Reduced Rank Autoregressive Models for Matrix Time Series

Han Xiao[a]*, Yuefeng Han[a], Rong Chen[a] and Chengcheng Liu[b]
[a]Rutgers University, [b]Capital University of Economics and Business

Abstract. Matrix time series, a sequence of observations in matrix form, is often observed in finance, economics, engineering and many other fields. To avoid the use of vectorization of the matrices which loses the column and row information, the Matrix Autoregressive (MAR) Model of Chen et al (2019) maintains and utilizes the matrix structure, leading to a substantial dimensional reduction and admitting explicit interpretations. However, the model still encounters difficulties in dealing with large dimensional matrix time series as the coefficient matrices in MAR models are also large. In this paper we propose to achieve further dimension reduction through reduced rank constraints of the coefficient matrices in the MAR model. The model also has a close connection to matrix factor models, but with a generative mechanism. Estimation and rank determination procedures with their theoretical properties are presented and demonstrated with simulated and real examples.

Keywords: Forecasting; Matrix time series; Rank determination; Reduced rank regression

*Rong Chen is Professor, Department of Statistics, Rutgers University, Piscataway, NJ 08854. E-mail: rongchen@stat.rutgers.edu. Yuefeng Han is Postdoctoral Fellow, Department of Statistics, Rutgers University, Piscataway, NJ 08854. Email: yuefeng.han@rutgers.edu. Chengcheng Liu is Assistant Professor, Department of Statistics, Capital University of Economics and Business, Beijing, China. Email: ccliu@cueb.edu.cn. Han Xiao is Associate Professor, Department of Statistics, Rutgers University, Piscataway, NJ 08854. E-mail: hxiao@stat.rutgers.edu. Han Xiao is the corresponding author.

1 Introduction

Observations in matrix and tensor (multi-dimensional array) forms have been generated and collected more and more abundantly in many fields including biological/medical research, economics, engineering, finance, signal processing, social sciences etc. In response to the urgent need of analytical tools for analyzing such type of data in various applications, many optimization and statistical methods/procedures have been proposed and studied. Similar to the use of matrix decomposition for analysis of vector observations, tensor decomposition and estimation methods play a principal role in analyzing matrix/tensor data (Anandkumar et al., 2014; Cichocki et al., 2015, 2009; De Lathauwer et al., 2000a,b; De Silva and Lim, 2008; Sidiropoulos et al., 2017).

In many applications, the matrices are observed through time, and hence form a matrix-valued time series. Although it is possible and perhaps convenient to treat time as another mode and apply the tensor methods to such a three-way tensor, the time dimension is intrinsically different, and the temporal dependence requires careful modeling and analysis to aid practitioners on acquiring a diagnostic understanding of the dynamics and making reliable forecasts. It has been witnessed that multilinear models can reduce the dimension and improve the estimation stability for matrix/tensor data (Ding and Dennis Cook, 2018; Raskutti and Yuan, 2015; Zhao and Leng, 2014; Zhou et al., 2013). For dependent data, there has been many recent works on factor models of matrix/tensor time series, see Chen et al. (2019b); Gao and Tsay (2020); Han et al. (2020b,c); Wang et al. (2019) among others. On the other hand, Hoff (2015) pioneered in suggesting the multilinear model for longitudinal tensor data. Chen et al. (2019a) proposed the matrix autoregressive model (MAR), which retains the matrix form of the data and specifies the autoregressive relationship through a bilinear matrix product. Besides the interpretations adherent to its matrix and bilinear form, the MAR also reduces the model complexity significantly, comparing to the approach of concatenating the matrix observation into a long vector and then using the traditional vector autoregressive (VAR) model (Hannan, 1970; Lütkepohl, 2005; Tiao and Box, 1981; Tsay, 2014).

When the matrix observations are themselves of large dimensions, the MAR model still involves a large number of parameters. It is desirable and sometimes necessary to reduce the dimension even further. In this paper, we propose the reduced rank matrix autoregressive model (RRMAR), which assumes the form of MAR, but requires in addition that the coefficient matrices have ranks smaller than their dimensions.

We note that another natural approach to reducing the model complexity is to impose sparsity on the MAR. This approach is closely related to recent works on sparse VAR models (Basu et al., 2015; Davis et al., 2016; Han et al., 2015; Kock and Callot, 2015; Lin and Michailidis, 2017; Loh and Wainwright, 2011; Melnyk and Banerjee, 2016; Nicholson et al., 2017). In addition, Basu et al. (2019) and Lin and Michailidis (2020) considered additional low rank constraints on the coefficient matrices, Hall et al. (2018) introduced the generalized VAR model, and Han et al. (2020a) focused on the nonlinear sparse VAR model. Ghosh et al. (2019) and Ghosh et al. (2020) considered the high dimensional VAR from a Bayesian perspective.

In contrast to the aforementioned works based on sparsity, the thrust of the present paper hinges upon the low rank structure of the coefficient matrices in MAR. As will be elaborated in Section 2, the low rank matrices in RRMAR continue to admit natural interpretations, and lead to a greater dimension reduction compared to MAR. It also relates to and provides a generative mechanism for the dynamic matrix factor models of Wang et al. (2019), and has a close connection to the hierarchical factor models.

The proposed model and estimation procedure are related to the reduced rank regression (Anderson, 1951; Izenman, 1975; Reinsel and Velu, 1998). We consider two estimators: one based on least squares (RR.LS) and one based on maximum likelihood (RR.CC). It is worth pointing out that they correspond to two different algorithms for vector reduced rank regression: one minimizing the trace of the sample covariance matrix, and the other the determinant, where the latter also corresponds to the canonical correlation analysis (see for example Velu and Reinsel (2013) for more details). The likelihood-based RR.CC is indeed the maximum likelihood estimator if the covariance tensor of the error matrix has the form of a tensor product of two covariance matrices. Even when this assumption does not hold, the RR.CC nevertheless can be viewed as an estimator obtained together with a regularized estimation of the covariance tensor, and can still potentially lead to superior performances over the RR.LS.

Here we shall emphasize two significant differences between our model and the classical reduced rank regression. First, the observations are in matrix form, and the model takes a bilinear form. Second, the algorithms requires running reduced rank least squares/maximum likelihood iteratively. We develop central limit theorems for the estimators of the coefficient matrices, as well as their singular vectors. The bilinear form of the matrix model also makes the analysis substantially

different from the vector case.

The estimation procedures depend on the ranks of the two coefficient matrices in the RRMAR model. We propose to use information criterion based procedures to identify the ranks of these matrices. Since two ranks are to be determined, a thorough search over all possible pairs of ranks can be very costly, so we also introduce procedures to select the two ranks separately. Asymptotic consistency of these selection procedures are established.

The rest of this article is organized as follows. The RRMAR model is introduced in Section 2, together with its basic properties, interpretations and connections with other models. In Section 3 we propose two estimators, RR.LS and RR.CC. Asymptotic distributions are provided for both of them, as well as corresponding estimators of the leading singular vectors of the coefficient matrices. The model/rank selection procedures based on information criterion are introduced in Section 5, with their consistency properties. We use an extensive numerical study and an example in finance to demonstrate the performances of the proposed models and estimators in Section 6. All the proofs are collected in the Appendix.

1.1 Notations

We gather the notations and the definitions of some special matrices in this section.

We use $\|\cdot\|_F$ to denote the Frobenius norm of a matrix, and $\rho(\cdot)$ the spectral radius. We use $\otimes$ to denote the Kronecker product, and $\circ$ the (point-wise) Hadamard product of two matrices. The notation $\mathbf{1}_k$ stands for a k-dimensional vector with all entries equal to one. For any matrix $\boldsymbol{M}$, we use $\boldsymbol{M}[i,]$ and $\boldsymbol{M}[,j]$ to denote its i-th row and j-th column respectively. The column space of $\boldsymbol{M}$ is denoted by $\mathrm{col}(\boldsymbol{M})$. The matrix vectorization, denoted by $\mathrm{vec}(\cdot)$, turns a matrix into a vector by stacking its columns.

For any positive integer p, let $\boldsymbol{e}_{p,j} \in \mathbb{R}^p$ be the j-th base vector whose j-the entry is 1, and others zero. For any two positive integers p and q, let $\boldsymbol{J}_{p,q}$ be the $(pq) \times (pq)$ permutation matrix defined as

$$\boldsymbol{J}_{p,q} = [\boldsymbol{I}_q \otimes \boldsymbol{e}_{p,1}, \boldsymbol{I}_q \otimes \boldsymbol{e}_{p,2}, \ldots, \boldsymbol{I}_q \otimes \boldsymbol{e}_{p,p}]\,. \tag{1}$$

The permutation $\boldsymbol{J}_{p,q}$ does the following: for any $p \times q$ matrix $\boldsymbol{M}$, $\boldsymbol{J}_{p,q}\mathrm{vec}(\boldsymbol{M}') = \mathrm{vec}(\boldsymbol{M})$. In other words, $\boldsymbol{J}_{p,q}$ connects the vectorization of a matrix and its transpose.

Let $\boldsymbol{L}_p$ be the $p^2 \times p$ matrix whose j-th column is given by $\boldsymbol{e}_{p,j} \otimes \boldsymbol{e}_{p,j}$, i.e.

$$\boldsymbol{L}_p = [\boldsymbol{e}_{p,1} \otimes \boldsymbol{e}_{p,1}, \boldsymbol{e}_{p,2} \otimes \boldsymbol{e}_{p,2}, \ldots, \boldsymbol{e}_{p,p} \otimes \boldsymbol{e}_{p,p}]. \tag{2}$$

For any $p \times p$ matrix $\boldsymbol{M} = (m_{jk})$, the following operation extracts its diagonals:

$$\boldsymbol{L}_p' \text{vec}(\boldsymbol{M}) = (m_{11}, \ldots, m_{pp})',$$

and furthermore,

$$\text{vec}^{-1}[\boldsymbol{L}_p \boldsymbol{L}_p' \text{vec}(\boldsymbol{M})] = \text{diag}(\boldsymbol{M}),$$

where $\text{diag}(\boldsymbol{M})$ is the $p \times p$ diagonal matrix keeping $\boldsymbol{M}$'s diagonal elements.

