 1 reports the empirical probabilities of  $P(\hat{r} = r)$  based on 500 repetitions for each  $(N, T)$  configuration when  $p = 20$ , where we use the method described in Section 2.3 with  $\bar{k} = 10$  and  $\delta_0 = 0.3$ . From Table 1, we see that the auto-correlation based method can successfully recover the number of common stochastic trends. This is understandable because all the factors used in the simulation are strong ones. Similar results can also be found in Bai and Ng (2002) and Lam and Yao (2012).

Table 1: Empirical probabilities of  $P(\hat{r} = r)$  for various  $(N, T)$  configurations, where the value is  $r = 3$  and the dimension is  $p = 20$ . The estimation method of Section 2.3 with  $\bar{k} = 10$  and  $\delta_0 = 0.3$  is used, and the results are based on 500 iterations.

$T$	$N$		
	20	40	60
400	100.00%	100.00%	100.00%
800	100.00%	100.00%	100.00%
1200	100.00%	100.00%	100.00%

Next, we consider the estimation accuracy of the loading matrix  $\mathbf{B}$ , which is measured by  $\|\mathbf{B}\mathbf{B}' - \widehat{\mathbf{B}}\widehat{\mathbf{B}}'\|_2$  over 500 replications. For the same reason mentioned before, we only

show the results for the case of  $p = 20$ . Boxplots of the discrepancies are shown in Figure 1, from which we see that for each  $N$ , the discrepancy between the estimated loading matrix and the true one decreases as the sample size  $T$  increases. This result is in agreement with our theorems. Furthermore, we also evaluate the estimation errors of the extracted factors. For each  $(N, T)$  configuration, we define the the root-mean-squared-error (RMSE) of the estimated factors as

$$RMSE = \left( \frac{1}{NT} \sum_{t=1}^T \|\mathbf{B}\mathbf{f}_t - \widehat{\mathbf{B}}\widehat{\mathbf{f}}_t\|_2^2 \right)^{1/2}, \quad (4.1)$$

which quantifies the accuracy in recovering the common stochastic trends. Figure 2 shows the results via boxplots using 500 replications. From Figure 2, we see clearly that, the recovery errors of the common factors decrease as the sample size  $T$  increases, which is consistent with the theoretical results in Theorem 1.

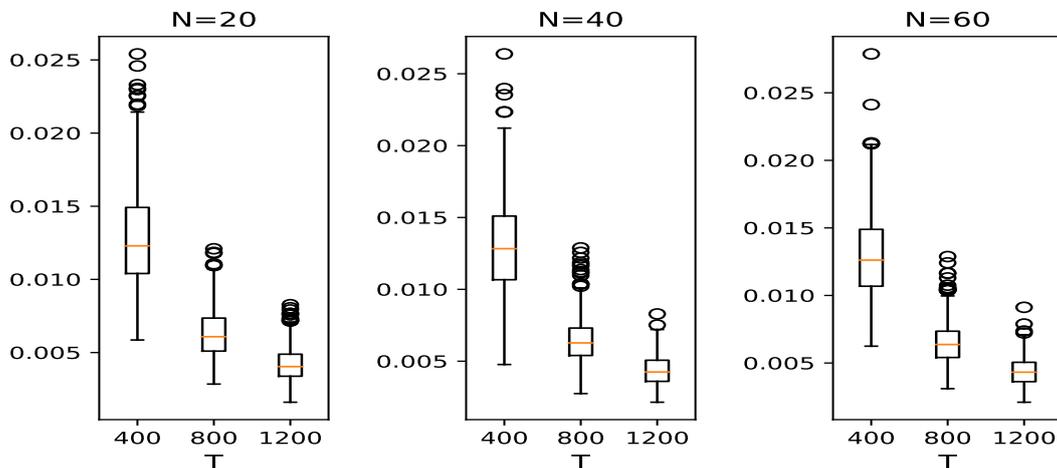


Figure 1: Boxplots of  $\|\mathbf{B}\mathbf{B}' - \widehat{\mathbf{B}}\widehat{\mathbf{B}}'\|_2$  with  $r = 3$  and  $p = 20$  in Example 1. For each  $N$  (dimension of  $\mathbf{x}_t$ ), the sample sizes used are 400, 800 and 1200, respectively. The results are based on 500 replications.

We then study the estimation accuracy of the low-rank matrix  $\mathbf{A}$  and the sparse matrix  $\Phi$  using the procedure in Algorithm 1. For simplicity, we set the tuning parameters  $\lambda_{\mathbf{A}} = \sqrt{(p + N)/T}$  and  $\lambda_{\Phi} = \sqrt{\log(p)/T}$ , which are just taken from the rates discussed in Remark 2(ii) by setting  $C_* = 1$  and this choice is good enough to produce satisfactory performance in the simulation. In practice, we may choose an optimal  $C_*$  from an interval using grid search. Due to the identification issue as that for  $\mathbf{B}$ , we also use  $\|\mathbf{A}\mathbf{A}' - \widehat{\mathbf{A}}\widehat{\mathbf{A}}'\|_2$  to evaluate the discrepancy between  $\widehat{\mathbf{A}}$  and  $\mathbf{A}$ . Because there is no identification issues with  $\Phi$  and the estimated  $\widehat{\Phi}$ , we use  $\|\Phi - \widehat{\Phi}\|_2$  to measure the estimation accuracy of the autoregressive coefficients. Boxplots of the estimation errors for  $\widehat{\mathbf{A}}$  and  $\widehat{\Phi}$  are presented in the Figures 3 and 4, respectively. As expected from Theorem 3, in each case of  $(p, N)$ , the

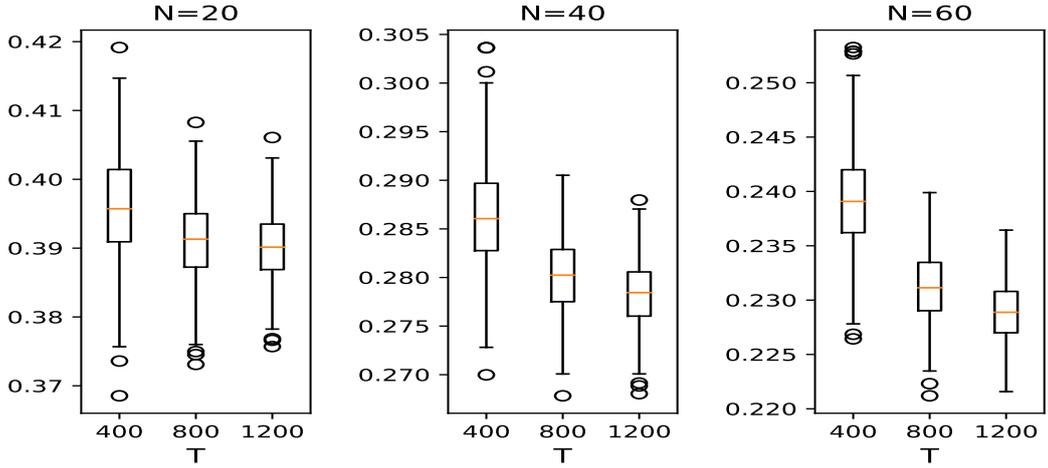


Figure 2: Boxplots for RMSE of the extracted factors defined in (4.1) with  $r = 3$  and  $p = 20$  in Example 1. For each  $N$ , the sample sizes used are 400, 800, and 1200, respectively. The results are based on 500 replications.

estimation errors of  $\mathbf{A}$  and  $\Phi$  both decrease as the sample size  $T$  increases, which is also consistent with our theoretical properties.

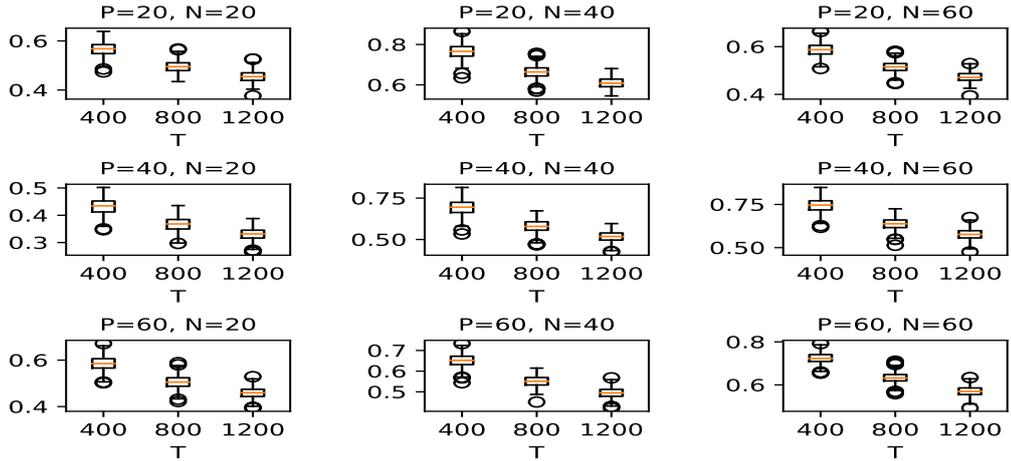


Figure 3: Boxplots for  $\|\mathbf{A}\mathbf{A}' - \widehat{\mathbf{A}}\widehat{\mathbf{A}}'\|_2$  of Example 1. For each  $(p, N)$  configuration, the sample sizes used are  $T = 400, 800, 1200$ , and the number of repetitions is 500.

Finally, we consider the estimation errors of the estimated explanatory variables and the true ones in Model (2.5). Similarly to that in (4.1), we define the RMSE for the regression model (2.5) as

$$RMSE = \left( \frac{1}{pT} \sum_{t=1}^T \|\mathbf{A}\mathbf{z}_{t-1} + \Phi\mathbf{y}_{t-1} - (\widehat{\mathbf{A}}\widehat{\mathbf{z}}_{t-1} + \widehat{\Phi}\mathbf{y}_{t-1})\|_2^2 \right)^{1/2}, \quad (4.2)$$

which is similar to the in-sample errors of a regression model. Figure 5 displays boxplots of the RMSEs in (4.2). From the plot, we see that the patterns of the boxplots are similar to those obtained before. For each given  $(p, N)$ , the RMSEs decrease as the sample size  $T$

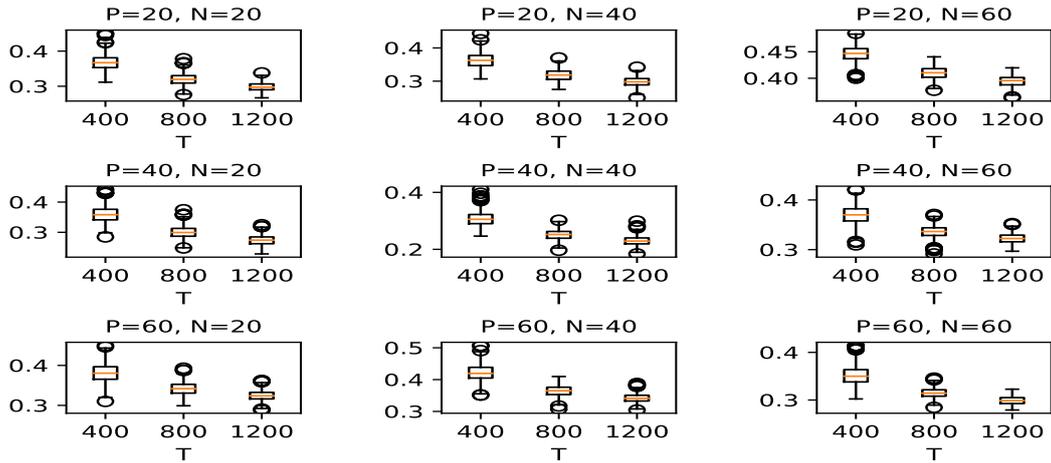


Figure 4: Boxplots for  $\|\Phi - \hat{\Phi}\|_2$  of Example 1. In each case of  $(p, N)$ , the sample sizes used are  $T = 400, 800, 1200$ , and the results are based on 500 repetitions.

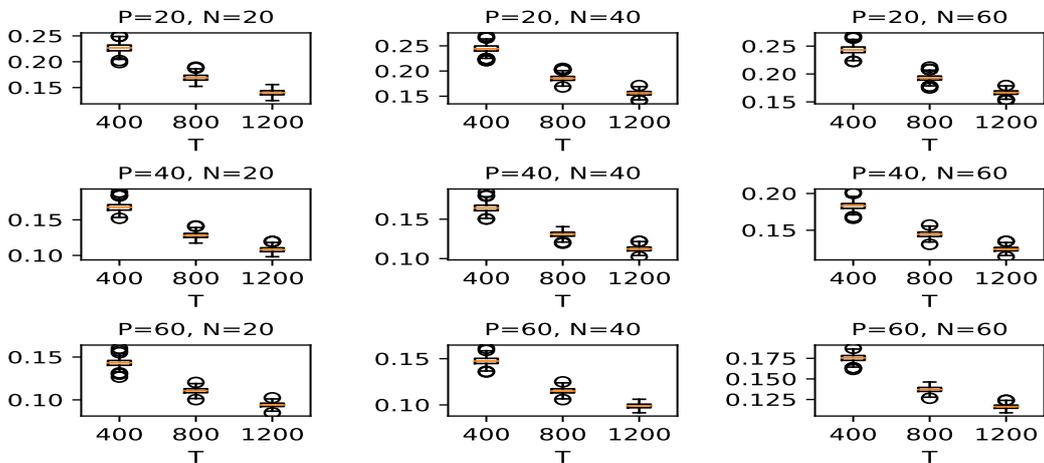


Figure 5: Boxplots for denoised RMSE of  $\mathbf{y}_t$  defined in (4.2) of Example 1. For each pair of  $(p, N)$ , the sample sizes used are 400, 800 and 1200 and the number of repetitions is 500.

increases, illustrating the efficacy of the proposed method. Overall, the simulation results indicate that the proposed procedure works well in recovering the estimated coefficients.

## 4.2 Example 2: The Integrative Reduced-Rank Approach

In this example, we investigate the performance of IRRA of Section 2.2.2. First, we generate the data  $\mathbf{x}_t$  using the same method as that of Section 4.1.1. Second, unlike the sparse autoregressive matrices in Example 1, we generate two low-rank matrices  $\Phi_1 \in \mathbb{R}^{p \times p}$  and  $\Phi_2 \in \mathbb{R}^{p \times p}$  under the context of the IRRA. Without loss of generality, we generate the those low-rank matrices in the same way as that of  $\mathbf{A}$ , and set  $\text{rank}(\Phi_1) = \text{rank}(\Phi_2) = 3$ . Third, the process  $\mathbf{y}_t$  is then generated according to (2.5) with the coefficients given above, where we choose  $d = 2$ .

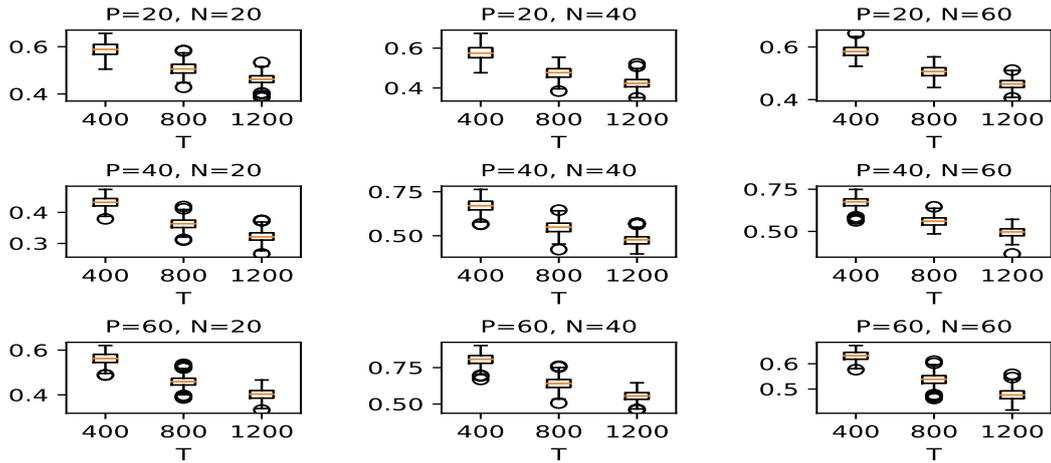


Figure 6: Boxplots for  $\|\mathbf{A}\mathbf{A}' - \widehat{\mathbf{A}}\widehat{\mathbf{A}}'\|_2$  of Example 2, where  $\widehat{\mathbf{A}}$  is estimated by Algorithm 2. For each pair of  $(p, N)$ , the sample sizes used are  $T = 400, 800$  or  $1200$ , and the number of repetitions is 500.

Similarly to the procedure in Example 1, we apply (2.9) and Algorithm 2 to estimate the number of factors and the coefficients, respectively. Since the performance of the auto-correlation based method is shown in Example 1, we omit the details here. Figures 6, 7 and 8 show the discrepancies between the estimated coefficients and the true ones using Algorithm 2. From these boxplots, we see that, for each configuration of  $(p, N)$ , all three coefficient estimates  $\widehat{\mathbf{A}}$ ,  $\widehat{\Phi}_1$  and  $\widehat{\Phi}_2$  converge to the true ones as  $T \rightarrow \infty$ , which is consistent with our theory. For comparison, we also test the ADMM algorithm of Li, Liu and Chen (2019), and find that the results of ADMM are quite close to those of the Algorithm 2 in the sense that the distance  $\|\widehat{\Theta}^{(\text{Ite})} - \widehat{\Theta}^{(\text{ADMM})}\|_2 / \|\widehat{\Theta}^{(\text{Ite})}\|_2$  is less than 10% in most cases, where  $\widehat{\Theta} = \widehat{\mathbf{A}}, \widehat{\Phi}_1$  or  $\widehat{\Phi}_2$ , and  $\widehat{\Theta}^{(\text{Ite})}$  and  $\widehat{\Theta}^{(\text{ADMM})}$  are the coefficient matrices estimated by Algorithm 2 and ADMM, respectively. Therefore, we omit the results obtained by the ADMM algorithm to save space.

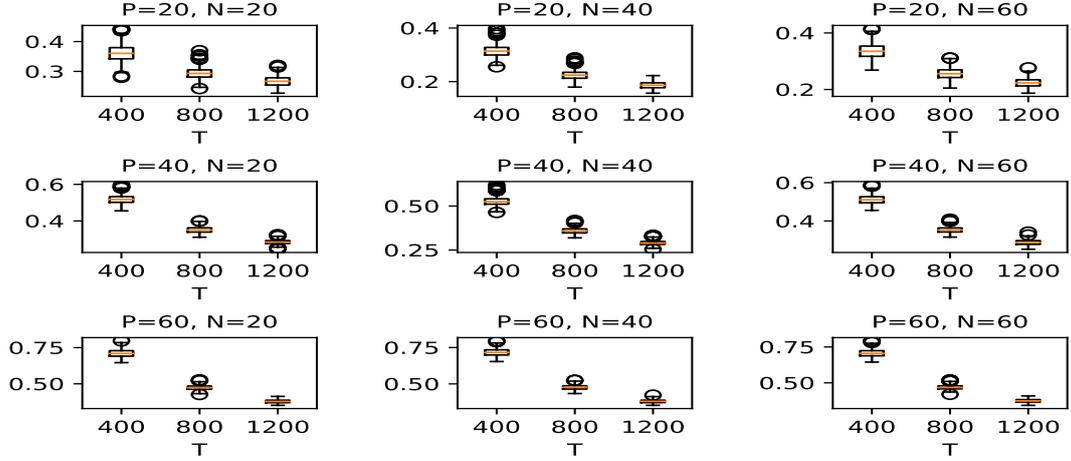


Figure 7: Boxplots for  $\|\Phi_1 - \hat{\Phi}_1\|_2$  in Example 2, where  $\hat{\Phi}$  is estimated by Algorithm 2. For each pair of  $(p, N)$ , the sample sizes used are  $T = 400, 800$  and  $1200$ , and the number of repetitions is 500.

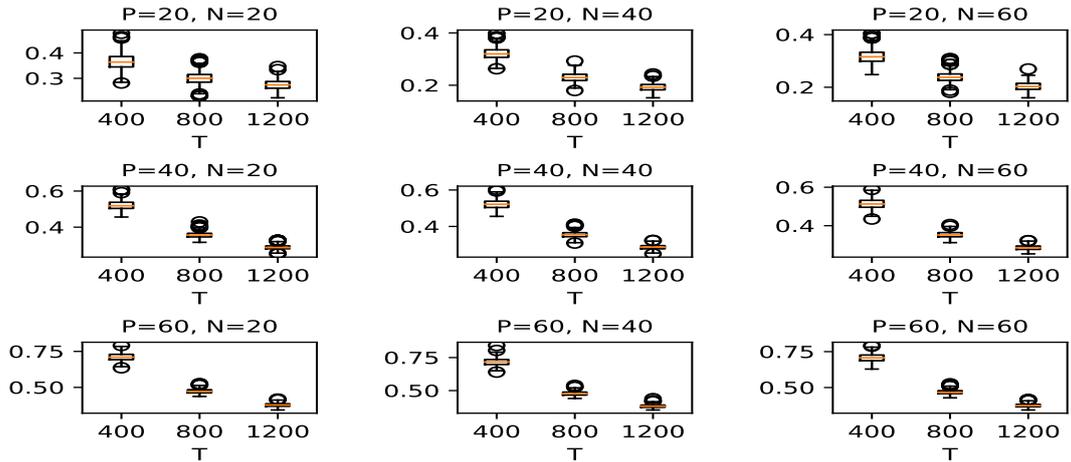


Figure 8: Boxplots for  $\|\Phi_2 - \hat{\Phi}_2\|_2$  in Example 2, where  $\hat{\Phi}$  is estimated by Algorithm 2. For each pair of  $(p, N)$ , the sample sizes used are  $T = 400, 800$  and  $1200$ , and the results are based on 500 repetitions.

## 5 Real Data Analysis

In this section, we apply the proposed method to predicting monthly stock returns. Welch and Goyal (2008) examined the predictability of some macroeconomic variables to the equity premium, and concluded that the performance of the predictions, both in-sample and out-of-sample, is poor and unstable. Using the same set of predictors, Koo et al. (2020) exploited the cointegration relationship of the predictors, and showed that LASSO can improve the predictability of the macroeconomic variables in forecasting the equity premium of S&P 500 index. We use an extended data set and conduct a forecasting experiment using the proposed method. Note that we predict the stock returns of a cross-section, instead of the equity premium of an individual stock or index.

### 5.1 Data and Empirical Strategy

Consider the monthly returns of selected stocks in the S&P 500 index. Using the constituents of the index in January 2011 and the data structures in the CRSP (Center for Research in Security Prices) database, we select 79 stocks, which have no missing values during the time span from January 1960 to December 2019, as our sample. Therefore, we have 720 monthly observations of 79  $I(0)$  processes. We also collect the monthly macroeconomic variables in Welch and Goyal (2008) as the predictors. An updated version of the data can be downloaded from Prof. Amit Goyal's personal website (<https://sites.google.com/view/agoyal145>), where we choose 13 macroeconomic predictors as the  $I(1)$  processes  $\mathbf{x}_t$  from December 1959 to November 2019. Therefore, we have  $N = 13$ ,  $p = 79$  and  $T = 720$  in this illustration.

Table 2 presents some descriptive statistics of the macro predictors, including their first-order sample autocorrelation coefficients  $\rho(1)$  over the entire sample period. As shown in the table, nine predictors have a first-order sample autocorrelation coefficient higher than 0.95, but four variables (inflation, long-term yield, corporate bond returns, and stock variance) show little persistence. Therefore, we see that most of the variables are highly persistent and can be used as the  $I(1)$  predictors in our model.

For the purpose of evaluating the forecasting performance of the proposed method, we adopt the out-of-sample  $R^2$  measure commonly used in the literature regarding the prediction of stock returns, see Gu, Kelly and Xiu (2020). At each time point, we define

Table 2: Descriptive statistics of the macroeconomic predictors, and their first-order sample autocorrelation coefficients over the entire sample period. The sample size is  $T = 720$ .

variable	mean	std	min	25%	50%	75%	max	$\rho(1)$
D12	14.5619	13.4247	1.8667	3.6175	11.0988	19.5073	58.2406	0.9919
E12	34.2835	34.3074	3.0300	8.1850	18.1334	51.0358	139.4700	0.9921
b/m	0.4899	0.2565	0.1205	0.2872	0.4382	0.6395	1.2065	0.9937
tbl	0.0454	0.0316	0.0001	0.0229	0.0462	0.0612	0.1630	0.9904
AAA	0.0705	0.0263	0.0298	0.0488	0.0698	0.0856	0.1549	0.9943
BAA	0.0806	0.0287	0.0387	0.0566	0.0789	0.0960	0.1718	0.9952
lty	0.0631	0.0274	0.0163	0.0423	0.0598	0.0800	0.1482	0.9921
ntis	0.0100	0.0199	-0.0560	-0.0022	0.0130	0.0245	0.0512	0.9809
Rfree	0.0037	0.0026	0.0000	0.0018	0.0037	0.0050	0.0135	0.9719
infl	0.0030	0.0036	-0.0192	0.0007	0.0029	0.0050	0.0181	0.5700
ltr	0.0061	0.0291	-0.1124	-0.0104	0.0040	0.0228	0.1523	0.0310
corpr	0.0063	0.0257	-0.0949	-0.0071	0.0052	0.0191	0.1560	0.1077
svar	0.0021	0.0043	0.0001	0.0006	0.0011	0.0021	0.0709	0.4717

the out-of-sample  $R^2$  as

$$R_{\text{OOS}}^2(t) = 1 - \frac{\|\mathbf{y}_t - (\hat{\mathbf{A}}\hat{\mathbf{z}}_{t-1} + \hat{\mathbf{\Phi}}\mathbf{P}_{t-1})\|_2^2}{\|\mathbf{y}_t\|_2^2},$$

which differs from that in Gu, Kelly and Xiu (2020) by being a function of the time index  $t$ , which denotes the forecasting origin.