2 Reduced Rank MAR Model

The *reduced rank matrix autoregressive model* (RRMAR) takes the form

$$\boldsymbol{X}_t = \boldsymbol{A}_1 \boldsymbol{X}_{t-1} \boldsymbol{A}_2' + \boldsymbol{E}_t, \tag{3}$$

where $\boldsymbol{A}_i$ are $d_i \times d_i$ autoregressive coefficient matrices of ranks $k_i \leqslant d_i$, and $\boldsymbol{E}_t \in \mathbb{R}^{d_1 \times d_2}$ is a matrix white noise. It is the same as the MAR model proposed by Chen et al. (2019a), except the additional low rank assumption that $\text{rank}(\boldsymbol{A}_i) = k_i \leqslant d_i$, for $i = 1, 2$. It is worth observing that the number of parameters to determine $\boldsymbol{A}_i$ under the rank constraint is $(2d_i - k_i)k_i$, as opposed to d_i^2 for the unconstrained $\boldsymbol{A}_i$, and the former can be much smaller if $k_i \ll d_i$. We also assume that $\|\boldsymbol{A}_1\|_F = 1$, so that $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are identified up to a sign change. To guarantee that the model (3) is causal and stationary, we require that $\rho(\boldsymbol{A}_1) \cdot \rho(\boldsymbol{A}_2) < 1$.

To better understand the implication of the low rank assumption, we write $\boldsymbol{A}_i = \boldsymbol{A}_{il} \boldsymbol{A}_{ic}'$, where $\boldsymbol{A}_{il}$ and $\boldsymbol{A}_{ic}$ are both $d_i \times k_i$ full rank matrices. The model (3) is then written as

$$\boldsymbol{X}_t = \boldsymbol{A}_{1l} \boxed{\boldsymbol{A}_{1c}' \boldsymbol{X}_{t-1} \boldsymbol{A}_{2c}} \boldsymbol{A}_{2l}' + \boldsymbol{E}_t. \tag{4}$$

The boxed part $\boldsymbol{F}_t := \boldsymbol{A}_{1c}' \boldsymbol{X}_{t-1} \boldsymbol{A}_{2c}$ is a $k_1 \times k_2$ matrix, which can be viewed as a composite and much smaller version of the $d_1 \times d_2$ matrix $\boldsymbol{X}_t$. The conditional expectation of $\boldsymbol{X}_t$ given $\boldsymbol{X}_{t-1}$ is then given by loading on $\boldsymbol{F}_t$ from left by $\boldsymbol{A}_{1l}$, and from right by $\boldsymbol{A}_{2l}'$. The RRMAR model therefore provides a generating mechanism for the matrix factor model $\boldsymbol{X}_t = \boldsymbol{A}_{l1} \boldsymbol{F}_t \boldsymbol{A}_{l2}' + \boldsymbol{E}_t$, introduced in Wang et al. (2019), in which the factor process $\boldsymbol{F}_t$ is assumed to be latent and unobserved. In the

RRMAR model in (4), $\boldsymbol{F}_t$ depends on $\boldsymbol{X}_{t-1}$ hence is observed given the parameters. Due to this connection, we call $\boldsymbol{A}_{ic}$ the composition matrix, and $\boldsymbol{A}_{il}$ the loading matrix.

The RRMAR model is also related to the hierarchical factor models in the econometrics literature (Moench et al., 2013). Specifically, let $\boldsymbol{F}_t^* = \boldsymbol{A}_{1c}'\boldsymbol{X}_{t-1}\boldsymbol{A}_2'$, then

$$\boldsymbol{X}_t = \boldsymbol{A}_{1l}\boldsymbol{F}_t^* + \boldsymbol{E}_t,$$

which means that the j-th column of $\boldsymbol{X}_t$ follows a factor model with loading $\boldsymbol{A}_{1l}$, and factors $\boldsymbol{F}_t^*[,j]$, the j-th column of $\boldsymbol{F}_t^*$. In the next layer, we have

$$\boldsymbol{F}_t^{*'} = \boldsymbol{A}_{2l}\boldsymbol{F}_t',$$

which says that the $d_2 \times k_1$ factor matrix $\boldsymbol{F}_t^{*'}$ is further driven by a smaller factor matrix $\boldsymbol{F}_t'$ (defined by the boxed part in (4)), the j-th column of $\boldsymbol{F}_t^{*'}$ corresponding to the loading $\boldsymbol{A}_{2l}$, and factors $\boldsymbol{F}_t'[,j]$. Therefore, the model (3) also gives a generating mechanism of a special instance of the hierarchical factor model.

There are many potential extensions of model (3). The first is the extension to RRMAR(p) model:

$$\boldsymbol{X}_t = \sum_{l=1}^{p} \boldsymbol{A}_{1l}\boldsymbol{X}_{t-l}\boldsymbol{A}_{2l}' + \boldsymbol{E}_t, \tag{5}$$

where all $\boldsymbol{A}_{ij}$ are of low ranks. The second extension is more subtle. Although only the lag-1 observation $\boldsymbol{X}_{t-1}$ is involved on the right hand side of (3), there can be multiple terms in the form

$$\boldsymbol{X}_t = \sum_{j=1}^{r} \boldsymbol{A}_1^{(j)}\boldsymbol{X}_{t-1}\left(\boldsymbol{A}_2^{(j)}\right)' + \boldsymbol{E}_t. \tag{6}$$

To see this extension more clearly, we take vectorization on both sides of (6):

$$\mathrm{vec}(\boldsymbol{X}_t) = \left(\sum_{j=1}^{r} \boldsymbol{A}_2^{(j)} \otimes \boldsymbol{A}_1^{(j)}\right)\mathrm{vec}(\boldsymbol{X}_{t-1}) + \boldsymbol{E}_t.$$

It is seen from the preceding equation that the RRMAR model (3) amounts to restricting the coefficient matrix of the VAR(1) model to the form of a Kronecker product, and the model (6) is more flexible by representing the coefficient matrix as a sum of r Kronecker products.

The extension to (5) is quite straightforward, so in order to fix ideas, we focus in this paper on model (3) only. On the other hand, the extension to (6) is more intricate, which we shall leave for future studies. Finally, we add that the two extensions (5) and (6) can be combined to give a more comprehensive model.

3 Estimation

Suppose a matrix time series $\{\boldsymbol{X}_t\}$ of length T is observed. To estimate the coefficient matrices $\boldsymbol{A}_i$, we propose to use the alternating reduced rank regression, updating one, while holding the other fixed. Specifically, suppose $\boldsymbol{A}_2$ is given, we discuss how to estimate $\boldsymbol{A}_1$. Recall that $\boldsymbol{A}[, j]$ denotes the j-th column of a matrix $\boldsymbol{A}$. We also make the convention that $\boldsymbol{A}'[, j]$ denotes the j-th column of $\boldsymbol{A}'$, i.e. the j-th row of $\boldsymbol{A}$ as a column vector. The j-th column of the model equation (3) is

$$\boldsymbol{X}_t[, j] = \boldsymbol{A}_1 \boxed{\boldsymbol{X}_{t-1}\boldsymbol{A}_2'[, j]} + \boldsymbol{E}_t[, j].$$

Since $\boldsymbol{A}_2$ is fixed, the preceding equation can be viewed as the reduced rank regression involving $(T-1)d_2$ sample units, where each column $\boldsymbol{X}_t[, j]$ is a response vector, the boxed vector is the covariate, and $\boldsymbol{A}_1$ is the coefficient matrix. In Section 3.1 we consider the estimation of $\boldsymbol{A}_1$ by least squares. On the other hand, under normality, the classical reduced rank regression minimizes the determinant of the sample covariance matrix of the error vectors, under the rank constraint, which is related and in fact equivalent to canonical correlation analysis (Anderson, 2003; Reinsel and Velu, 1998). In Section 3.2 we introduce a special covariance structure of $\boldsymbol{E}_t$, under which we seek to estimate $\boldsymbol{A}_1$ by the Gaussian MLE.

Note that each step of the algorithms reduces the corresponding objective functions, hence the algorithms will converge, though often converge to a local minimum. We suggest to use the projection estimator of $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ in Chen et al. (2019a), *i.e.,* without rank constraints, as the initial values of both alternating algorithms. In this section we focus on the estimation of the coefficient matrices given the ranks k_1 and k_2. The determination of the ranks will be discussed in Section 5.

3.1 Alternating least squares

The least squares method minimizes the trace of the sample covariance matrix of the residuals under the rank constraint:

$$\min_{\boldsymbol{A}_1:\, \mathrm{rank}(\boldsymbol{A}_1)=k_1} \sum_{t=2}^{T} \left\| \boldsymbol{X}_t - \boldsymbol{A}_1 \boldsymbol{X}_{t-1} \boldsymbol{A}_2' \right\|_F^2 \tag{7}$$

$$\iff \quad \min_{\boldsymbol{A}_1:\, \mathrm{rank}(\boldsymbol{A}_1)=k_1} \mathrm{tr}\left[\sum_{t=2}^{T} \sum_{j=1}^{d_2} \left(\boldsymbol{X}_t[, j] - \boldsymbol{A}_1 \boldsymbol{X}_{t-1} \boldsymbol{A}_2'[, j] \right) \left(\boldsymbol{X}_t[, j] - \boldsymbol{A}_1 \boldsymbol{X}_{t-1} \boldsymbol{A}_2'[, j] \right)' \right].$$

We denote the least squares estimator by $\hat{\boldsymbol{A}}_i^{\text{ls}}$, and will refer to it as the RR.LS estimator. Suppose $\boldsymbol{A}_2$ is given, let $\boldsymbol{S}_{xx} = \sum_t \boldsymbol{X}_{t-1}\boldsymbol{A}_2'\boldsymbol{A}_2\boldsymbol{X}_{t-1}'$, $\boldsymbol{S}_{yx} = \sum_t \boldsymbol{X}_t\boldsymbol{A}_2\boldsymbol{X}_{t-1}'$, and $\boldsymbol{U} := [U_1, U_2, \ldots, U_{k_1}]$, where U_j is the j-th leading normalized eigenvector of $\boldsymbol{S}_{yx}\boldsymbol{S}_{xx}^{-1}\boldsymbol{S}_{xy}$. Then $\boldsymbol{A}_1$ can be updated as

$$\check{\boldsymbol{A}}_1^{\text{ls}} = \boldsymbol{U}\boldsymbol{U}'\boldsymbol{S}_{yx}\boldsymbol{S}_{xx}^{-1}.$$

Given $\boldsymbol{A}_1$, an update of $\boldsymbol{A}_2$ can be similarly obtained. We therefore use the alternating least squares to find the minimizer of (7).