Our empirical analysis works as follows. First, we divide the time span of 60 years into two periods. The first period is the initial estimation period from January 1960 to December 2010, and the second one is the testing period from January 2011 to December 2019. Specifically, we conduct the empirical test according to the following procedure, which is similar to that commonly used in the asset pricing literature. At the beginning, we use the data from January 1960 to December 2010 to estimate the coefficient matrices  $\mathbf{A}$  and  $\mathbf{\Phi}$  of the model (2.5), then predict the returns of the 79 stocks of January 2011 and calculate the out-of-sample  $R^2$  of the prediction. We then add the returns of January 2011 to the estimation period, and refit the model (2.5) with data from January 1960 to January 2011 to obtain updated coefficient matrices. The updated model is used to predict the returns of February 2011 with the model (2.5) and to calculate the out-of-sample  $R^2$  again. We repeat this estimation-prediction process by adding one-month returns to the estimation period in each iteration until November 2019, which enables us to predict the returns for December 2019. In addition, we choose the tuning parameters  $\lambda_{\mathbf{A}}$  and  $\lambda_{\mathbf{\Phi}}$  based on the procedure described in Section 2.4 but letting  $\lambda_{\mathbf{A}}$  and  $\lambda_{\mathbf{\Phi}}$  be proportional to

$\sqrt{(p+N)/T}$  and  $\sqrt{\log(p)/T}$ , respectively. See Remark 2(ii).

For comparison, we consider some alternative models commonly seen in the literature as benchmarks. The first benchmark is the naive VAR( $d$ ) model, that is,

$$\mathbf{y}_t = \mathbf{\Psi}\mathbf{P}_{t-1} + \mathbf{e}_t, \quad t = 1, 2, \dots, T.$$

For each series of  $\mathbf{y}_t$  and the corresponding row of  $\mathbf{\Psi}$ , we can treat the above equation as a simple regression problem with  $dp$  explanatory variables and  $T$  observations. Therefore, we may use the LS method to estimate each row of  $\mathbf{\Psi}$ , then put them together to construct an estimator of  $\mathbf{\Psi}$ . We use the model in Koo et al. (2020) as another benchmark, where the tuning parameter  $\lambda$  is fixed to  $\frac{\log(p)}{10\sqrt{T}}$ . Since the original model in Koo et al. (2020) is developed for predicting a scalar time series, we apply their model with 13 macroeconomic predictors described above to predict each stock separately, then stack the predictions together to calculate the out-of-sample  $R^2$  for all 79 stocks. Our final benchmark is the random walk model, in which we predict the returns of the next period using returns of the current period. For a more comprehensive comparison, we conduct the experiment for  $d = 1, 2$  and 3, respectively, for the naive VAR, the RRSRA, and the IRRRA models.

## 5.2 Prediction Performance of RRSRA

We evaluate the empirical performance of the RRSRA in this section. To begin, Figure 9 shows a time plot of the estimated number of common trends by the method described in Section 2.3 in the testing period. The figure shows that, except for the first three months of 2011 that may be affected by some economic crisis, the estimated number of common trends within  $\mathbf{x}_t$  is four, which is fairly stable over the entire test period. Thus, we have nine cointegrating vectors to produce the stationary process  $\hat{\mathbf{z}}_t$  as a proxy of macroeconomic predictors. Before analyzing the forecasting performance of these estimated  $\hat{\mathbf{z}}_t$  variables, we take a look at the number of parameters to be estimated in the two coefficient matrices  $\hat{\mathbf{A}}$  and  $\hat{\mathbf{\Phi}}$ . Suppose that  $\hat{r} = 4$ , then there are 9 cointegrating vectors and hence, the matrix  $\hat{\mathbf{A}}$  has  $79 \times 9 = 711$  entries, and the matrix  $\hat{\mathbf{\Phi}}$  has  $79 \times 79 = 6241$  entries to be estimated, both of which are relatively large. Therefore, we expect that the dimensions of the two matrices can further be reduced to low-rank or sparse ones, which is commonly assumed in the literature to avoid over-fitting and to produce better forecasting performance. For this reason, we expect that the tuning parameters  $\lambda_{\mathbf{A}}$  and  $\lambda_{\mathbf{\Phi}}$  in our framework should be

relatively large to guarantee that the dimensions can be reduced.

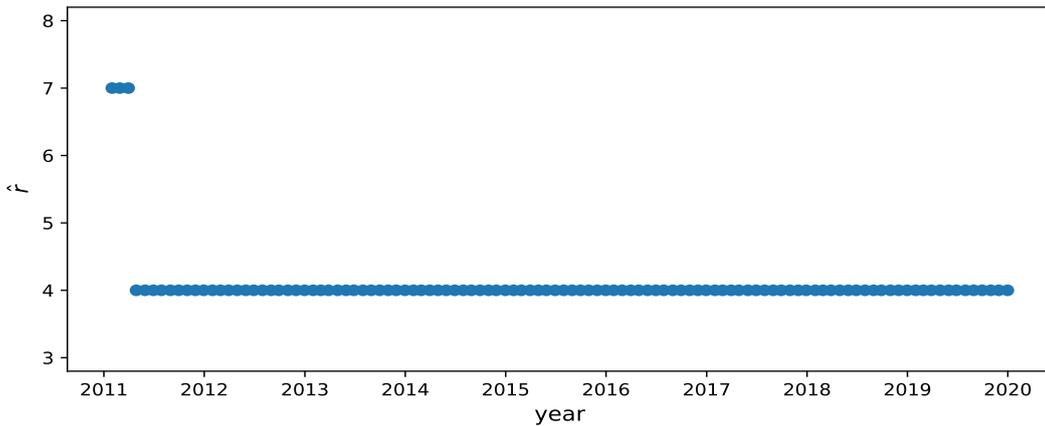


Figure 9: Time plot of  $\hat{r}$  obtained using all the data before the corresponding time point on the horizontal axis.

Figure 10 shows the estimated rank of  $\hat{\mathbf{A}}$  and the estimated number of non-zero entries of  $\hat{\Phi}$  for the proposed model with  $d = 1$  at each prediction time point. The average rank of  $\hat{\mathbf{A}}$  is 1.97 and the average number of non-zero entries of  $\hat{\Phi}$  is 5.88. Except for the first three months in 2011, the estimated rank of  $\hat{\mathbf{A}}$  is 2 in the estimation period. In addition,  $\hat{\Phi}$  has at most 13 non-zero entries at all time points, which is extremely small compared to 6241 of the total number of entries. Overall, the proposed method provides an effective way to reduce the number of parameters and the dimension of the coefficient matrices.

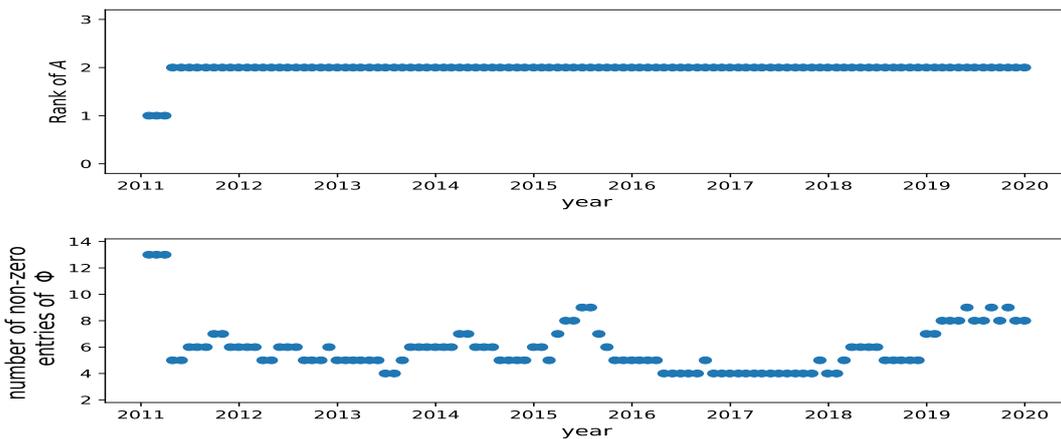


Figure 10: The estimated rank of  $\hat{\mathbf{A}}$  and the number of non-zero entries of  $\hat{\Phi}$ . The two coefficient matrices are estimated with  $d = 1$  using all the data prior to the time point.

Next we show the forecasting results in detail by following the method described in Section 2.2 and the forecasting procedure mentioned above to evaluate the performance of different models. Table 3 reports the overall comparisons of our proposed method against

the three benchmarks mentioned before, in terms of  $R_{\text{OOS}}^2$ . From Panels A and C of Table 3, we see that the proposed RRSRA substantially outperforms the naive VAR model (denoted by VAR( $d$ )), the method of Koo et al. (2020) (denoted by Koo), and the random walk model (denoted by RW). The mean of out-of-sample  $R^2$  of our method with  $d = 1$  is 0.91% and the result is nearly the same for  $d = 2$  or 3. We also note that our results are slightly better than those in Gu, Kelly and Xiu (2020), where the highest monthly out-of-sample  $R^2$  for all stocks is 0.40% among all machine learning methods considered in their paper, and it is 0.70% for the top 1,000 stocks and 0.47% for the bottom 1,000 stocks by market values. In particular, the VAR model performs relatively poorly, as the out-of-sample  $R^2$ s with different lags all assume negative values with large magnitudes, which are  $-22.54\%$ ,  $-41.08\%$  and  $-87.81\%$ , respectively. One possible reason is that the number of parameters to be estimated in VAR models is significantly large and this often leads to severe over-fitting, which in turn produces high variations in out-of-sample forecasting. When the lag order  $d$  increases, the number of parameters also increases, so the performance would further deteriorate. For the model in Koo et al. (2020), the results in Table 3 imply that the Koo method has limited predictive power when forecasting the returns of individual stocks. Finally, we see that the random walk model performs the worst. In summary, the proposed model has marked advantages in prediction over the three benchmark models considered.

To explore the in-sample goodness of fit, we apply all entertained models except the random walk to the entire data set, and calculate the in-sample  $R^2$  at each time point. The results over the time are shown in Table 4. As expected, the VAR model, which has many more degrees of freedom than the others, produces the highest in-sample  $R^2$ . Our model and the one of Koo et al. (2020) provide a robust in-sample fit, and the result by the Koo method is only slightly worse than those of the proposed models.

To check whether our model outperforms the benchmarks uniformly over the in-sample and the out-of-sample periods, we plot the in-sample  $R^2$ s and out-of-sample ones of the RRSRA(1) model in Figures 11 and 12, respectively, where the time index is on the horizontal axis. For a better illustration, the points in Figure 11 are the in-sample  $R^2$ s based on the data of each year from 1960 to 2019. In both figures, we also plot the VAR(1) as a benchmark. An additional plot of the random walk model is also included in Figure 12 as another benchmark. Because the results produced by the Koo method in Koo et al. (2020) are very close to ours, they are omitted. From Figure 11, we see that our method

Table 3: Comparison of the RRSRA model, IRRA model and several benchmarks in terms of out-of-sample  $R^2$ . Panel A shows the results for RRSRA model with lag  $d = 1, 2, 3$ . Panel B is the results for IRRA model, where we use both Algorithm 2 and ADMM method in estimation, and with  $d = 1, 2, 3$ , respectively. Panel C reports several benchmark models, including the naive VAR model with lag  $d = 1, 2, 3$ , the method in Koo et al. (2020) denoted by Koo, and the random walk model denoted by RW. For each time point in the test sample, we calculate the value of  $R_{\text{OOS}}^2$ . The size of the test sample is  $108 - d + 1$  for all models.

	Out-of-sample $R^2(t)$ (in percentage)						
	mean	std	min	25%	50%	75%	max
<i>Panel A. Method (2.7) with different <math>d</math></i>							
RRSRA(1)	0.91	18.73	-85.75	-6.42	4.61	13.25	31.45
RRSRA(2)	0.92	18.70	-85.76	-6.38	4.66	13.20	31.53
RRSRA(3)	0.90	18.83	-86.44	-6.32	4.67	13.33	31.78
<i>Panel B. Method (2.8) with different <math>d</math>, fitted using both iterative method and ADMM method</i>							
Iterative(1)	0.75	18.29	-80.36	-6.77	4.79	13.09	30.90
Iterative(2)	0.66	18.39	-76.71	-7.01	3.96	13.18	29.99
Iterative(3)	0.59	18.53	-77.88	-7.71	4.21	13.66	28.77
ADMM(1)	0.80	18.18	-79.28	-6.57	4.83	13.11	30.61
ADMM(2)	0.65	18.38	-76.52	-7.01	3.96	13.16	30.03
ADMM(3)	0.60	18.53	-77.87	-7.67	4.11	13.66	29.13
<i>Panel C. Benchmark models</i>							
VAR(1)	-22.54	31.23	-133.10	-34.86	-18.96	0.05	24.27
VAR(2)	-41.08	49.96	-344.47	-65.06	-31.93	-12.30	37.23
VAR(3)	-87.81	97.26	-797.91	-117.46	-66.46	-35.08	46.18
Koo	-5.23	21.67	-91.40	-17.98	-0.53	11.32	26.52
RW	-127.08	117.45	-665.03	-163.63	-111.06	-49.64	37.24

Table 4: In-sample  $R^2$  of method (2.7), method (2.8) and some benchmarks. The three models in Panel A are the RRSRA model with lag  $d = 1, 2, 3$ . Panel B shows the results for IRRA, where we use both iterative method and ADMM method in estimation with  $d = 1, 2, 3$ , respectively. Panel C reports some benchmark models, that is, the naive VAR model with  $d = 1, 2, 3$ , the method in Koo et al. (2020) denoted by Koo, and the random walk model denoted by RW. We fit each model with the entire data set and obtain fitted values for returns of individual stocks, then calculate  $R^2$  at each time point. The sample size is  $719 - d + 1$  for all models.

	In-sample $R^2$ (in percentage)						
	mean	std	min	25%	50%	75%	max
<i>Panel A. Method (2.7) with different d</i>							
RRSRA(1)	1.53	16.02	-68.32	-9.36	3.80	13.40	42.66
RRSRA(2)	1.52	15.99	-68.44	-9.37	3.77	13.32	42.47
RRSRA(3)	1.55	16.04	-68.63	-9.29	3.75	13.42	42.63
<i>Panel B. Method (2.8) with different d, fitted using both iterative method and ADMM method</i>							
Iterative(1)	1.61	16.00	-68.16	-8.95	3.85	13.25	43.03
Iterative(2)	1.58	15.99	-65.52	-8.84	3.89	13.27	43.42
Iterative(3)	1.71	16.03	-65.41	-9.10	3.81	13.40	42.95
ADMM(1)	1.59	15.98	-67.62	-8.99	3.91	13.25	42.84
ADMM(2)	1.58	15.97	-65.14	-8.87	3.92	13.26	43.34
ADMM(3)	1.69	16.01	-65.19	-8.96	3.78	13.33	42.98
<i>Panel C. Several benchmark models</i>							
VAR(1)	8.10	24.69	-137.87	-3.87	11.85	25.34	69.86
VAR(2)	15.96	30.09	-121.38	0.80	20.39	37.38	84.32
VAR(3)	24.69	35.48	-303.66	9.14	30.57	48.35	89.87
Koo	-1.13	16.30	-48.29	-14.18	2.03	12.90	34.17

fits the data relatively poorly compared to the VAR model in most years according to the in-sample  $R^2$ . This is understandable since the VAR model fits the data via the LS method to minimize the squared distance between the fitted values and the true ones, while our method adopts regularization, which often introduces some in-sample biases in order to provide more stable predictions in out-of-samples. Furthermore, Figure 12 shows that our method produces more robust predictions than the VAR and the random walk model, and outperforms them over most of the time points based on the out-of-sample  $R^2$ . This illustrates the predictive advantages of using the proposed method.

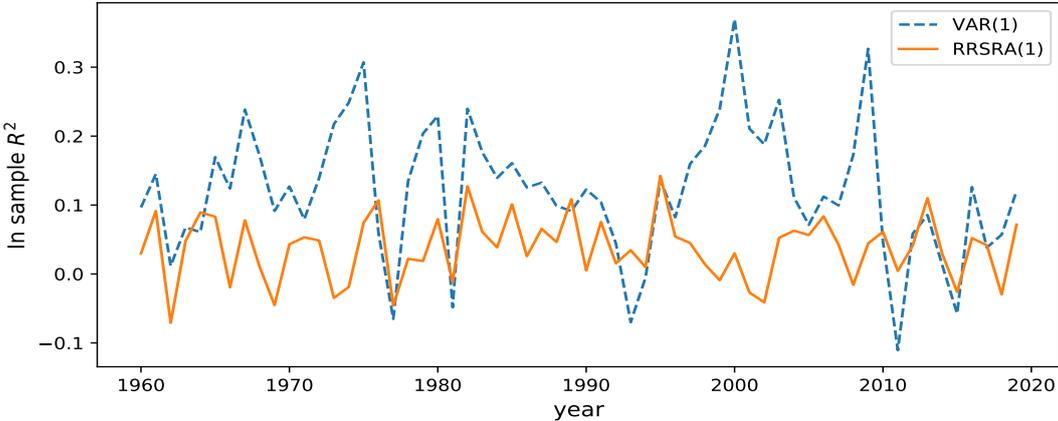


Figure 11: In-sample  $R^2$  for the VAR(1) model and our method with  $d = 1$ . We use the entire data set in estimation and calculate the in-sample  $R^2$  using the data of each year from 1960 to 2019.

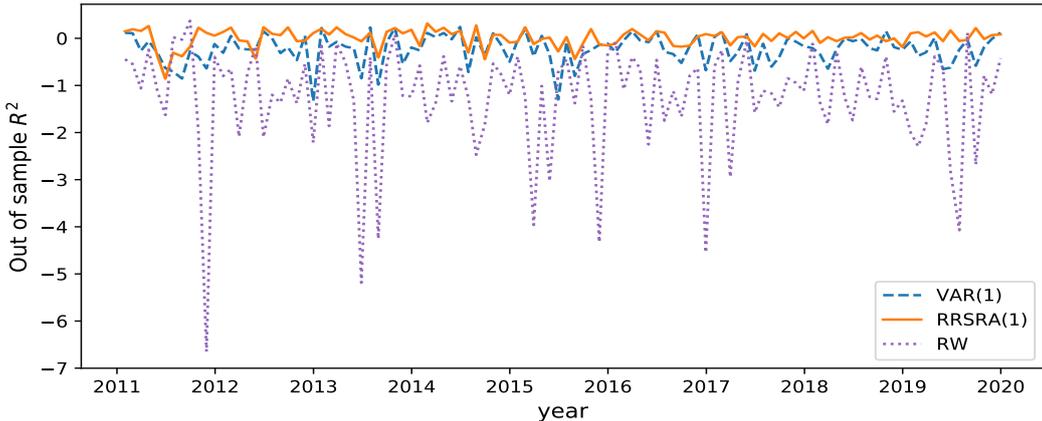


Figure 12: Out-of-sample  $R^2$  for the proposed model with  $d = 1$ , the VAR(1) model and the random walk model. For each time point in the test period, we fit each model using the data prior to that time point, make predictions for the returns, and calculate the out-of-sample  $R^2$  of the predictions.

### 5.3 Prediction Performance of IRRA

In this section, we evaluate the predictive performance of IRRA models described in Section 2.2.2. The procedure of estimating the factor model (2.2) and obtaining  $\mathbf{z}_t$  are exactly the same as those in Section 5.2. With the estimated  $\hat{\mathbf{z}}_t$ , we fit the data via (2.8) using both iterative method and ADMM method, and evaluate its predictive performance. We expect that the results of IRRA are close to those of RRSRA, because the tuning parameter  $\lambda_{\Phi}$  selected by a grid search is relatively large in both models. When the tuning parameter  $\lambda_{\Phi}$  in both methods tends to infinity, the estimated coefficients obtained by the two algorithms tend to be the same.

We first examine the estimated coefficient matrices. Figure 13 shows the rank of  $\hat{\mathbf{A}}$  and  $\hat{\Phi}$  estimated by Algorithm 2 in the case of  $d = 1$  at each time point. We find that the rank of  $\hat{\mathbf{A}}$  is reduced to 1 or 2 over the entire time horizon, which is the same as that of the RRSRA, implying that the efficient cointegration rank is low in this particular application. In addition, the rank of  $\hat{\Phi}$  is also 1 or 2 over time.

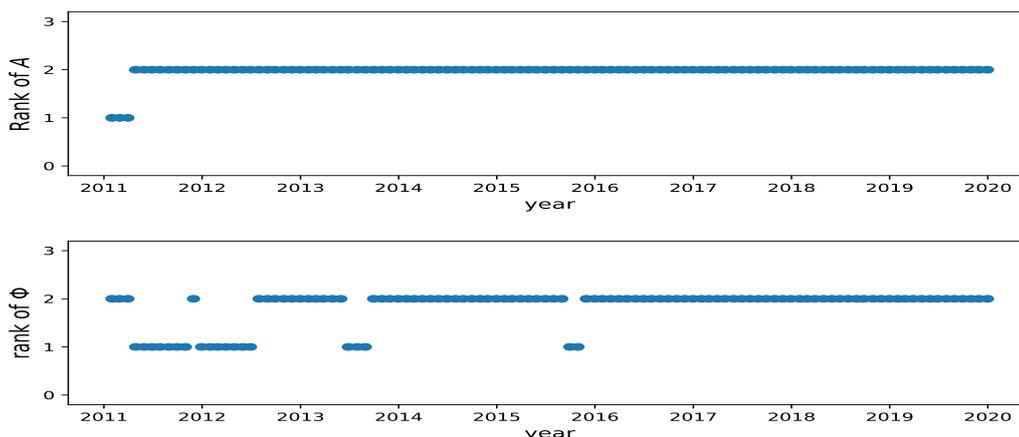


Figure 13: The estimated rank of  $\hat{\mathbf{A}}$  and the rank of  $\hat{\Phi}$ . The two coefficient matrices are estimated by Algorithm 2 with  $d = 1$  at each time point using all the data prior to it.

Panel B of Table 3 shows the predictive performance of (2.8), where the coefficients are estimated by both Algorithm 2 and the ADMM method with  $d = 1, 2, 3$ , respectively. All six results are positive and close to each other. They are also close to, but a little worse than, those of RRSRA. One possible reason is that both IRRA and RRSRA are constrained regressions but the latter one produces sparse solutions and reduces the model complexity more substantially compared to the low-rank structures. Similarly to the conclusion of the RRSRA procedure, the  $R_{\text{OOS}}^2$  results of (2.8) also outperform those of the benchmarks.

Finally, we see that the performance of IRRA is close to that of RRSRA not only with respect to the out-of-sample  $R^2$ , but also with respect to the in-sample  $R^2$ ; see Tables 3 and 4. Panel B of Table 4 shows the in-sample  $R^2$  results for IRRA, with all six estimation settings. Once again, we find that the results are close to those in Panel A, but IRRA fits the data slightly better, which may be a consequence of higher degrees of freedom in IRRA.

## 6 Concluding Remarks

Finding proper cointegration relationships is an important topic in Econometrics and Statistics, yet the interpretation of cointegrating structures might become complicated if the dimension of the system under study is high. This paper introduced the concept of *effective cointegration rank* and considered a new method to identify the important cointegration relationships among a high-dimensional  $I(1)$  series from a predictive perspective. In a nutshell, the effective cointegration rank is the number of cointegrating relationships that can produce useful predictors in a given forecasting application. The proposed method consists of a two-step estimation procedure, where we first use the Principal Component Analysis to estimate the common stochastic trends of the  $I(1)$  series and to identify all possible cointegrating vectors. We then employ all stationary series obtained via the cointegrating vectors and some lagged values of dependent variables to form predictors of the second-step estimation. A reduced-rank regression technique is applied to the co-integrated predictors and the dimension of relevant cointegrating vectors is defined as the effective cointegration rank. We also applied the LASSO penalty or reduced rank constraints to the coefficients of the lagged variables in the second step, and an iterative procedure is proposed to estimate the unknown coefficients.

Our proposed method has a wide range of applications in many scientific areas, including Economics, Finance, and Environmental studies, because it is common in these areas to use nonstationary variables or factors to predict stationary series in empirical applications. We applied the proposed method to the problem of predicting cross-sectional stock returns, and illustrated clearly the predictive advantages of the proposed procedure over some commonly used benchmarks available in the literature.

## References

- Agarwal, A., Negahban, S. and Wainwright, M.J., 2012. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics*, 40(2), pp.1171–1197.
- Aznar, A. and Salvador, M., 2002. Selecting the rank of the cointegration space and the form of the intercept using an information criterion. *Econometric Theory*, 18(4), pp.926–947.
- Bai, J., 2004. Estimating cross-section common stochastic trends in nonstationary panel data. *Journal of Econometrics*, 122(1), pp.137–183.
- Bai, J. and Ng, S., 2002. Determining the number of factors in approximate factor models. *Econometrica*, 70(1), pp.191–221.
- Banerjee, A., Marcellino, M. and Masten, I., 2014. Forecasting with factor-augmented error correction models. *International Journal of Forecasting*, 30(3), pp.589–612.
- Billingsley, P., 1999. *Convergence of probability measures*. John Wiley & Sons.
- Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J. et al., 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine learning*, 3(1), pp.1–122.
- Boyd, S. and Vandenberghe, L., 2004. *Convex optimization*. Cambridge university press.
- Burai, P., 2013. Necessary and sufficient condition on global optimality without convexity and second order differentiability. *Optimization Letters*, 7(5), pp.903–911.
- Chen, K., Dong, H. and Chan, K.-S., 2013. Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, 100(4), pp.901–920.
- Engle, R.F. and Granger, C.W., 1987. Co-integration and error correction: representation, estimation, and testing. *Econometrica*, pp.251–276.
- Fan, J., Liao, Y. and Mincheva, M., 2013. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4), pp.603–680.

- 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(471), pp.830–840.
- Gao, Z., Ma, Y., Wang, H. and Yao, Q., 2019. Banded spatio-temporal autoregressions. *Journal of Econometrics*, 208(1), pp.211–230.
- Gao, Z. and Tsay, R.S., 2019. A structural-factor approach to modeling high-dimensional time series and space-time data. *Journal of Time Series Analysis*, 40(3), pp.343–362.
- Gao, Z. and Tsay, R.S., 2021a. A two-way transformed factor model for matrix-variate time series. *Econometrics and Statistics*.
- Gao, Z. and Tsay, R.S., 2021b. Modeling high-dimensional time series: A factor model with dynamically dependent factors and diverging eigenvalues. *Journal of the American Statistical Association*, pp.1–17.
- Gao, Z. and Tsay, R.S., 2021c. Modeling high-dimensional unit-root time series. *International Journal of Forecasting*, 37(4), pp.1535–1555.
- Gao, Z. and Tsay, R.S., 2022. Divide-and-conquer: a distributed hierarchical factor approach to modeling large-scale time series data. *Journal of the American Statistical Association*, forthcoming.
- Gu, S., Kelly, B. and Xiu, D., 2020. Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5), pp.2223–2273.
- Hastie, T., Tibshirani, R. and Wainwright, M.J., 2015. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press.
- Ji, S. and Ye, J., 2009. An accelerated gradient method for trace norm minimization. *Proceedings of the 26th Annual International Conference on Machine Learning*, pp.457–464.
- Johansen, S., 1988. Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, 12(2-3), pp.231–254.
- Johansen, S., 1991. Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica*, pp.1551–1580.