3.2 Alternating canonical correlation analysis

The classical reduced rank regression has also been situated under normality, leading to the Gaussian MLE of the coefficient matrix. To introduce the MLE for the RRMAR model, we need to assume that the covariance matrix Σ_e of $\text{vec}(\boldsymbol{E}_t)$ takes the form of a product

$$\Sigma_e = \Sigma_2 \otimes \Sigma_1, \tag{8}$$

where Σ_1 and Σ_2 are $d_1 \times d_1$ and $d_2 \times d_2$ positive definite matrices respectively. This assumption allows us to separate the row and column dependence within the error matrix, with Σ_1 and Σ_2 corresponding to the row-wise and column-wise correlations among the entries of $\boldsymbol{E}_t$ respectively. This type of covariance model has been proposed and studied in the literature as the "transposable" (Allen and Tibshirani, 2010), "array normal" (Hoff et al., 2011), "separable" (Tsiligkaridis and Hero, 2013; Zhou, 2014), and "Kronecker product" (Hafner et al., 2020; Linton and Tang, 2019) covariance structure. We refer the readers to Hoff et al. (2011) and Linton and Tang (2019) for a more detailed account on the history of the separable covariance matrix. Chen and Chen (2019) also considered the MAR model under this covariance structure.

Under the normality, the log likelihood of the RRMAR model is (up to some additive constants)

$$-(T-1)(d_2 \log|\Sigma_1| + d_1 \log|\Sigma_2|) - \sum_{t=2}^{T} \text{tr}\left[\Sigma_1^{-1}(\boldsymbol{X}_t - \boldsymbol{A}_1\boldsymbol{X}_{t-1}\boldsymbol{A}_2')\Sigma_2^{-1}(\boldsymbol{X}_t - \boldsymbol{A}_1\boldsymbol{X}_{t-1}\boldsymbol{A}_2')'\right]. \tag{9}$$

We will introduce an alternating algorithm to find the MLE, maximizing (9) alternatively over one pair $(\boldsymbol{A}_i, \Sigma_i)$ while holding the other fixed. As will be seen, each iteration can be viewed as a reduced rank regression, which is equivalent to the canonical correlation analysis (Reinsel and Velu, 1998). We therefore denote the minimizer of (9) by $\hat{\boldsymbol{A}}_i^{\text{cc}}$ and $\hat{\Sigma}_i$, and refer to it as the RR.CC estimator.

We now describe how to estimate $\boldsymbol{A}_1$ and Σ_1 when $\boldsymbol{A}_2$ and Σ_2 are known. Under assumption (8), we can rewrite the model as

$$\left(\boldsymbol{X}_t\Sigma_2^{-1/2}\right)[,j] = \boldsymbol{A}_1\left(\boldsymbol{X}_{t-1}\boldsymbol{A}_2'\Sigma_2^{-1/2}\right)[,j] + \left(\boldsymbol{E}_t\Sigma_2^{-1/2}\right)[,j].$$

Note that the columns of the transformed error matrix $\boldsymbol{E}_t\Sigma_2^{-1/2}$ are iid $N(\boldsymbol{0},\Sigma_1)$. If we let $\boldsymbol{y}_{tj} = \left(\boldsymbol{X}_t\Sigma_2^{-1/2}\right)[,j]$, $\boldsymbol{x}_{tj} = \left(\boldsymbol{X}_{t-1}\boldsymbol{A}_2'\Sigma_2^{-1/2}\right)[,j]$ and $\boldsymbol{\epsilon}_{tj} = \left(\boldsymbol{E}_t\Sigma_2^{-1/2}\right)[,j]$, then the preceding equation can be viewed as a reduced rank regression with i.i.d. errors:

$$\boldsymbol{y}_{tj} = \boldsymbol{A}_1\boldsymbol{x}_{tj} + \boldsymbol{\epsilon}_{tj}, \quad 2 \leqslant t \leqslant T,\ 1 \leqslant j \leqslant d_2. \tag{10}$$

The MLE of $\boldsymbol{A}_1$ based on (10) with i.i.d. normal errors has been well studied in the classical reduced rank regression. Here we only define necessary notations to introduce the final expression of the MLE. We refer the readers to the classical texts Anderson (2003) and Reinsel and Velu (1998) for more details. Let

$$\tilde{\boldsymbol{S}}_{xx} = \sum_t\sum_j \boldsymbol{x}_{tj}\boldsymbol{x}_{tj}' = \sum_t \boldsymbol{X}_{t-1}\boldsymbol{A}_2'\Sigma_2^{-1}\boldsymbol{A}_2\boldsymbol{X}_{t-1}',$$
$$\tilde{\boldsymbol{S}}_{yx} = \sum_t\sum_j \boldsymbol{y}_{tj}\boldsymbol{x}_{tj}' = \sum_t \boldsymbol{X}_t\Sigma_2^{-1}\boldsymbol{A}_2\boldsymbol{X}_{t-1}'.$$

The least squares estimator (with no rank constraint) of $\boldsymbol{A}_1$ based on (10) is then given by $\tilde{\boldsymbol{A}}_1 = \tilde{\boldsymbol{S}}_{yx}\tilde{\boldsymbol{S}}_{xx}^{-1}$. To get the MLE of $\boldsymbol{A}_1$ under the constraint $\text{rank}(\boldsymbol{A}_1) = k_1$, let

$$\tilde{\Sigma}_{\epsilon\epsilon} = \sum_t\sum_j \left(\boldsymbol{y}_{tj} - \tilde{\boldsymbol{A}}_1\boldsymbol{x}_{tj}\right)\left(\boldsymbol{y}_{tj} - \tilde{\boldsymbol{A}}_1\boldsymbol{x}_{tj}\right)' = \sum_t \left(\boldsymbol{X}_t - \tilde{\boldsymbol{A}}_1\boldsymbol{X}_{t-1}\boldsymbol{A}_2'\right)\Sigma_2^{-1}\left(\boldsymbol{X}_t - \tilde{\boldsymbol{A}}_1\boldsymbol{X}_{t-1}\boldsymbol{A}_2'\right)'.$$

Take $\tilde{\boldsymbol{U}} := [\tilde{U}_1, \tilde{U}_2, \ldots, \tilde{U}_{k_1}]$, where $\tilde{U}_j$ is the j-th leading unit eigenvector of $\tilde{\Sigma}_{\epsilon\epsilon}^{-1/2}\tilde{\boldsymbol{S}}_{yx}\tilde{\boldsymbol{S}}_{xx}^{-1}\tilde{\boldsymbol{S}}_{xy}\tilde{\Sigma}_{\epsilon\epsilon}^{-1/2}$. Then $\boldsymbol{A}_1$ is updated as

$$\check{\boldsymbol{A}}_1^{\text{cc}} = \tilde{\Sigma}_{\epsilon\epsilon}^{1/2}\boldsymbol{U}\boldsymbol{U}'\tilde{\Sigma}_{\epsilon\epsilon}^{-1/2}\tilde{\boldsymbol{S}}_{yx}\tilde{\boldsymbol{S}}_{xx}^{-1}.$$

Subsequently, the covariance matrix Σ_1 is updated as

$$\check{\Sigma}_1 = \frac{1}{T-1}\sum_t \left(\boldsymbol{X}_t - \check{\boldsymbol{A}}_1^{\text{cc}}\boldsymbol{X}_{t-1}\boldsymbol{A}_2'\right)\Sigma_2^{-1}\left(\boldsymbol{X}_t - \check{\boldsymbol{A}}_1^{\text{cc}}\boldsymbol{X}_{t-1}\boldsymbol{A}_2'\right)'.$$

Given $\boldsymbol{A}_1$ and Σ_1, an update of $\boldsymbol{A}_2$ and Σ_2 can be similarly obtained. Therefore, we use the alternating algorithm to find the minimizer of (9).

4 Asymptotics

The asymptotic analysis is substantially different from the classical reduced rank regression, due to the alternating nature of the estimation. For example, the gradient condition for the LSE $\hat{\boldsymbol{A}}_1^{\text{ls}}$ is $\hat{\boldsymbol{A}}_1^{\text{ls}} = \hat{\boldsymbol{U}}\hat{\boldsymbol{U}}'\hat{\boldsymbol{S}}_{yx}\hat{\boldsymbol{S}}_{xx}^{-1}$, where $\hat{\boldsymbol{U}}$, $\hat{\boldsymbol{S}}_{xx}$ and $\hat{\boldsymbol{S}}_{yx}$ are defined as the $\boldsymbol{U}$, $\boldsymbol{S}_{xx}$ and $\boldsymbol{S}_{yx}$ in Section 3 with the modification that all $\boldsymbol{A}_2$ therein need to be replaced by $\hat{\boldsymbol{A}}_2^{\text{ls}}$. In other words, the asymptotic behaviors of $\hat{\boldsymbol{A}}_1^{\text{ls}}$ and $\hat{\boldsymbol{A}}_2^{\text{ls}}$ are interwoven.