- Johansen, S., 2002. A small sample correction for the test of cointegrating rank in the vector autoregressive model. *Econometrica*, 70(5), pp.1929–1961.
- Koo, B., Anderson, H.M., Seo, M.H. and Yao, W., 2020. High-dimensional predictive regression in the presence of cointegration. *Journal of Econometrics*, 219(2), pp.456–477.
- Lam, C. and Yao, Q., 2012. Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics*, pp.694–726.
- Lam, C., Yao, Q. and Bathia, N., 2011. Estimation of latent factors for high-dimensional time series. *Biometrika*, 98(4), pp.901–918.
- Li, G., Liu, X. and Chen, K., 2019. Integrative multi-view regression: Bridging group-sparse and low-rank models. *Biometrics*, 75(2), pp.593–602.
- Lin, J. and Michailidis, G., 2017. Regularized estimation and testing for high-dimensional multi-block vector-autoregressive models. *Journal of Machine Learning Research*, 18.
- Lütkepohl, H., 2006. *New introduction to multiple time series analysis*. Springer Science & Business Media.
- Merlevède, F., Peligrad, M. and Rio, E., 2011. A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, 151(3), pp.435–474.
- Negahban, S. and Wainwright, M.J., 2011. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2), pp.1069–1097.
- Pan, J. and Yao, Q., 2008. Modelling multiple time series via common factors. *Biometrika*, 95(2), pp.365–379.
- Peña, D. and Poncela, P., 2006. Nonstationary dynamic factor analysis. *Journal of Statistical Planning and Inference*, 136(4), pp.1237–1257.
- Reinsel, G.C., Velu, R.P. and Chen, K., 2022+. *Multivariate reduced-rank regression: theory and applications (2nd ed.)*. Springer.
- Saikkonen, P. and Lütkepohl, H., 2000. Testing for the cointegrating rank of a VAR process with structural shifts. *Journal of Business & Economic Statistics*, 18(4), pp.451–464.

- Stock, J.H., 1987. Asymptotic properties of least squares estimators of cointegrating vectors. *Econometrica*, pp.1035–1056.
- Stock, J.H. and Watson, M.W., 2005. *Implications of dynamic factor models for VAR analysis*. National Bureau of Economic Research Cambridge, Mass., USA.
- Tiao, G.C. and Tsay, R.S., 1989. Model specification in multivariate time series (with discussion). *Journal of the Royal Statistical Society: Series B (Methodological)*, 51(2), pp.157–195.
- Tsay, R.S., 2014. *Multivariate time series analysis: with R and financial applications*. John Wiley & Sons.
- Tseng, P., 2001. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3), pp.475–494.
- Vandenberghe, L. and Boyd, S., 1996. Semidefinite programming. *SIAM Review*, 38(1), pp.49–95.
- Wainwright, M.J., 2019. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press.
- Welch, I. and Goyal, A., 2008. A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 21(4), pp.1455–1508.
- Zhang, R., Robinson, P. and Yao, Q., 2019. Identifying cointegration by eigenanalysis. *Journal of the American Statistical Association*, 114(526), pp.916–927.

# Appendices

In this supplement, we provide proofs of all theorems stated in Section 3 of the main article. We use  $c$  or  $C$  to denote a generic positive constant, its value may change for different places.

## Appendix A Proof of Theorems

### A.1 Proofs of Theorem 1 and 2

To begin, we first introduce a useful lemma, which is commonly seen in matrix perturbation theory. See Golub and Van Loan (2013) (Theorem 8.1.10), Johnstone and Lu (2009), and Lam, Yao and Bathia (2011), among others.

**Lemma A1.** *Suppose  $\mathbf{A}$  and  $\mathbf{A} + \mathbf{E}$  are  $n \times n$  symmetric matrices, and  $\mathbf{Q} = [\mathbf{Q}_1 \ \mathbf{Q}_2]$ , with  $\mathbf{Q}_1 \in \mathbb{R}^{n \times r}$  and  $\mathbf{Q}_2 \in \mathbb{R}^{n \times (n-r)}$ , is an  $n \times n$  orthogonal matrix such that  $\text{range}(\mathbf{Q}_1)$  is an invariant subspace for  $\mathbf{A}$ . Partition the matrices  $\mathbf{Q}'\mathbf{A}\mathbf{Q}$  and  $\mathbf{Q}'\mathbf{E}\mathbf{Q}$  as follows:*

$$\mathbf{Q}'\mathbf{A}\mathbf{Q} = \begin{bmatrix} \mathbf{Q}'_1\mathbf{A}\mathbf{Q}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}'_2\mathbf{A}\mathbf{Q}_2 \end{bmatrix} \quad \text{and} \quad \mathbf{Q}'\mathbf{E}\mathbf{Q} = \begin{bmatrix} \mathbf{Q}'_1\mathbf{E}\mathbf{Q}_1 & \mathbf{Q}'_1\mathbf{E}\mathbf{Q}_2 \\ \mathbf{Q}'_2\mathbf{E}\mathbf{Q}_1 & \mathbf{Q}'_2\mathbf{E}\mathbf{Q}_2 \end{bmatrix}.$$

If  $\text{sep}(\mathbf{Q}'_1\mathbf{A}\mathbf{Q}_1, \mathbf{Q}'_2\mathbf{A}\mathbf{Q}_2) = \min_{\mu \in \lambda(\mathbf{Q}'_1\mathbf{A}\mathbf{Q}_1), \nu \in \lambda(\mathbf{Q}'_2\mathbf{A}\mathbf{Q}_2)} |\mu - \nu| > 0$ , where  $\lambda(\mathbf{M})$  denotes the set of eigenvalues of matrix  $\mathbf{M}$ , and

$$\|\mathbf{E}\|_{\text{F}} \leq \frac{1}{5} \text{sep}(\mathbf{Q}'_1\mathbf{A}\mathbf{Q}_1, \mathbf{Q}'_2\mathbf{A}\mathbf{Q}_2),$$

then there exists a matrix  $\mathbf{P} \in \mathbb{R}^{(n-r) \times r}$  with

$$\|\mathbf{P}\|_{\text{F}} \leq \frac{4}{\text{sep}(\mathbf{Q}'_1\mathbf{A}\mathbf{Q}_1, \mathbf{Q}'_2\mathbf{A}\mathbf{Q}_2)} \|\mathbf{Q}'_1\mathbf{E}\mathbf{Q}_2\|_{\text{F}}$$

such that the columns of  $\widehat{\mathbf{Q}}_1 = (\mathbf{Q}_1 + \mathbf{Q}_2\mathbf{P})(\mathbf{I}_r + \mathbf{P}'\mathbf{P})^{-1/2}$  define an orthonormal basis for a subspace that is invariant for  $\mathbf{A} + \mathbf{E}$ .

**Proof of Theorem 1.** From (2.2), we have the identity

$$\widehat{\Sigma}_{\mathbf{x}} = \mathbf{B}\widehat{\Sigma}_{\mathbf{f}}\mathbf{B}' + \mathbf{B}\widehat{\Sigma}_{\mathbf{f}\varepsilon} + \widehat{\Sigma}_{\varepsilon\mathbf{f}}\mathbf{B}' + \widehat{\Sigma}_{\varepsilon},$$

where  $\widehat{\Sigma}_{\mathbf{x}} = T^{-1} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t'$ ,  $\widehat{\Sigma}_{\mathbf{f}} = T^{-1} \sum_{t=1}^T \mathbf{f}_t \mathbf{f}_t'$ ,  $\widehat{\Sigma}_{\mathbf{f}\boldsymbol{\varepsilon}} = T^{-1} \sum_{t=1}^T \mathbf{f}_t \boldsymbol{\varepsilon}_t'$ ,  $\widehat{\Sigma}_{\boldsymbol{\varepsilon}\mathbf{f}} = T^{-1} \sum_{t=1}^T \boldsymbol{\varepsilon}_t \mathbf{f}_t'$ ,  $\widehat{\Sigma}_{\boldsymbol{\varepsilon}} = T^{-1} \sum_{t=1}^T \boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t'$ . The sample means  $\bar{\mathbf{x}}$ ,  $\bar{\mathbf{f}}$ , and  $\bar{\boldsymbol{\varepsilon}}$  are set to  $\mathbf{0}$ , because we assume the data  $\mathbf{X}_t$  are properly centered in advance.

By Assumption 4, we have

$$\begin{aligned} \|\widehat{\Sigma}_{\mathbf{x}} - \mathbf{B}\widehat{\Sigma}_{\mathbf{f}}\mathbf{B}'\|_2 &\leq 2\|\mathbf{B}\|_2\|\widehat{\Sigma}_{\mathbf{f}\boldsymbol{\varepsilon}}\|_2 + \|\widehat{\Sigma}_{\boldsymbol{\varepsilon}} - \Sigma_{\boldsymbol{\varepsilon}}\|_2 + \|\Sigma_{\boldsymbol{\varepsilon}}\|_2 \\ &= O_p(N) + O_p(NT^{-1/2}) + O_p(1) \\ &= O_p(N), \end{aligned}$$

Because  $[\mathbf{B} \ \mathbf{B}_c]$  is an orthogonal matrix, we have

$$\begin{aligned} \begin{bmatrix} \mathbf{B}' \\ \mathbf{B}_c' \end{bmatrix} (\mathbf{B}\widehat{\Sigma}_{\mathbf{f}}\mathbf{B}') [\mathbf{B} \ \mathbf{B}_c] &= \begin{bmatrix} \widehat{\Sigma}_{\mathbf{f}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \\ \begin{bmatrix} \mathbf{B}' \\ \mathbf{B}_c' \end{bmatrix} (\widehat{\Sigma}_{\mathbf{x}} - \mathbf{B}\widehat{\Sigma}_{\mathbf{f}}\mathbf{B}') [\mathbf{B} \ \mathbf{B}_c] &= \begin{bmatrix} \mathbf{B}'\widehat{\Sigma}_{\mathbf{x}}\mathbf{B} - \widehat{\Sigma}_{\mathbf{f}} & \mathbf{B}'\widehat{\Sigma}_{\mathbf{x}}\mathbf{B}_c \\ \mathbf{B}_c'\widehat{\Sigma}_{\mathbf{x}}\mathbf{B} & \mathbf{B}_c'\widehat{\Sigma}_{\mathbf{x}}\mathbf{B}_c \end{bmatrix}. \end{aligned}$$

By Theorem 1 in Peña and Poncela (2006) and Assumptions 3 and 4, we can show that  $\text{sep}(\widehat{\Sigma}_{\mathbf{f}}) = \lambda_r(\widehat{\Sigma}_{\mathbf{f}}) > CNT$  almost surely for some constant  $C > 0$ . By Lemma A1, there is an  $(N - r) \times r$  matrix  $\mathbf{P}$  with an upper bounded  $\|\mathbf{P}\|_2$ , such that  $\widehat{\mathbf{B}} = (\mathbf{B} + \mathbf{B}_c\mathbf{P})(\mathbf{I}_r + \mathbf{P}'\mathbf{P})^{-1/2}$  defines an orthonormal basis for a subspace that is invariant for  $\widehat{\Sigma}_{\mathbf{x}}$ . Therefore,

$$\begin{aligned} \|\widehat{\mathbf{B}} - \mathbf{B}\|_2 &\leq \|\mathbf{B}(\mathbf{I}_r - (\mathbf{I}_r + \mathbf{P}'\mathbf{P})^{-1/2})\|_2 + \|\mathbf{B}_c\mathbf{P}(\mathbf{I}_r + \mathbf{P}'\mathbf{P})^{-1/2}\|_2 \\ &\leq 2\|\mathbf{P}\|_2 \\ &\leq \frac{8}{\lambda_r(\widehat{\Sigma}_{\mathbf{f}})} \|\mathbf{B}'(\widehat{\Sigma}_{\mathbf{x}} - \mathbf{B}\widehat{\Sigma}_{\mathbf{f}}\mathbf{B}')\|_2 \\ &= O_p(T^{-1}). \end{aligned}$$

A similar result also holds for  $\|\widehat{\mathbf{B}}_c - \mathbf{B}_c\|_2$  as we can exchange the position of  $\mathbf{B}$  and  $\mathbf{B}_c$  in the orthogonal matrix  $[\mathbf{B}, \mathbf{B}_c]$ , and apply the same argument as above again.

Consequently, we have

$$\begin{aligned}
\|\mathbf{B}\mathbf{B}' - \widehat{\mathbf{B}}\widehat{\mathbf{B}}'\|_2 &= \|(\mathbf{B} - \widehat{\mathbf{B}})\mathbf{B}' + \widehat{\mathbf{B}}(\mathbf{B} - \widehat{\mathbf{B}})'\|_2 \\
&\leq \|\mathbf{B} - \widehat{\mathbf{B}}\|_2 (\|\mathbf{B}\|_2 + \|\widehat{\mathbf{B}}\|_2) \\
&= O_p(T^{-1}).
\end{aligned}$$

Furthermore, from a least-squares perspective in (2.6), the factors are estimated as  $\widehat{\mathbf{f}}_t = \widehat{\mathbf{B}}'\mathbf{x}_t = \widehat{\mathbf{B}}'(\mathbf{B}\mathbf{f}_t + \boldsymbol{\varepsilon}_t)$ . Then,

$$\begin{aligned}
\|\widehat{\mathbf{B}}\widehat{\mathbf{f}}_t - \mathbf{B}\mathbf{f}_t\|_2 &= \|\widehat{\mathbf{B}}\widehat{\mathbf{B}}'\mathbf{B}\mathbf{f}_t + \widehat{\mathbf{B}}\widehat{\mathbf{B}}'\boldsymbol{\varepsilon}_t - \mathbf{B}\mathbf{f}_t\|_2 \\
&\leq \|\widehat{\mathbf{B}}\widehat{\mathbf{B}}'(\mathbf{B} - \widehat{\mathbf{B}})\mathbf{f}_t\|_2 + \|(\widehat{\mathbf{B}} - \mathbf{B})\mathbf{f}_t\|_2 + \|\widehat{\mathbf{B}}\widehat{\mathbf{B}}'\boldsymbol{\varepsilon}_t\|_2 \\
&\leq 2\|(\widehat{\mathbf{B}} - \mathbf{B})\mathbf{f}_t\|_2 + \|\widehat{\mathbf{B}}'\boldsymbol{\varepsilon}_t\|_2 \\
&= O_p(\sqrt{N/T}) + O_p(1),
\end{aligned}$$

where the last line follows from the fact that  $\|\mathbf{f}_t\|_2 = \|\sum_{s=1}^t \mathbf{u}_s\|_2 = O_p(\sqrt{NT})$  and  $\widehat{\mathbf{B}}'\boldsymbol{\varepsilon}_t$  is an  $r$ -dimensional random vector with finite variance. Therefore,

$$N^{-1/2}\|\widehat{\mathbf{B}}\widehat{\mathbf{f}}_t - \mathbf{B}\mathbf{f}_t\|_2 = O_p(N^{-1/2} + T^{-1/2}).$$

This completes the proof. ■

**Proof of Theorem 2.** For any column vector  $\widehat{\mathbf{b}}$  in  $\widehat{\mathbf{B}}_c$ , denote its corresponding true one by  $\mathbf{b}$ . Then,

$$\widehat{\mathbf{b}}'\mathbf{X}_t = \widehat{\mathbf{b}}'\mathbf{B}\mathbf{f}_t + \widehat{\mathbf{b}}'\boldsymbol{\varepsilon}_t = (\widehat{\mathbf{b}} - \mathbf{b})'\mathbf{B}\mathbf{f}_t + (\widehat{\mathbf{b}} - \mathbf{b})'\boldsymbol{\varepsilon}_t + \mathbf{b}'\boldsymbol{\varepsilon}_t.$$

By a similar argument as the proof of theorem 4 in Gao and Tsay (2021), the autocorrelations of  $\widehat{\mathbf{b}}'\mathbf{X}_t$  will only depend on those of the third terms if the magnitudes of the first two terms are asymptotically negligible. Therefore, we only need to show

$$\max_{1 \leq t \leq T} |(\widehat{\mathbf{b}} - \mathbf{b})'\mathbf{B}\mathbf{f}_t| = o_p(1),$$

as the second term is obviously dominated by the third one according to Theorem 1. Note

that  $\mathbf{f}_t$  has an additional strength of order  $\sqrt{N}$ , and Model (2.3) implies that

$$\mathbf{f}_t = \sum_{i=1}^t \mathbf{u}_i,$$

is a partial sum of weakly dependent variables. Under Assumptions 1–4, we see that the conditions for Theorem 1 of Merlevède, Peligrad and Rio (2011) hold. By aforementioned Theorem 1 or the proof of Theorem 2 in Gao and Tsay (2021), we can show that

$$\begin{aligned} \max_{1 \leq t \leq T} |(\widehat{\mathbf{b}} - \mathbf{b})' \mathbf{B} \mathbf{f}_t| &\leq \|\widehat{\mathbf{b}} - \mathbf{b}\|_2 \max_{1 \leq t \leq T} \|\mathbf{B} \mathbf{f}_t\|_2 \\ &\leq CT^{-1} \sqrt{NT}^{1/2} \log(T) \\ &\leq CN^{1/2} \log(T) T^{-1/2}. \end{aligned}$$

Therefore, it suffices to require  $N^{1/2} \log(T) T^{-1/2} = o(1)$ . This completes the proof.  $\blacksquare$

## Appendix B Proof of Theorem 3

Recall the SVD of any  $m \times n$  matrix  $\Theta$  in (3.5), the subspaces in (3.6) and the decomposition of any  $m \times n$  matrix  $\mathbf{M}$  in (3.7). When  $\mathbf{M} = \Theta$ , we obviously have  $\|\mathbf{M}\|_* = \|\mathbf{M}_1\|_* + \|\mathbf{M}_2\|_*$ . In addition, for the decomposition of a general matrix  $\mathbf{M}$  that may be different from  $\Theta$ , we introduce the following lemma.

**Lemma A2.** *Given the SVD of  $\Theta$  in (3.5), for any matrix  $\mathbf{M} \in \mathbb{R}^{m \times n}$ , the decomposition*

$$\|\Theta_1 + \mathbf{M}_2\|_* = \|\Theta_1\|_* + \|\mathbf{M}_2\|_*$$

*holds.*

**Proof.** Given the SVD of  $\Theta$  in (3.5), we have

$$\begin{aligned} \|\Theta_1 + \mathbf{M}_2\|_* &= \left\| \begin{bmatrix} \mathbf{D}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{U}'_{k,c} \mathbf{M} \mathbf{V}_{k,c} \end{bmatrix} \right\|_* \\ &= \left\| \begin{bmatrix} \mathbf{D}_k \end{bmatrix} \right\|_* + \left\| \begin{bmatrix} \mathbf{U}'_{k,c} \mathbf{M} \mathbf{V}_{k,c} \end{bmatrix} \right\|_* \\ &= \|\Theta_1\|_* + \|\mathbf{M}_2\|_*. \end{aligned} \quad \blacksquare$$

**Lemma A3.** *Let the conditions of Theorem 3 hold. As  $N, p, T \rightarrow \infty$ ,*

$$\frac{1}{2T} \|\Delta_{\mathbf{A}} \widehat{\mathbf{Z}} + \Delta_{\Phi} \mathbf{P}\|_{\mathbb{F}}^2 \geq C_1 \kappa_1 \|\Delta\|_{\mathbb{F}}^2 - C_2 \tau_T \Psi^2(\Delta),$$

*with probability tending to 1, where  $\Psi$  is defined in (3.2) and  $\Delta = [\Delta_{\mathbf{A}}, \Delta_{\Phi}]$ .*

**Proof.** Note that

$$\begin{aligned} \|\Delta_{\mathbf{A}} \mathbf{Z} + \Delta_{\Phi} \mathbf{P}\|_{\mathbb{F}}^2 &= \|\Delta_{\mathbf{A}} \widehat{\mathbf{Z}} + \Delta_{\Phi} \mathbf{P} + \Delta_{\mathbf{A}} (\mathbf{Z} - \widehat{\mathbf{Z}})\|_{\mathbb{F}}^2 \\ &\leq 2 \left( \|\Delta_{\mathbf{A}} \widehat{\mathbf{Z}} + \Delta_{\Phi} \mathbf{P}\|_{\mathbb{F}}^2 + \|\Delta_{\mathbf{A}} (\mathbf{Z} - \widehat{\mathbf{Z}})\|_{\mathbb{F}}^2 \right), \end{aligned}$$

we have

$$\frac{1}{2T} \|\Delta_{\mathbf{A}} \widehat{\mathbf{Z}} + \Delta_{\Phi} \mathbf{P}\|_{\mathbb{F}}^2 \geq \frac{1}{4T} \|\Delta_{\mathbf{A}} \mathbf{Z} + \Delta_{\Phi} \mathbf{P}\|_{\mathbb{F}}^2 - \frac{1}{2T} \|\Delta_{\mathbf{A}} (\mathbf{Z} - \widehat{\mathbf{Z}})\|_{\mathbb{F}}^2. \quad (\text{B.1})$$

We consider the last term on the right-hand side of the above inequality,

$$\frac{1}{2T} \|\Delta_{\mathbf{A}} (\mathbf{Z} - \widehat{\mathbf{Z}})\|_{\mathbb{F}}^2 \leq \frac{1}{2T} \|\Delta_{\mathbf{A}}\|_{\mathbb{F}}^2 \|\mathbf{Z} - \widehat{\mathbf{Z}}\|_2^2 = \frac{1}{2T} \|\Delta_{\mathbf{A}}\|_{\mathbb{F}}^2 \|(\mathbf{B}_c - \widehat{\mathbf{B}}_c)' \mathbf{X}\|_2^2,$$

where we used the inequality  $\|\mathbf{MN}\|_{\mathbb{F}} \leq \|\mathbf{M}\|_2 \|\mathbf{N}\|_{\mathbb{F}}$ . Notice that

$$\|\mathbf{X}\|_{\mathbb{F}}^2 = \sum_{t=1}^T \text{tr}(\mathbf{x}_t \mathbf{x}_t') = \sum_{t=1}^T (\mathbf{f}_t' \mathbf{f}_t + 2\mathbf{f}_t' \mathbf{B}' \boldsymbol{\varepsilon}_t + \boldsymbol{\varepsilon}_t' \boldsymbol{\varepsilon}_t).$$

By Assumptions 2 and 4, we have  $\sum_{t=1}^T (\mathbf{f}_t' \mathbf{f}_t) = O_p(NT^2)$ ,  $\sum_{t=1}^T (2\mathbf{f}_t' \mathbf{B}' \boldsymbol{\varepsilon}_t) = O_p(NT)$ , and  $\sum_{t=1}^T (\boldsymbol{\varepsilon}_t' \boldsymbol{\varepsilon}_t) = O_p(N\sqrt{T})$ . Then, by the results in Theorem 1,

$$\frac{1}{2T} \|\Delta_{\mathbf{A}} (\mathbf{Z} - \widehat{\mathbf{Z}})\|_{\mathbb{F}}^2 \leq \|\Delta_{\mathbf{A}}\|_{\mathbb{F}}^2 \|\mathbf{B}_c - \widehat{\mathbf{B}}_c\|_2^2 \frac{1}{2T} \|\mathbf{X}\|_{\mathbb{F}}^2 = \|\Delta_{\mathbf{A}}\|_{\mathbb{F}}^2 O_p\left(\frac{N}{T}\right) = o_p(\|\Delta_{\mathbf{A}}\|_{\mathbb{F}}^2),$$

under the assumption that  $N/T \rightarrow 0$ . It follows from (B.1) and the above rate that

$$\frac{1}{2T} \|\Delta_{\mathbf{A}} \widehat{\mathbf{Z}} + \Delta_{\Phi} \mathbf{P}\|_{\mathbb{F}}^2 \geq \frac{1}{4T} \|\Delta_{\mathbf{A}} \mathbf{Z} + \Delta_{\Phi} \mathbf{P}\|_{\mathbb{F}}^2 - o_p(\|\Delta_{\mathbf{A}}\|_{\mathbb{F}}^2).$$

By the RSC condition specified in Theorem 3, we have

$$\begin{aligned}
\frac{1}{2T} \|\Delta_{\mathbf{A}} \widehat{\mathbf{Z}} + \Delta_{\Phi} \mathbf{P}\|_{\mathbb{F}}^2 &\geq \frac{1}{4T} \|\Delta_{\mathbf{A}} \mathbf{Z} + \Delta_{\Phi} \mathbf{P}\|_{\mathbb{F}}^2 - o_p(\|\Delta_{\mathbf{A}}\|_{\mathbb{F}}^2) \\
&\geq C_1 \kappa_1 \|\Delta\|_{\mathbb{F}}^2 - C_2 \tau_T \Psi^2(\Delta) - o_p(\|\Delta_{\mathbf{A}}\|_{\mathbb{F}}^2) \\
&\geq C_1 \kappa_1 \|\Delta\|_{\mathbb{F}}^2 - C_2 \tau_T \Psi^2(\Delta),
\end{aligned}$$

where we assume  $\kappa_1 > 0$  in the last step. This completes the proof.  $\blacksquare$