The asymptotics for $\boldsymbol{A}_i^{\text{ls}}$ and $\boldsymbol{A}_i^{\text{cc}}$ involve heavy notations. First of all, recall that we assume $\|\boldsymbol{A}_1\|_F = 1$ for the parameter identifiability, so we rescale $\hat{\boldsymbol{A}}_i^{\text{ls}}$ and $\hat{\boldsymbol{A}}_i^{\text{cc}}$ so that $\|\hat{\boldsymbol{A}}_1^{\text{ls}}\|_F = 1$ and $\|\hat{\boldsymbol{A}}_1^{\text{cc}}\|_F = 1$. In the table below, we list notations that appear in both Theorem 1 and Theorem 2, but with different definitions in these theorems.

Notations	Theorem 1	Theorem 2
Γ_1	$\mathbb{E}(\boldsymbol{X}_t'\boldsymbol{A}_1'\boldsymbol{A}_1\boldsymbol{X}_t)$	$\mathbb{E}(\boldsymbol{X}_t'\boldsymbol{A}_1'\Sigma_1^{-1}\boldsymbol{A}_1\boldsymbol{X}_t)$
Γ_2	$\mathbb{E}(\boldsymbol{X}_t\boldsymbol{A}_2'\boldsymbol{A}_2\boldsymbol{X}_t')$	$\mathbb{E}(\boldsymbol{X}_t\boldsymbol{A}_2'\Sigma_2^{-1}\boldsymbol{A}_2\boldsymbol{X}_t')$
$\mathbb{P}_i$	orthogonal projection to $\text{col}(\boldsymbol{A}_i)$	orthogonal projection to $\text{col}(\Sigma_i^{-1/2}\boldsymbol{A}_i)$
$\mathcal{P}_i$	$\mathbb{P}_i$	$\Sigma_i^{-1/2}\mathbb{P}_i\Sigma_i^{1/2}$
$\boldsymbol{\alpha}_1$	$\text{vec}(\boldsymbol{A}_1)$	Same
$\boldsymbol{\gamma}_1$	$(\boldsymbol{\alpha}_1', \boldsymbol{0}')'$	Same
$\boldsymbol{W}_t$	$[(\boldsymbol{A}_2\boldsymbol{X}_t') \otimes \boldsymbol{I}_{d_1}, \boldsymbol{I}_{d_2} \otimes (\boldsymbol{A}_1\boldsymbol{X}_t)]'$	Same
$\boldsymbol{H}$	$\mathbb{E}(\boldsymbol{W}_t\boldsymbol{W}_t') + \boldsymbol{\gamma}_1\boldsymbol{\gamma}_1'$	$\mathbb{E}(\boldsymbol{W}_t\Sigma_e^{-1}\boldsymbol{W}_t') + \boldsymbol{\gamma}_1\boldsymbol{\gamma}_1'$

Define

$$\boldsymbol{Q}_t := \begin{pmatrix} \boldsymbol{X}_t\boldsymbol{A}_2' \otimes \mathcal{P}_1 + [\Gamma_2\boldsymbol{A}_1'(\boldsymbol{A}_1\Gamma_2\boldsymbol{A}_1')^+\boldsymbol{A}_1\boldsymbol{X}_t\boldsymbol{A}_2'] \otimes (\boldsymbol{I} - \mathcal{P}_1) \\ \mathcal{P}_2 \otimes \boldsymbol{X}_t'\boldsymbol{A}_1' + (\boldsymbol{I} - \mathcal{P}_2) \otimes [\Gamma_1\boldsymbol{A}_2'(\boldsymbol{A}_2\Gamma_1\boldsymbol{A}_2')^+\boldsymbol{A}_2\boldsymbol{X}_t'\boldsymbol{A}_1'] \end{pmatrix},$$

where $\boldsymbol{M}^+$ denotes the Moore-Penrose inverse of $\boldsymbol{M}$. Note that $\boldsymbol{Q}_t$ will appear in both Theorem 1 and Theorem 2. Although it seems to have the same definition in both theorems, the two versions actually differ because the Γ_i and $\mathcal{P}_i$ involved have different definitions.

Theorem 1. *Assume that $\{\boldsymbol{E}_t\}$ are i.i.d. with mean zero and finite second moments, and absolutely continuous. Assume that $0 < \text{rank}\,\boldsymbol{A}_i = k_i \leqslant d_i$, $\rho(\boldsymbol{A}_1)\rho(\boldsymbol{A}_2) < 1$, where $\rho(\cdot)$ denotes the spectral radius of a matrix, and Σ_e is non-singular. Also assume that the nonzero eigenvalues of $\boldsymbol{A}_1\Gamma_2\boldsymbol{A}_1'$*

are distinct, and the same for $\boldsymbol{A}_2\Gamma_1\boldsymbol{A}_2'$. *Then*

$$\sqrt{T}\begin{pmatrix}\operatorname{vec}\left[\hat{\boldsymbol{A}}_1^{ls}-\boldsymbol{A}_1\right]\\ \operatorname{vec}\left[(\hat{\boldsymbol{A}}_2^{ls})'-\boldsymbol{A}_2'\right]\end{pmatrix}\Rightarrow N(\boldsymbol{0},\Xi^{ls}),$$

where

$$\Xi^{ls}:=\boldsymbol{H}^{-1}\mathbb{E}(\boldsymbol{Q}_t\Sigma_e\boldsymbol{Q}_t')\boldsymbol{H}^{-1}. \tag{11}$$

Theorem 2. *Assume that* $\{\boldsymbol{E}_t\}$ *are i.i.d. with mean zero and finite second moments, and absolutely continuous. Assume that* $0<\operatorname{rank}\boldsymbol{A}_i=k_i\leqslant d_i$, $\rho(\boldsymbol{A}_1)\rho(\boldsymbol{A}_2)<1$, *and* Σ_e *is of the form* (8), *and is non-singular. Also assume that the nonzero eigenvalues of* $\Sigma_1^{-1/2}\boldsymbol{A}_1\Gamma_2\boldsymbol{A}_1'\Sigma_1^{-1/2}$ *are distinct, and the same for* $\Sigma_2^{-1/2}\boldsymbol{A}_2\Gamma_1\boldsymbol{A}_2'\Sigma_2^{-1/2}$. *Then*

$$\sqrt{T}\begin{pmatrix}\operatorname{vec}\left[\hat{\boldsymbol{A}}_1^{cc}-\boldsymbol{A}_1\right]\\ \operatorname{vec}\left[(\hat{\boldsymbol{A}}_2^{cc})'-\boldsymbol{A}_2'\right]\end{pmatrix}\Rightarrow N(\boldsymbol{0},\Xi^{cc}),$$

where

$$\Xi^{cc}:=\boldsymbol{H}^{-1}\mathbb{E}(\boldsymbol{Q}_t\Sigma_e^{-1}\boldsymbol{Q}_t')\boldsymbol{H}^{-1}. \tag{12}$$

We now consider the asymptotics of the composition and loading matrices $\boldsymbol{A}_{il}$ and $\boldsymbol{A}_{ic}$ in (4). The following discussion works the same for either $\hat{\boldsymbol{A}}_i^{\text{ls}}$ or $\hat{\boldsymbol{A}}_i^{\text{cc}}$. Therefore, we will use the unified notations $\hat{\boldsymbol{A}}_i$ and Ξ, dropping the superscripts $^{\text{ls}}$ and $^{\text{cc}}$. Since $\boldsymbol{A}_{ic}$ and $\boldsymbol{A}_{il}$ cannot be identified as seen from $\boldsymbol{A}_{il}\boldsymbol{A}_{ic}'=\boldsymbol{A}_{il}\boldsymbol{M}\boldsymbol{M}^{-1}\boldsymbol{A}_{ic}'$ for any invertible $k_i\times k_i$ matrix $\boldsymbol{M}$, we consider instead the singular value decomposition (SVD) of $\boldsymbol{A}_i$. Write $\boldsymbol{A}_i=\boldsymbol{U}_i\boldsymbol{D}_i\boldsymbol{V}_i'$, where both $\boldsymbol{U}_i$ and $\boldsymbol{V}_i$ are $d_i\times k_i$ ortho-normal matrices. Denote the j-th diagonal element of $\boldsymbol{D}_i$ by d_{ij}, and define $\boldsymbol{d}_i=(d_{i1},\ldots,d_{i,k_i})'$. Comparing (4), we see that $\boldsymbol{V}_i$ corresponds to the composition matrix $\boldsymbol{A}_{ic}$, $\boldsymbol{U}_i$ corresponds to the loading matrix $\boldsymbol{A}_{il}$, and $\boldsymbol{D}_i$ can be absorbed into either $\boldsymbol{A}_{il}$ or $\boldsymbol{A}_{ic}$. Let $\hat{\boldsymbol{A}}_i=\hat{\boldsymbol{U}}_i\hat{\boldsymbol{D}}_i(\hat{\boldsymbol{V}}_i)'$ be the SVD of $\hat{\boldsymbol{A}}_i$. Since $\hat{\boldsymbol{A}}_i\hat{\boldsymbol{A}}_i'=\hat{\boldsymbol{U}}_i\hat{\boldsymbol{D}}_i^2(\hat{\boldsymbol{U}}_i)'$, the asymptotic distribution of $\hat{\boldsymbol{U}}_i$ can be obtained based on that of $\hat{\boldsymbol{A}}_i\hat{\boldsymbol{A}}_i'$. Similarly, the asymptotic distribution of $\hat{\boldsymbol{V}}_i$ can be derived from that of $\hat{\boldsymbol{A}}_i'\hat{\boldsymbol{A}}_i$. Note that the asymptotic covariance matrix of $\operatorname{vec}(\hat{\boldsymbol{A}}_i)$ (for $i=1,2$) is a submatrix of Ξ and can be extracted from (11) or (12). Following that, we let Ξ_{i1} be the asymptotic covariance matrix of $\operatorname{vec}(\hat{\boldsymbol{A}}_i\hat{\boldsymbol{A}}_i')$, which can be obtained through the expansion

$$\hat{\boldsymbol{A}}_i\hat{\boldsymbol{A}}_i'=\boldsymbol{A}_i\boldsymbol{A}_i'+(\hat{\boldsymbol{A}}_i-\boldsymbol{A}_i)\boldsymbol{A}_i'+\boldsymbol{A}_i(\hat{\boldsymbol{A}}_i-\boldsymbol{A}_i)'+o_P(T^{-1/2}).$$