**Proof of Theorem 3.** Let

$$Loss_1(\mathbf{A}, \Phi) = \frac{1}{2T} \|\mathbf{Y} - \mathbf{A} \widehat{\mathbf{Z}} - \Phi \mathbf{P}\|_{\mathbb{F}}^2 + \lambda_{\mathbf{A}} \|\mathbf{A}\|_* + \lambda_{\Phi} \|\text{vec}(\Phi)\|_1$$

be the loss function defined in (2.7). Because the solutions  $\widehat{\mathbf{A}}$  and  $\widehat{\Phi}$  are obtained by solving the minimization problem

$$\widehat{\mathbf{A}}, \widehat{\Phi} = \arg \min_{\mathbf{A}, \Phi} Loss_1(\mathbf{A}, \Phi),$$

implying that

$$Loss_1(\widehat{\mathbf{A}}, \widehat{\Phi}) \leq Loss_1(\mathbf{A}, \Phi),$$

where  $\mathbf{A}$  and  $\Phi$  are the corresponding true ones. Recall that  $\Delta_{\mathbf{A}} = \widehat{\mathbf{A}} - \mathbf{A}$  and  $\Delta_{\Phi} = \widehat{\Phi} - \Phi$ , it follows from the above inequality that

$$\begin{aligned}
\frac{1}{2T} \|\Delta_{\mathbf{A}} \widehat{\mathbf{Z}} + \Delta_{\Phi} \mathbf{P} + \mathbf{A}(\widehat{\mathbf{Z}} - \mathbf{Z})\|_{\mathbb{F}}^2 &\leq \frac{1}{T} \langle \mathbf{E}, \Delta_{\mathbf{A}} \widehat{\mathbf{Z}} + \Delta_{\Phi} \mathbf{P} \rangle + \frac{1}{2T} \|\mathbf{A}(\widehat{\mathbf{Z}} - \mathbf{Z})\|_{\mathbb{F}}^2 \\
&\quad + \lambda_{\mathbf{A}} (\|\mathbf{A}\|_* - \|\mathbf{A} + \Delta_{\mathbf{A}}\|_*) \\
&\quad + \lambda_{\Phi} (\|\text{vec}(\Phi)\|_1 - \|\text{vec}(\Phi + \Delta_{\Phi})\|_1) \\
&= \frac{1}{T} \langle \mathbf{E}, \Delta_{\mathbf{A}} \mathbf{Z} + \Delta_{\Phi} \mathbf{P} \rangle + \frac{1}{T} \langle \mathbf{E}, \Delta_{\mathbf{A}}(\widehat{\mathbf{Z}} - \mathbf{Z}) \rangle \\
&\quad + \frac{1}{2T} \|\mathbf{A}(\widehat{\mathbf{Z}} - \mathbf{Z})\|_{\mathbb{F}}^2 + \lambda_{\mathbf{A}} (\|\mathbf{A}\|_* - \|\mathbf{A} + \Delta_{\mathbf{A}}\|_*) \\
&\quad + \lambda_{\Phi} (\|\text{vec}(\Phi)\|_1 - \|\text{vec}(\Phi + \Delta_{\Phi})\|_1).
\end{aligned} \tag{B.2}$$

Notice that the second term of the right-hand side of (B.2) satisfies

$$\begin{aligned}
\frac{1}{T} \langle \mathbf{E}, \Delta_{\mathbf{A}}(\widehat{\mathbf{Z}} - \mathbf{Z}) \rangle &\leq \frac{1}{T} \|\Delta_{\mathbf{A}}\|_* \|(\widehat{\mathbf{Z}} - \mathbf{Z})\mathbf{E}'\|_2 \\
&\leq \frac{1}{T} \|\Delta_{\mathbf{A}}\|_* \|(\widehat{\mathbf{B}}_c - \mathbf{B}_c)'\mathbf{B}\|_2 \left\| \sum_{t=1}^T \mathbf{f}_t \mathbf{e}'_t \right\|_2 \\
&\quad + \frac{1}{T} \|\Delta_{\mathbf{A}}\|_* \|\widehat{\mathbf{B}}_c - \mathbf{B}_c\|_2 \left\| \sum_{t=1}^T \varepsilon_t \mathbf{e}'_t \right\|_2 \\
&= O_p \left( \sqrt{\frac{p}{T^3}} + \sqrt{\frac{pN}{T^3}} \right) \|\Delta_{\mathbf{A}}\|_* \\
&= o_p(1) \|\Delta_{\mathbf{A}}\|_*,
\end{aligned}$$

and under the assumption of  $\|\mathbf{A}\|_2 = O_p(1)$ , the third term in the right-hand side of (B.2) satisfies

$$\frac{1}{2T} \|\mathbf{A}(\widehat{\mathbf{Z}} - \mathbf{Z})\|_{\mathbb{F}}^2 \leq \frac{1}{2T} \|\mathbf{A}\|_2^2 \|\widehat{\mathbf{B}}_c - \mathbf{B}_c\|_2^2 \|\mathbf{X}\|_{\mathbb{F}}^2 = O_p \left( \frac{N}{T} \right) \|\mathbf{A}\|_2^2 = o_p(1). \quad (\text{B.3})$$

Therefore, with probability tending to 1, (B.2) implies that

$$\begin{aligned}
&\frac{1}{2T} \|\Delta_{\mathbf{A}}\widehat{\mathbf{Z}} + \Delta_{\Phi}\mathbf{P} + \mathbf{A}(\widehat{\mathbf{Z}} - \mathbf{Z})\|_{\mathbb{F}}^2 \\
&\leq \frac{1}{T} \|\Delta_{\mathbf{A}}\|_* \|\mathbf{E}\mathbf{Z}'\|_2 + \frac{1}{T} \|\text{vec}(\Delta_{\Phi})\|_1 \|\text{vec}(\mathbf{E}\mathbf{P}')\|_{\infty} + o_p(1) \|\Delta_{\mathbf{A}}\|_* + o_p(1) \\
&\quad + \lambda_{\mathbf{A}} (\|\mathbf{A}\|_* - \|\mathbf{A} + \Delta_{\mathbf{A}}\|_*) + \lambda_{\Phi} (\|\text{vec}(\Phi)\|_1 - \|\text{vec}(\Phi) + \Delta_{\Phi}\|_1) \quad (\text{B.4}) \\
&\leq \frac{1}{2} \lambda_{\mathbf{A}} (\|\Delta_{\mathbf{A}}\|_* + 2\|\mathbf{A}\|_* - 2\|\Delta_{\mathbf{A}} + \mathbf{A}\|_*) \\
&\quad + \frac{1}{2} \lambda_{\Phi} (\|\text{vec}(\Delta_{\Phi})\|_1 + 2\|\text{vec}(\Phi)\|_1 - 2\|\text{vec}(\Delta_{\Phi} + \Phi)\|_1),
\end{aligned}$$

where we use the condition that  $\lambda_{\mathbf{A}} \geq \frac{3}{T} \|\mathbf{E}\mathbf{Z}'\|_2$  and  $\lambda_{\Phi} \geq \frac{2}{T} \|\text{vec}(\mathbf{E}\mathbf{P}')\|_{\infty}$  in the second inequality.

Let  $\Delta_{\mathbf{A},2} = \Pi_{\mathcal{S}_{\mathbf{A}}^{\perp}(r_{\mathbf{A}})}(\Delta_{\mathbf{A}})$  be the projection of  $\Delta_{\mathbf{A}}$  onto  $\mathcal{S}_{\mathbf{A}}^{\perp}(r_{\mathbf{A}})$ , where  $r_{\mathbf{A}} = \text{rank}(\mathbf{A})$ . Then we have the decomposition  $\Delta_{\mathbf{A}} = \Delta_{\mathbf{A},1} + \Delta_{\mathbf{A},2}$ , as discussed at the beginning of the Section B. Similarly, we can decompose  $\mathbf{A}$  as  $\mathbf{A} = \mathbf{A}_1 + \mathbf{A}_2$ , where  $\mathbf{A}_2 = \Pi_{\mathcal{S}_{\mathbf{A}}^{\perp}(r_{\mathbf{A}})}(\mathbf{A}) = \mathbf{0}$ .

Then,

$$\begin{aligned}
& \|\Delta_{\mathbf{A}}\|_* + 2\|\mathbf{A}\|_* - 2\|\Delta_{\mathbf{A}} + \mathbf{A}\|_* \\
&= \|\Delta_{\mathbf{A},1} + \Delta_{\mathbf{A},2}\|_* + 2\|\mathbf{A}_1\|_* - 2\|\Delta_{\mathbf{A},1} + \Delta_{\mathbf{A},2} + \mathbf{A}_1\|_* \\
&\leq \|\Delta_{\mathbf{A},1}\|_* + \|\Delta_{\mathbf{A},2}\|_* + 2\|\mathbf{A}_1\|_* - 2\|\Delta_{\mathbf{A},2} + \mathbf{A}_1\|_* + 2\|\Delta_{\mathbf{A},1}\|_* \\
&= 3\|\Delta_{\mathbf{A},1}\|_* - \|\Delta_{\mathbf{A},2}\|_*,
\end{aligned} \tag{B.5}$$

where the last line comes from Lemma A2.

Let  $\mathcal{S}_{\Phi}$  be the support set of  $\Phi$ , that is, the set of the indexes of the nonzero elements in  $\Phi$ , and  $s_{\Phi} = \text{card}(\mathcal{S}_{\Phi})$ . For  $\Delta_{\Phi}$ , use the decomposition  $\Delta_{\Phi} = \Delta_{\Phi,1} + \Delta_{\Phi,2}$ , where the entries of  $\Delta_{\Phi,1}$  can only be non-zero in the positions in  $\mathcal{S}_{\Phi}$ , and the entries of  $\Delta_{\Phi,2}$  can only be non-zero in the complement set of  $\mathcal{S}_{\Phi}$ . Similarly, we can decompose  $\Phi$  as  $\Phi = \Phi_1 + \Phi_2$ , and it is not hard to see that  $\Phi_2 = \mathbf{0}$ .

By a similar argument as (B.5), we have

$$\begin{aligned}
& \|\text{vec}(\Delta_{\Phi})\|_1 + 2\|\text{vec}(\Phi)\|_1 - 2\|\text{vec}(\Delta_{\Phi} + \Phi)\|_1 \\
&= \|\text{vec}(\Delta_{\Phi,1}) + \text{vec}(\Delta_{\Phi,2})\|_1 + 2\|\text{vec}(\Phi_1)\|_1 - 2\|\text{vec}(\Delta_{\Phi,1} + \Delta_{\Phi,2} + \Phi_1)\|_1 \\
&\leq \|\text{vec}(\Delta_{\Phi,1})\|_1 + \|\text{vec}(\Delta_{\Phi,2})\|_1 + 2\|\text{vec}(\Phi_1)\|_1 - 2\|\text{vec}(\Delta_{\Phi,2} + \Phi_1)\|_1 \\
&\quad + 2\|\text{vec}(\Delta_{\Phi,1})\|_1 \\
&= 3\|\text{vec}(\Delta_{\Phi,1})\|_1 - \|\text{vec}(\Delta_{\Phi,2})\|_1.
\end{aligned} \tag{B.6}$$

By (B.5) and (B.6), the right-hand side of (B.4) can be upper bounded by

$$\frac{1}{2}\lambda_{\mathbf{A}} (3\|\Delta_{\mathbf{A},1}\|_* - \|\Delta_{\mathbf{A},2}\|_*) + \frac{1}{2}\lambda_{\Phi} (3\|\text{vec}(\Delta_{\Phi,1})\|_1 - \|\text{vec}(\Delta_{\Phi,2})\|_1), \tag{B.7}$$

which also implies that

$$\frac{1}{2}\lambda_{\mathbf{A}} (3\|\Delta_{\mathbf{A},1}\|_* - \|\Delta_{\mathbf{A},2}\|_*) + \frac{1}{2}\lambda_{\Phi} (3\|\text{vec}(\Delta_{\Phi,1})\|_1 - \|\text{vec}(\Delta_{\Phi,2})\|_1) \geq 0.$$

Now we turn to the left-hand side of (B.2). Notice that

$$\frac{1}{2T}\|\Delta_{\mathbf{A}}\widehat{\mathbf{Z}} + \Delta_{\Phi}\mathbf{P} + \mathbf{A}(\widehat{\mathbf{Z}} - \mathbf{Z})\|_{\mathbb{F}}^2 \geq \frac{1}{4T}\|\Delta_{\mathbf{A}}\widehat{\mathbf{Z}} + \Delta_{\Phi}\mathbf{P}\|_{\mathbb{F}}^2 - \frac{1}{2T}\|\mathbf{A}(\widehat{\mathbf{Z}} - \mathbf{Z})\|_{\mathbb{F}}^2,$$

it follows from B.3 and Lemma A3 that

$$\begin{aligned} \frac{1}{2T} \|\Delta_{\mathbf{A}} \widehat{\mathbf{Z}} + \Delta_{\Phi} \mathbf{P} + \mathbf{A}(\widehat{\mathbf{Z}} - \mathbf{Z})\|_{\mathbb{F}}^2 &\geq C_1 \kappa_1 (\|\Delta_{\mathbf{A}}\|_{\mathbb{F}}^2 + \|\Delta_{\Phi}\|_{\mathbb{F}}^2) \\ &\quad - C_2 (\lambda_{\mathbf{A}} \|\Delta_{\mathbf{A}}\|_* + \lambda_{\Phi} \|\text{vec}(\Delta_{\Phi})\|_1)^2, \end{aligned} \quad (\text{B.8})$$

where the last term on the right-hand side in the parentheses satisfies

$$\begin{aligned} \lambda_{\mathbf{A}} \|\Delta_{\mathbf{A}}\|_* + \lambda_{\Phi} \|\text{vec}(\Delta_{\Phi})\|_1 &\leq \lambda_{\mathbf{A}} (\|\Delta_{\mathbf{A},1}\|_* + \|\Delta_{\mathbf{A},2}\|_*) \\ &\quad + \lambda_{\Phi} (\|\text{vec}(\Delta_{\Phi,1})\|_1 + \|\text{vec}(\Delta_{\Phi,2})\|_1) \\ &\leq 4(\lambda_{\mathbf{A}} \|\Delta_{\mathbf{A},1}\|_* + \lambda_{\Phi} \|\text{vec}(\Delta_{\Phi,1})\|_1) \\ &\leq 4(\lambda_{\mathbf{A}} \sqrt{2r_{\mathbf{A}}} \|\Delta_{\mathbf{A},1}\|_{\mathbb{F}} + \lambda_{\Phi} \sqrt{s_{\Phi}} \|\Delta_{\Phi,1}\|_{\mathbb{F}}) \\ &\leq 4(\lambda_{\mathbf{A}} \sqrt{2r_{\mathbf{A}}} \|\Delta_{\mathbf{A}}\|_{\mathbb{F}} + \lambda_{\Phi} \sqrt{s_{\Phi}} \|\Delta_{\Phi}\|_{\mathbb{F}}), \end{aligned}$$

where we use the inequality  $\text{rank}(\Delta_{\mathbf{A},1}) \leq 2r_{\mathbf{A}}$ . Under the assumptions that  $\kappa_1 \geq C_0 \lambda_{\mathbf{A}}^2 r_{\mathbf{A}} \tau_T$  and  $\kappa_1 \geq C_0 \lambda_{\Phi}^2 s_{\Phi} \tau_T$  for some  $C_0 > 0$ , it follows that

$$\tau_T \Psi^2(\Delta) \leq \frac{1}{4} C_0 \tau_T (r_{\mathbf{A}} \lambda_{\mathbf{A}}^2 \|\Delta_{\mathbf{A}}\|_{\mathbb{F}}^2 + s_{\Phi} \lambda_{\Phi}^2 \|\Delta_{\Phi}\|_{\mathbb{F}}^2) \leq \frac{1}{4} \kappa_1 (\|\Delta_{\mathbf{A}}\|_{\mathbb{F}}^2 + \|\Delta_{\Phi}\|_{\mathbb{F}}^2).$$