More specifically, when $i = 1$,

$$\Xi_{11} = [\boldsymbol{A}_1 \otimes \boldsymbol{I}_{d_1} + (\boldsymbol{I}_{d_1} \otimes \boldsymbol{A}_1)\boldsymbol{J}_{d_1,d_1}] \left\{\Xi[1 : d_1^2, 1 : d_1^2]\right\} [\boldsymbol{A}_1 \otimes \boldsymbol{I}_{d_1} + (\boldsymbol{I}_{d_1} \otimes \boldsymbol{A}_1)\boldsymbol{J}_{d_1,d_1}]',$$

where $\Xi[1 : d_1^2, 1 : d_1^2]$ is the upper left $d_1^2 \times d_1^2$ block of Ξ^{ls} or Ξ^{cc}, and the matrix $\boldsymbol{J}_{d_1,d_1}$ is defined in (1). The asymptotic covariance matrix of $\hat{\boldsymbol{A}}_1'\hat{\boldsymbol{A}}_1$, denoted by Ξ_{12}, has a similar expression. When $i = 2$, the matrices Ξ_{21} and Ξ_{22}, related to $\hat{\boldsymbol{A}}_2$, are also defined similarly.

Define the matrix $\boldsymbol{R}_{i1}$ as

$$\boldsymbol{R}_{i1} = (\boldsymbol{I}_{k_i} \otimes \boldsymbol{U}_i, \boldsymbol{I}_{k_i} \otimes \boldsymbol{U}_i^{\perp}) \begin{pmatrix} (\boldsymbol{D}_i^2 \otimes \boldsymbol{I}_{k_i} - \boldsymbol{I}_{k_i} \otimes \boldsymbol{D}_i^2 + \boldsymbol{L}_{k_i}\boldsymbol{L}_{k_i}')^{-1}(\boldsymbol{I}_{k_i^2} - \boldsymbol{L}_{k_i}\boldsymbol{L}_{k_i}')(\boldsymbol{U}_i' \otimes \boldsymbol{U}_i') \\ (\boldsymbol{D}_i^{-2}\boldsymbol{U}_i') \otimes (\boldsymbol{U}_i^{\perp})' \end{pmatrix},$$

and define $\boldsymbol{R}_{i2}$ similarly, but replacing $\boldsymbol{U}_i$ with $\boldsymbol{V}_i$.

As a consequence of Theorem 1 and Theorem 2, we have the following result regarding $\hat{\boldsymbol{U}}_i$ and $\hat{\boldsymbol{V}}_i$.

Corollary 3. *Assume the conditions of Theorem 1 or Theorem 2 hold, and that the singular values of $\boldsymbol{A}_1$ are distinct, and so are those of $\boldsymbol{A}_2$. For each of $i = 1, 2$, it holds that*

$$\sqrt{T}\text{vec}(\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i) \Rightarrow N\left(\boldsymbol{0}, \boldsymbol{R}_{i1}\Xi_{i1}\boldsymbol{R}_{i1}'\right),$$

and

$$\sqrt{T}\text{vec}(\hat{\boldsymbol{V}}_i - \boldsymbol{V}_i) \Rightarrow N\left(\boldsymbol{0}, \boldsymbol{R}_{i2}\Xi_{i2}\boldsymbol{R}_{i2}'\right),$$

and

$$\sqrt{T}(\hat{\boldsymbol{d}}_i - \boldsymbol{d}_i) \Rightarrow N\left(\boldsymbol{0}, \frac{1}{4}\boldsymbol{D}_i^{-1}\boldsymbol{L}_{k_i}'(\boldsymbol{U}_i' \otimes \boldsymbol{U}_i')\Xi_{i1}(\boldsymbol{U}_i \otimes \boldsymbol{U}_i)\boldsymbol{L}_{k_i}\boldsymbol{D}_i^{-1}\right).$$

Remark 1. Let $\boldsymbol{G}_i$ be the $k_i \times k_i$ matrix defined as

$$\boldsymbol{G}_i[j,k] = \begin{cases} (d_{ik}^2 - d_{ij}^2)^{-1} & \text{when } j \neq k, \\ 0 & \text{when } j = k. \end{cases}$$

Furthermore, let $\boldsymbol{U}_i^{\perp}$ be the $d_i \times (d_i - k_i)$ orthonormal matrix such that $(\boldsymbol{U}_i, \boldsymbol{U}_i^{\perp})$ is an orthogonal matrix. The asymptotic distribution of $\hat{\boldsymbol{U}}_i$ can also be obtained from the following equation

$$\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i = (\boldsymbol{U}_i, \boldsymbol{U}_i^{\perp}) \begin{pmatrix} \boldsymbol{G}_i \circ \left[\boldsymbol{U}_i'(\hat{\boldsymbol{A}}_i\hat{\boldsymbol{A}}_i' - \boldsymbol{A}_i\boldsymbol{A}_i')\boldsymbol{U}_i\right] \\ (\boldsymbol{U}_i^{\perp})'(\hat{\boldsymbol{A}}_i\hat{\boldsymbol{A}}_i' - \boldsymbol{A}_i\boldsymbol{A}_i')\boldsymbol{U}_i\boldsymbol{D}_i^{-2} \end{pmatrix} + o_P\left(\frac{1}{\sqrt{T}}\right),$$

where $\circ$ denotes the entry-wise Hadamard or Schur product of two matrices.

Remark 2. The joint distribution of $\hat{\boldsymbol{U}}_i$ and $\hat{\boldsymbol{V}}_i$ can also be derived. However, we choose not to spell the details out here for two reasons: the notations are already very complicated, and more importantly, there does not seem to be any direct applications of such a joint distribution.

5 Identification of the Rank

In most applications the ranks of $\boldsymbol{A}_i$ are unknown, and it is important to determine them from the data. This problem has been considered for multivariate reduced rank regression by Anderson (1951) and Anderson (2003), and for reduced rank autoregressive model by Kohn (1979), Reinsel and Velu (1998), Tiao and Tsay (1989) and Tsay and Tiao (1985), among others. For high dimensional reduced rank regression based on independent samples, penalized least squares can select the ranks along with the estimation, where the penalty is based on nuclear norm (Negahban and Wainwright, 2011; Yuan et al., 2007), ℓ_0 norm (Bunea et al., 2011), or Schatten-q quasi-norm (Rohde and Tsybakov, 2011). Basu et al. (2019) and Lin and Michailidis (2020) considered low rank VAR and tensor models, combining least squares estimation with the nuclear norm penalty.

We propose to use an information criterion to select the ranks. For a given pair of ranks (r_1, r_2), it is defined as

$$\begin{aligned} \mathrm{EBIC}(r_1, r_2) = &\log\left[\frac{1}{Td_1d_2}\sum_{t=2}^{T}\|\boldsymbol{X}_t - \hat{\boldsymbol{A}}_1^{\mathrm{ls}}\boldsymbol{X}_{t-1}(\hat{\boldsymbol{A}}_2^{\mathrm{ls}})'\|_F^2\right] \\ &+ \frac{1}{Td_1d_2}\cdot[\log(Td_2)\cdot r_1(2d_1 - r_1) + \log(Td_1)\cdot r_2(2d_2 - r_2)]. \end{aligned} \tag{13}$$

This can be viewed as an extended version of the Bayesian Information Criterion (Schwarz et al., 1978), so we use the acronym EBIC. Here the likelihood is calculated for the model where the entries of $\boldsymbol{E}_t$ are iid $N(0, \sigma^2)$, so it is best viewed as a "quasi"-likelihood. It is not precisely derived according to the posterior probability under the Bayesian framework (Haughton et al., 1988). Instead, we combine the quasi-log-likelihood and a penalty term where the numbers of parameters are multiplied by the logarithm of the sample sizes. Instead of simply counting the number of parameters, the effective number of parameters (Mukherjee et al., 2015; Yuan, 2016) can also be used in the EBIC. However, we choose the current version for simplicity. The selected pair of ranks $(\hat{k}_1, \hat{k}_2)$ minimizes the EBIC over all the pairs (r_1, r_2) such that $1 \leqslant r_1 \leqslant r_{1\max}$ and $1 \leqslant r_2 \leqslant r_{2\max}$, where $r_{1\max}$ and $r_{2\max}$ are pre-determined maximum ranks of $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$, respectively. If no information is available and the dimensions d_1 and d_2 are not too large, we can

simply use $r_{1\max} = d_1$ and $r_{2\max} = d_2$. The following Theorem 4 confirms that the EBIC in (13) does achieve the consistency. Its empirical performances are also outstanding, as will be shown in Section 6.1.

Since there are two ranks to be determined, a direct search via (13) over all possible pairs of ranks can be very costly when both $r_{1\max}$ and $r_{2\max}$ are large. We also consider selecting these two ranks separately. Specifically, the selected ranks are

$$\hat{r}_1 = \arg\min_{r_1} \mathrm{EBIC}(r_1, r_{2\max}) \qquad \hat{r}_2 = \arg\min_{r_2} \mathrm{EBIC}(r_{1\max}, r_2).$$

Theorem 4. *Assume that $\{\boldsymbol{E}_t\}$ are i.i.d. with mean zero and finite second moments. Also assume that $0 < \operatorname{rank} \boldsymbol{A}_i = k_i \leqslant d_i$, $\rho(\boldsymbol{A}_1)\rho(\boldsymbol{A}_2) < 1$, and Σ_e is non-singular. Then both the joint* EBIC *and the separate* EBIC_i *select the true ranks consistently, given that $k_i \leqslant r_{i\max}$.*

6 Numerical Studies

6.1 Simulations

In this section, we investigate the finite sample performance of the proposed estimation and rank determination procedures for the RRMAR models under various simulation setups. The simulation study consists of three parts. The first part is designed to compare the empirical behavior of the proposed alternating least square estimator $\hat{\boldsymbol{A}}_i^{\mathrm{ls}}$, labelled as RR.LS in the figures, and the alternating MLE $\hat{\boldsymbol{A}}_i^{\mathrm{cc}}$, labelled as RR.CC. The true ranks are taken as known. The least squares estimator without rank constraints (labelled as LSE) in Chen et al. (2019a) is also included as a benchmark for comparison. In the second part, we report the coverage probabilities of the confidence intervals constructed based on Theorem 1, Theorem 2 and Corollary 3. The third part examines the rank determination based on the EBIC proposed in Section 5. We also experiment with rank selection by rolling forecasting.

For given dimensions d_i and ranks k_i, the observed data $\boldsymbol{X}_t$ are simulated according to model (3). The matrix $\boldsymbol{A}_1$ is generated according to $\boldsymbol{A}_1 = \boldsymbol{Q}_1 \Lambda \boldsymbol{Q}_2'$, where the entries of the $k_1 \times k_2$ diagonal matrix Λ are sampled from the uniform distribution over the interval $[0.5, 1.5]$, and the $d_1 \times k_1$ orthonormal matrices $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$ are generated randomly from the Haar distribution. The matrix $\boldsymbol{A}_2$ is generated in the same way. The two matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are then rescaled so that

$\rho := \rho(\boldsymbol{A}_1)\rho(\boldsymbol{A}_2) < 1$ and $||\boldsymbol{A}_1||_F = 1$. Throughout all the simulation studies, two different settings of the covariance structure of the innovation matrix $\boldsymbol{E}_t$ are considered:

(I) The covariance matrix $\Sigma_e = \mathrm{Cov}(\mathrm{vec}(\boldsymbol{E}_t))$ is randomly generated according to $\Sigma_e = \boldsymbol{Q}\Lambda\boldsymbol{Q}'$, where the entries of the diagonal matrix Λ are equally spaced over $[1, 10]$, and $\boldsymbol{Q}$ is a random orthogonal matrix generated from the Haar distribution.

(II) The covariance matrix Σ_e takes the form (8), where each of Σ_1 and Σ_2 is generated in the same way as the Σ_e in Setting I, except that the diagonal entries of Λ are equally spaced over $[1, 5]$.

For a particular simulation setting with multiple repetitions, the matrices $\boldsymbol{A}_i$ and Σ_e are fixed.

In the first experiment, for each configuration of sample size T, dimensions d_i and ranks k_i, we repeat the simulation 100 times, and show a box plot of the estimation error

$$\log(\|\hat{\boldsymbol{A}}_2 \otimes \hat{\boldsymbol{A}}_1 - \boldsymbol{A}_2 \otimes \boldsymbol{A}_1\|_F^2).$$

The spectral radius is fixed at $\rho = .75$. Figure 1 and Figure 2 use these boxplots to compare LSE, RR.LS and RR.CC, under Settings (I) and (II) respectively. It is seen from both figures that the advantage of RR.LS and RR.CC over LSE gets bigger as the dimensions grow higher. On the other hand, for fixed dimensions, this advantage becomes smaller as the ranks increase. From Figure 1 we also find that, under Setting (I) of the error covariance matrix, RR.CC performs similarly as RR.LS does, even though the covariance matrix Σ_e does not have the form (8), which is assumed for RR.CC. On the other hand, Figure 2 clearly demonstrates the advantage of RR.CC over RR.LS under setting (II), when Σ_e does bear the form (8).

In the second part, we consider the coverage probabilities of the confidence intervals based on Theorem 1, Theorem 2 and Corollary 3. In this experiment the true ranks are fixed at $k_1 = 3$ and $k_2 = 2$. We run simulations 1000 times for sample size $T = 200, 400, 1000$, and consider the cases of dimension $(d_1, d_2) = (6, 4), (9, 6), (15, 10)$ and $\rho = 0.25, 0.5, 0.75$. For RR.LS, the error covariance matrix settings (I) and (II) are considered. For RR.CS, we consider two 'correct' settings (II') and (II), where in (II') we use $\Sigma_e = \boldsymbol{I}_{d_2} \otimes \boldsymbol{I}_{d_1}$. The confidence intervals of the entries of the matrices $\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{U}_1, \boldsymbol{V}_1, \boldsymbol{U}_2, \boldsymbol{V}_2$ are constructed. Table 1 shows the percentage that the true parameters fall within their corresponding marginal 95% confidence intervals. Each percentage records the average empirical coverage over all involved matrix entries. It can be seen from the table that the

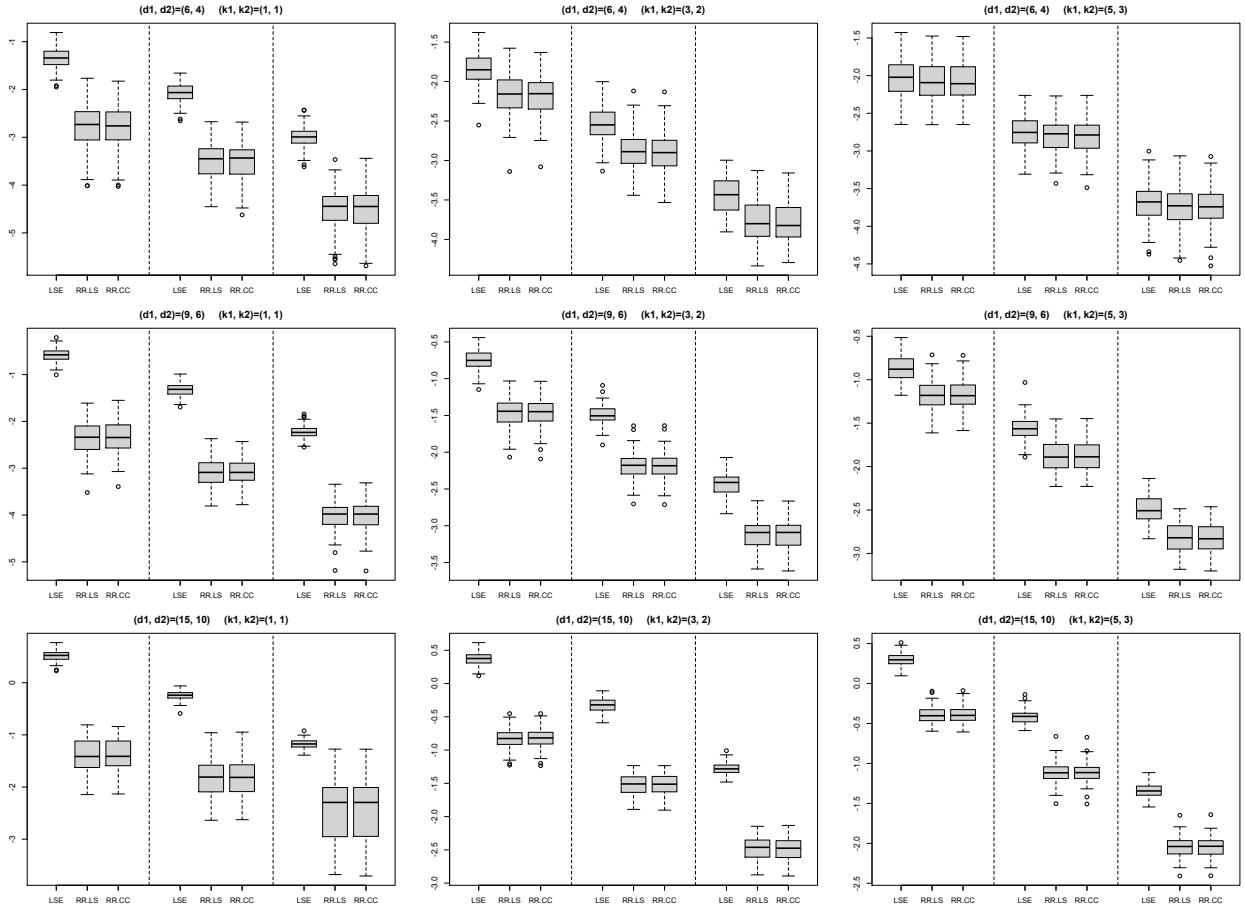

Figure 1: Comparison of LSE, RR.LS and RR.CC. The three panels in each figure correspond to sample sizes 200, 400 and 1000 respectively. The errors are generated according to Setting I.

coverage is quite accurate, especially when the sample size is large ($T = 1000$). The empirical coverage probabilities are closer to the nominal ones under Setting (I) for RR.LS and Setting (II') for RR.CS than those under Setting (II). This is probably due to the fact that the singulars values of Σ_e under Setting (II) are more spread out, making Σ_e more different from the scalar matrix.