Therefore, (B.8) implies that

$$\frac{1}{2T} \|\Delta_{\mathbf{A}} \mathbf{Z} + \Delta_{\Phi} \mathbf{P}\|_{\mathbb{F}}^2 \geq \frac{1}{4} \kappa_1 (\|\Delta_{\mathbf{A}}\|_{\mathbb{F}}^2 + \|\Delta_{\Phi}\|_{\mathbb{F}}^2).$$

By (B.7) and the above inequality,

$$\begin{aligned} \frac{1}{4} \kappa_1 (\|\Delta_{\mathbf{A}}\|_{\mathbb{F}}^2 + \|\Delta_{\Phi}\|_{\mathbb{F}}^2) &\leq \frac{1}{2} \lambda_{\mathbf{A}} (3\|\Delta_{\mathbf{A},1}\|_* - \|\Delta_{\mathbf{A},2}\|_*) \\ &\quad + \frac{1}{2} \lambda_{\Phi} (3\|\text{vec}(\Delta_{\Phi,1})\|_1 - \|\text{vec}(\Delta_{\Phi,2})\|_1) \\ &\leq \frac{3}{2} (\lambda_{\mathbf{A}} \sqrt{2r_{\mathbf{A}}} \|\Delta_{\mathbf{A}}\|_{\mathbb{F}} + \lambda_{\Phi} \sqrt{s_{\Phi}} \|\Delta_{\Phi}\|_{\mathbb{F}}) \\ &\leq \frac{3}{2} \sqrt{2r_{\mathbf{A}} \lambda_{\mathbf{A}}^2 + \lambda_{\Phi}^2 s_{\Phi}} \sqrt{\|\Delta_{\mathbf{A}}\|_{\mathbb{F}}^2 + \|\Delta_{\Phi}\|_{\mathbb{F}}^2}. \end{aligned}$$

Dividing both sides by  $\sqrt{\|\Delta_{\mathbf{A}}\|_{\mathbb{F}}^2 + \|\Delta_{\Phi}\|_{\mathbb{F}}^2}$ , we obtain

$$\|\Delta_{\mathbf{A}}\|_{\mathbb{F}}^2 + \|\Delta_{\Phi}\|_{\mathbb{F}}^2 \leq C (r_{\mathbf{A}} \lambda_{\mathbf{A}}^2 + s_{\Phi} \lambda_{\Phi}^2) / \kappa_1^2.$$

This completes the proof. ■

## Appendix C Proof of Theorem 4

We now provide a proof of Theorem 4, which is similar to the proof of Theorem 3 but with adjustments for different regularizations and different conditions.

**Proof of Theorem 4.** Let

$$Loss_2(\mathbf{A}, \Phi) = \frac{1}{2T} \|\mathbf{Y} - \mathbf{A}\hat{\mathbf{Z}} - \Phi\mathbf{P}\|_{\text{F}}^2 + \lambda_{\mathbf{A}} \|\mathbf{A}\|_* + \sum_{i=1}^d \lambda_i \|\Phi_i\|_*.$$

Then following  $Loss_2(\hat{\mathbf{A}}, \hat{\Phi}) \leq Loss_2(\mathbf{A}, \Phi)$ , we can obtain a similar result as (B.2) and (B.4), that is, with probability tending to 1,

$$\begin{aligned} \frac{1}{2T} \|\Delta_{\mathbf{A}}\hat{\mathbf{Z}} + \Delta_{\Phi}\mathbf{P} + \mathbf{A}(\hat{\mathbf{Z}} - \mathbf{Z})\|_{\text{F}}^2 &\leq \frac{1}{T} \langle \mathbf{E}, \Delta_{\mathbf{A}}\mathbf{Z} + \Delta_{\Phi}\mathbf{P} \rangle + \frac{1}{T} \langle \mathbf{E}, \Delta_{\mathbf{A}}(\hat{\mathbf{Z}} - \mathbf{Z}) \rangle \\ &\quad + \frac{1}{2T} \|\mathbf{A}(\hat{\mathbf{Z}} - \mathbf{Z})\|_{\text{F}}^2 + \lambda_{\mathbf{A}} (\|\mathbf{A}\|_* - \|\mathbf{A} + \Delta_{\mathbf{A}}\|_*) \\ &\quad + \sum_{i=1}^d \lambda_i (\|\Delta_{\Phi_i}\|_* - \|\Phi_i + \Delta_{\Phi_i}\|_*) \\ &\leq \frac{1}{T} \|\Delta_{\mathbf{A}}\|_* \|\mathbf{E}\mathbf{Z}'\|_2 + \frac{1}{T} \sum_{i=1}^d \|\Delta_{\Phi_i}\|_* \|\mathbf{E}L^i(\mathbf{Y})'\|_2 \\ &\quad + o_p(1) \|\Delta_{\mathbf{A}}\|_* + o_p(1) + \lambda_{\mathbf{A}} (\|\mathbf{A}\|_* - \|\mathbf{A} + \Delta_{\mathbf{A}}\|_*) \\ &\quad + \sum_{i=1}^d \lambda_i (\|\Delta_{\Phi_i}\|_* - \|\Phi_i + \Delta_{\Phi_i}\|_*) \\ &\leq \frac{1}{2} \lambda_{\mathbf{A}} (\|\Delta_{\mathbf{A}}\|_* + 2\|\mathbf{A}\|_* - 2\|\Delta_{\mathbf{A}} + \mathbf{A}\|_*) \\ &\quad + \frac{1}{2} \sum_{i=1}^d \lambda_i (\|\Delta_{\Phi_i}\|_* + 2\|\Phi_i\|_* - 2\|\Delta_{\Phi_i} + \Phi_i\|_*), \end{aligned} \tag{C.1}$$

where we use the condition  $\lambda_{\mathbf{A}} \geq \frac{3}{T} \|\mathbf{E}\mathbf{Z}'\|_2$  and  $\lambda_i \geq \frac{2}{T} \|\mathbf{E}L^i(\mathbf{Y})'\|_2$  in the last inequality.

Again, we decompose  $\Delta_{\mathbf{A}}$  as that in the proof of Theorem 3, and decompose  $\Delta_{\Phi_i}$  as  $\Delta_{\Phi_i} = \Delta_{\Phi_{i,1}} + \Delta_{\Phi_{i,2}}$ ,  $i = 1, \dots, d$ , where  $\Delta_{\Phi_{i,2}} = \Pi_{\mathcal{S}_{\Phi_i}^\perp(r_{\Phi_i})}(\Delta_{\Phi_i})$ . Then, by a similar

argument as (B.5), the right-hand side of the (C.1) has an upper bound

$$\frac{1}{2}\lambda_{\mathbf{A}}(3\|\Delta_{\mathbf{A},1}\|_* - \|\Delta_{\mathbf{A},2}\|_*) + \frac{1}{2}\sum_{i=1}^d\lambda_i(3\|\Delta_{\Phi_i,1}\|_* - \|\Delta_{\Phi_i,2}\|_*), \quad (\text{C.2})$$

which also implies that  $\Delta$  is in the restricted set  $\mathcal{C}$  defined in (3.8). Furthermore, by the RE condition defined in Assumption 6, the left-hand side of (C.1) has a lower bound, that is,

$$\begin{aligned} \frac{1}{2T}\|\Delta_{\mathbf{A}}\widehat{\mathbf{Z}} + \Delta_{\Phi}\mathbf{P} + \mathbf{A}(\widehat{\mathbf{Z}} - \mathbf{Z})\|_{\text{F}}^2 &\geq \frac{1}{4T}\|\Delta_{\mathbf{A}}\widehat{\mathbf{Z}} + \Delta_{\Phi}\mathbf{P}\|_{\text{F}}^2 - \frac{1}{2T}\|\mathbf{A}(\widehat{\mathbf{Z}} - \mathbf{Z})\|_{\text{F}}^2 \\ &\geq \frac{1}{8T}\|\Delta_{\mathbf{A}}\mathbf{Z} + \Delta_{\Phi}\mathbf{P}\|_{\text{F}}^2 - \frac{1}{4T}\|\Delta_{\mathbf{A}}(\widehat{\mathbf{Z}} - \mathbf{Z})\|_{\text{F}}^2 \\ &\quad - \frac{1}{2T}\|\mathbf{A}(\widehat{\mathbf{Z}} - \mathbf{Z})\|_{\text{F}}^2 \\ &\geq \frac{1}{8T}\|\Delta_{\mathbf{A}}\mathbf{Z} + \Delta_{\Phi}\mathbf{P}\|_{\text{F}}^2 - o_p(1) \\ &\geq C\kappa_2(\|\Delta_{\mathbf{A}}\|_{\text{F}}^2 + \sum_{i=1}^d\|\Delta_{\Phi_i}\|_{\text{F}}^2). \end{aligned}$$

Then by (C.2) and the above inequality,

$$\begin{aligned} C\kappa_2\left(\|\Delta_{\mathbf{A}}\|_{\text{F}}^2 + \sum_{i=1}^d\|\Delta_{\Phi_i}\|_{\text{F}}^2\right) &\leq \frac{1}{2}\lambda_{\mathbf{A}}(3\|\Delta_{\mathbf{A},1}\|_* - \|\Delta_{\mathbf{A},2}\|_*) \\ &\quad + \frac{1}{2}\sum_{i=1}^d\lambda_i(3\|\Delta_{\Phi_i,1}\|_* - \|\Delta_{\Phi_i,2}\|_*) \\ &\leq \frac{3}{2}\left(\lambda_{\mathbf{A}}\sqrt{2r_{\mathbf{A}}}\|\Delta_{\mathbf{A}}\|_{\text{F}} + \sum_{i=1}^d\lambda_i\sqrt{2r_{\Phi_i}}\|\Delta_{\Phi_i}\|_{\text{F}}\right) \\ &\leq \frac{3}{2}\sqrt{2r_{\mathbf{A}}\lambda_{\mathbf{A}}^2 + 2\sum_{i=1}^dr_{\Phi_i}\lambda_i^2}\sqrt{\|\Delta_{\mathbf{A}}\|_{\text{F}}^2 + \sum_{i=1}^d\|\Delta_{\Phi_i}\|_{\text{F}}^2}. \end{aligned}$$

Therefore,

$$\|\Delta_{\mathbf{A}}\|_{\text{F}}^2 + \sum_{i=1}^d\|\Delta_{\Phi_i}\|_{\text{F}}^2 \leq C\left(r_{\mathbf{A}}\lambda_{\mathbf{A}}^2 + \sum_{i=1}^dr_{\Phi_i}\lambda_i^2\right)/\kappa_2^2.$$

This completes the proof. ■

## References

- Gao, Z. and Tsay, R.S., 2021. Modeling high-dimensional unit-root time series. *International Journal of Forecasting*, 37(4), pp.1535–1555.
- Golub, G.H. and Van Loan, C.F., 2013. *Matrix computations*. JHU press.
- Johnstone, I.M. and Lu, A.Y., 2009. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486), pp.682–693.
- Lam, C., Yao, Q. and Bathia, N., 2011. Estimation of latent factors for high-dimensional time series. *Biometrika*, 98(4), pp.901–918.
- Merlevède, F., Peligrad, M. and Rio, E., 2011. A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, 151(3), pp.435–474.
- Peña, D. and Poncela, P., 2006. Nonstationary dynamic factor analysis. *Journal of Statistical Planning and Inference*, 136(4), pp.1237–1257.