The third part of the simulation considers the performance of the rank determination procedure using the joint EBIC (13) and the separate EBIC. Simulations are conducted under various configurations of the sample size, dimensions, ranks and signal strength ρ, and the empirical probabilities of selecting the correct ranks out of 100 repetitions are recorded. When the true ranks are $k_1 = k_2 = 1$, or when the signal strength ρ is not too small ($\rho \geqslant 0.1$), both selection procedures are able to determine the ranks perfectly. Therefore, we choose to report in Table 2 only the results

			RR.LS						RR.CC					
		Setting	I			II			II′			II		
		T	200	400	1000	200	400	1000	200	400	1000	200	400	1000
	ρ	(d_1, d_2)												
$(\boldsymbol{A}_1, \boldsymbol{A}_2)$	0.75	$(6, 4)$	93.5	93.8	94.3	92.1	92.3	92.5	93.9	94.2	95.0	91.0	91.3	91.3
		$(9, 6)$	94.3	94.6	95.3	92.2	92.6	92.8	94.0	94.4	95.0	91.8	92.1	92.4
		$(15, 10)$	94.5	94.7	95.5	92.5	92.8	93.3	93.8	94.1	94.9	92.2	92.5	93.0
	0.5	$(6, 4)$	93.5	93.8	94.5	91.4	91.9	92.1	93.7	94.0	94.9	90.5	91.0	91.3
		$(9, 6)$	94.3	94.6	95.4	91.7	92.4	92.6	93.8	94.3	95.0	91.2	92.0	92.2
		$(15, 10)$	94.4	94.7	95.4	92.3	92.8	93.2	93.7	94.1	94.9	92.0	92.5	92.9
	0.25	$(6, 4)$	92.8	93.7	94.6	88.0	90.1	91.3	92.6	93.6	94.7	87.2	89.5	90.8
		$(9, 6)$	93.3	94.4	95.3	88.8	90.9	91.9	92.8	93.8	94.8	88.4	90.7	91.8
		$(15, 10)$	93.8	94.5	95.3	90.6	92.0	92.9	93.1	93.8	94.8	90.4	91.8	92.6
$(\boldsymbol{U}_1, \boldsymbol{V}_1)$	0.75	$(6, 4)$	94.4	94.2	94.2	92.4	92.7	92.5	94.4	94.3	94.5	91.3	91.5	91.1
		$(9, 6)$	95.0	95.0	95.1	92.4	92.8	92.6	94.5	94.4	94.8	92.0	92.2	92.3
		$(15, 10)$	94.9	95.2	95.3	92.7	92.8	93.3	94.2	94.5	94.8	92.8	93.0	93.3
	0.5	$(6, 4)$	94.6	94.2	94.1	92.9	92.5	92.2	94.5	94.3	94.2	91.8	91.5	91.1
		$(9, 6)$	95.1	95.0	95.1	92.7	93.0	92.6	94.8	94.5	94.9	92.2	92.4	92.3
		$(15, 10)$	94.9	95.0	95.4	92.8	92.7	93.0	94.1	94.4	94.8	92.7	92.6	92.8
	0.25	$(6, 4)$	95.2	94.8	94.7	93.0	92.6	92.4	95.0	94.6	94.4	92.3	91.7	91.6
		$(9, 6)$	95.3	95.3	95.3	92.5	92.8	92.6	94.9	94.8	94.9	92.0	92.4	92.2
		$(15, 10)$	95.0	95.0	95.1	92.4	92.4	92.6	94.3	94.2	94.5	92.3	92.3	92.6
$(\boldsymbol{U}_2, \boldsymbol{V}_2)$	0.75	$(6, 4)$	94.1	94.3	94.2	91.6	91.7	91.1	94.5	94.5	94.3	91.2	91.3	90.6
		$(9, 6)$	94.7	94.6	95.5	92.0	91.9	92.5	94.4	94.2	94.8	91.9	92.1	92.9
		$(15, 10)$	95.1	95.6	95.3	92.6	93.2	93.1	94.3	95.0	94.7	93.2	93.7	93.6
	0.5	$(6, 4)$	94.3	94.4	94.6	90.9	91.4	91.0	94.1	94.7	94.6	90.8	91.1	90.7
		$(9, 6)$	94.8	94.7	95.3	91.8	91.8	92.3	94.4	94.2	94.5	91.5	91.9	92.4
		$(15, 10)$	95.0	95.5	95.2	92.4	93.1	92.8	94.2	94.9	94.6	92.7	93.4	93.1
	0.25	$(6, 4)$	93.7	93.7	94.8	88.8	89.8	90.6	93.3	93.9	94.7	88.3	89.9	91.0
		$(9, 6)$	93.8	94.7	95.3	89.4	91.0	91.9	93.5	94.1	94.4	89.2	91.2	92.1
		$(15, 10)$	94.6	95.6	95.1	91.0	92.5	92.4	93.7	94.8	94.5	91.5	92.9	92.7

Table 1: Empirical coverage probabilities (in percentage) of the 95% confidence intervals.

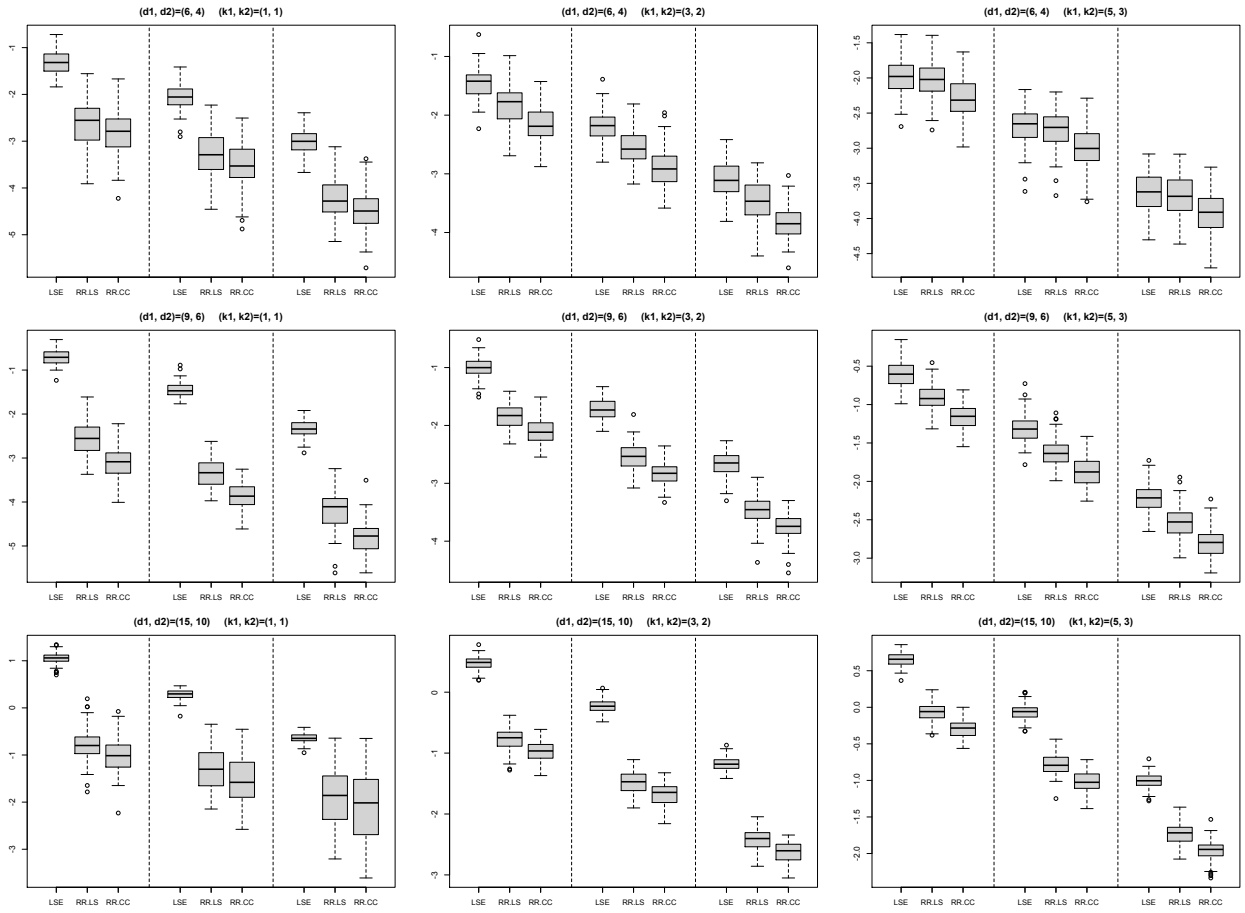

Figure 2: Comparison of LSE, RR.LS and RR.CC. The three panels in each figure correspond to sample sizes 200, 400 and 1000 respectively. The errors are generated according to Setting II.

for the configurations that are more challenging, with $\rho = .15$ when the true ranks are $(3, 2)$, and $\rho = .25$ when the true ranks are $(5, 3)$. A closer look of the simulation results (not shown in the table) reveals that, in these very low signal to noise ratio cases, both EBIC procedures tend to select ranks smaller than the true ranks. However, larger sampling sizes significantly enhance the performance. Moreover, when the autocorrelation strength ρ is larger than those reported in Table 2, both procedures make nearly perfect choices of the ranks for all configurations and covariance settings. We also note that the performances of the joint and separate procedures are almost the same.

It is also observed that the performance under the error covariance setting (II) is worse than that under Setting (I). This is due to the design of Σ_e in these two settings. The eigenvalues of Σ_e

spread over $[1, 10]$ in Setting (I), and over $[1, 25]$ in Setting (II). Therefore, both the estimation and model selection are more challenging under Setting (II).

We also experiment with using rolling forecasting to choose the ranks. We consider the range $1 \leqslant r_i \leqslant \min\{d_i, k_i + 2\}$, $i = 1, 2$, as the candidate set of $\text{rank}(\boldsymbol{A}_i)$. For each configuration of $(\rho, T, k_1, k_2, d_1, d_2)$, we choose $T/4$ as the rolling forecast origin, calculate the entry-wise squared forecast error (SFE) of the one-step ahead prediction of $\boldsymbol{X}_{s+1}$, $T/4 \leqslant s \leqslant T$, then take the average over the $d_1 d_2$ series and over the time,

$$\text{MSFE}(r_1, r_2) = \frac{1}{d_1 d_2 (3T/4)} \sum_{s=T/4}^{T-1} \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \left| \hat{\boldsymbol{X}}_{s+1}^{(r_1, r_2)}[i, j] - \boldsymbol{X}_{s+1}[i, j] \right|^2.$$

The estimated ranks $(\widehat{k}_1, \widehat{k}_2)$ is the pair (r_1, r_2) with the smallest $\text{MSFE}(r_1, r_2)$. Table 3 shows the proportion of the correct selection out of 100 repetitions. It is seen that, although rolling forecast criterion still performs very well in most cases, it has a much higher variability than EBIC. It performs better than EBIC in the case $\rho = .25$ and $(k_1, k_2) = (5, 3)$, when the sample size is small.

		$(r_1, r_2) = (3, 2)$, $\rho = .15$			$(r_1, r_2) = (5, 3)$, $\rho = .25$		
	(d_1, d_2)	200	400	1000	200	400	1000
I	$(6, 4)$	(.01, .01)	(.67, .68)	(1, 1)	(.36, .36)	(.96, .96)	(1, 1)
	$(9, 6)$	(.15, .15)	(.90, .92)	(1, 1)	(.07, .07)	(.71, .71)	(1, 1)
	$(15, 10)$	(1, 1)	(1, 1)	(1, 1)	(.00, .00)	(.47, .48)	(1, 1)
II	$(6, 4)$	(.02, .04)	(.38, .39)	(.99, .99)	(.25, .23)	(.92, .92)	(1, 1)
	$(9, 6)$	(.00, .00)	(.45, .47)	(1, 1)	(.10, .09)	(.70, .69)	(1, 1)
	$(15, 10)$	(.98, .99)	(1, 1)	(1, 1)	(.00, .00)	(.03, .04)	(1, 1)

Table 2: Empirical probabilities of the correct rank selection by the EBIC. I, II stand for different covariance structures of $\boldsymbol{E}_t$. For each cell, two numbers correspond to the joint and separate selections respectively. The second row shows the sample sizes.

			(1,1)			(3,2)			(5,3)		
		(d_1, d_2)	200	400	1000	200	400	1000	200	400	1000
$\rho = .5$	I	(6,4)	0.94	0.98	0.98	0.86	0.91	0.92	0.68	0.74	0.78
		(9,6)	0.99	1	1	0.97	0.99	0.99	0.94	0.98	1
		(15,10)	1	1	1	0.99	1	1	1	1	1
	II	(6,4)	0.95	0.97	0.99	0.87	0.88	0.92	0.74	0.77	0.79
		(9,6)	0.99	0.99	1	0.96	0.98	0.98	0.94	0.95	0.96
		(15,10)	1	1	1	0.99	1	1	1	1	1
$\rho = .25$	I	(6,4)	0.95	0.98	0.98	0.81	0.88	0.90	0.62	0.68	0.74
		(9,6)	1	1	1	0.96	0.98	0.99	0.94	0.98	0.98
		(15,10)	1	1	1	1	1	1	0.96	1	1
	II	(6,4)	0.96	0.97	0.98	0.44	0.79	0.93	0.64	0.76	0.82
		(9,6)	0.99	1	1	0.69	0.98	0.98	0.95	0.95	0.95
		(15,10)	1	1	1	0.99	1	1	0.94	1	1

Table 3: Empirical probabilities of the correct rank selection by rolling forecast. I, II stands for different covariance structures of $\boldsymbol{E}_t$

6.2 Example

We use the RRMAR model to study the 100 portfolios formed with publicly traded stocks sorted according to the deciles of ME (Market Equity) and BE/ME (Book-to-Market Ratio), leading to a 10×10 matrix (row and column for BE/ME and ME *resp.*) of the portfolio returns at each month. These portfolios have been constructed based on Fama and French (1993), and the data is from Prof. French's publicly available data library. Monthly data from 1964 to 2018 is used, with total 660 observations. The market return is subtracted from each entry of the tensor, i.e. we are modeling the market excess returns. We set the ranks as $k_1 = 2$ and $k_2 = 1$ according to the rolling forecast performance. Using RR.CC approach, the estimated right singular vectors $\hat{\boldsymbol{V}}_1$ of $\boldsymbol{A}_1$ and the right singular vector $\hat{\boldsymbol{V}}_2$ of $\boldsymbol{A}_2$, and their corresponding estimated standard errors are shown in Table 4. The $\hat{\boldsymbol{V}}_1$ shown is after the `varimax` transformation (Kaiser, 1958). In each panel of the table, the first row corresponds to a column of $\hat{\boldsymbol{V}}_i$, and the second row gives the standard errors.

Entries which are not significant at 5% level are shown in light gray color.

$\hat{\boldsymbol{V}}_1[,1]$	-0.17	0.36	-0.42	0.36	0.06	-0.01	0.04	0.04	0.28	-0.67
	0.05	0.05	0.06	0.06	0.07	0.07	0.07	0.06	0.05	0.02
$\hat{\boldsymbol{V}}_1[,2]$	0.17	-0.23	-0.20	-0.38	-0.03	0.11	-0.24	0.70	0.41	-0.05
	0.04	0.05	0.04	0.05	0.05	0.05	0.05	0.03	0.04	0.05
$\hat{\boldsymbol{V}}_2$	0.13	0.21	-0.11	0.13	-0.01	0.19	0.07	0.24	0.00	-0.90
	0.04	0.04	0.04	0.04	0.04	0.04	0.03	0.04	0.04	0.02

Table 4: Estimated right singular vectors of the coefficient matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$, with their corresponding estimated standard errors, for the Fama-French 10×10 portfolio return series.

It is very interesting to compare this result with the classical Fama-French three-factor model (Fama and French, 1993). First, from the RRMAR model, it is seen that $\boldsymbol{F}_t$, defined in the representation (4), contains two factors, consistent with the Fama-French three-factor model (with the third one being the market return, which is removed here). Second, the middle portion of the estimates $\hat{\boldsymbol{V}}_1$ and $\hat{\boldsymbol{V}}_2$ are in general smaller (as shown in the light gray color) compared with the two ends, which means the two factors are obtained by contrasting small and big ME, and high and low BE/ME, again agreeing with the Fama-French model. Therefore, the RRMAR model provides a generative and data-driven construction of the Fama-French model, relating the factors to the returns from the previous month. We note that Fama-French factors are constructed using the concurrent observations, while the factors obtained under RRMAR model are constructed using the observations at the previous time period, hence allowing for predictions. In Figure 3 we plot the estimated factors in $\boldsymbol{F}_t$ and the Fama-French factors SMB (Small minus Big ME) and HML (High minus Low BE/ME). For the precise definitions and constructions of Fama-French factors, we refer the readers to Fama and French (1993). It is seen the estimated factors $\boldsymbol{F}_t$ exhibit some co-movement with the Fama-French factors. On the other hand, the factors under the RRMAR model are obtained by a composite contrasting: ME and BE/ME are contrasted simultaneously in $\boldsymbol{F}_t := \boldsymbol{V}_1' \boldsymbol{X}_{t-1} \boldsymbol{V}_2$.

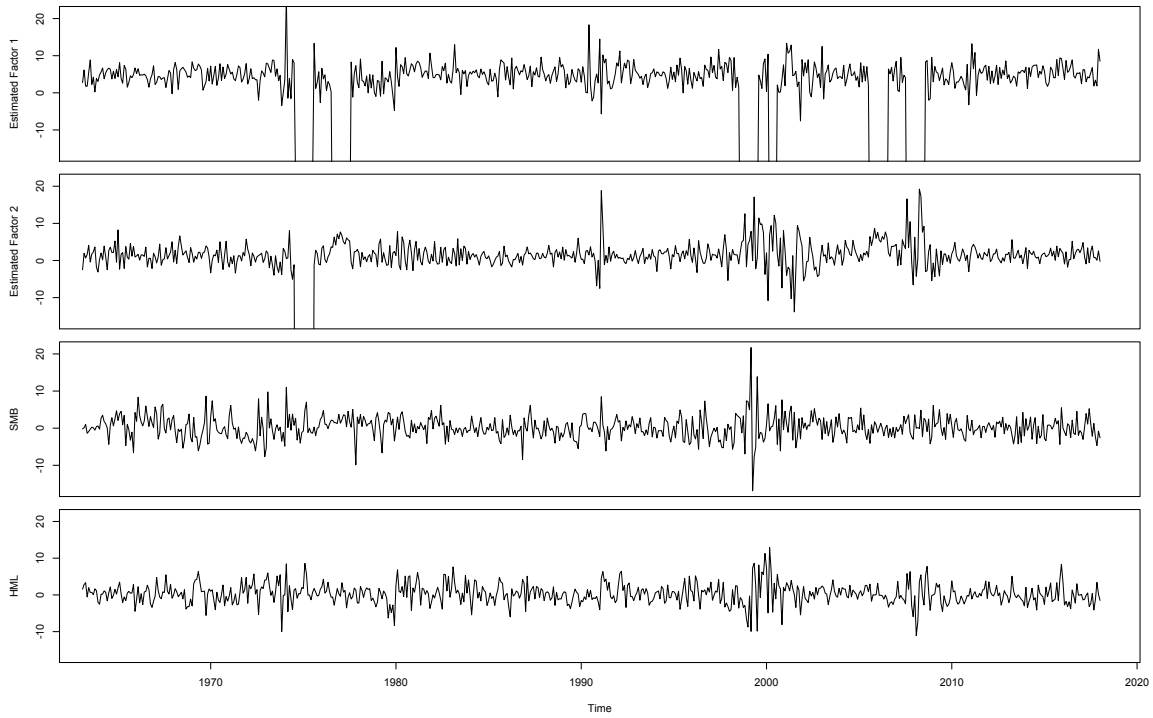

Figure 3: Estimated factors in $\boldsymbol{F}_t$ and the Fama-French factors SMB and HML.

	FF	iAR(1)	iAR(2)	VAR(1)	PROJ	LSE	MLE	RR.LS	RR.CC
MSE	17.51	13.67	13.54	17.62	17.55	13.71	13.56	13.39	13.19
# par	300	100	200	10,000	200	200	200	55	55

Table 5: Out-sample prediction performance comparison of various models for the Fama-French matrix series.

The one-step rolling forecast performance on the last 20 years are summarized in Table 5 in which we compare the following nine methods.

(i) **FF:** Using the Fama-French factors HMF and SMB at time $t-1$ to predict $\boldsymbol{X}_t$, using individual linear regressions.

(ii) **iAR(1) and iAR(2):** Fit an AR(1) or AR(2) model to each individual series.

(iii) **VAR(1):** Fit a VAR(1) model to $\text{vec}(\boldsymbol{X}_t)$.

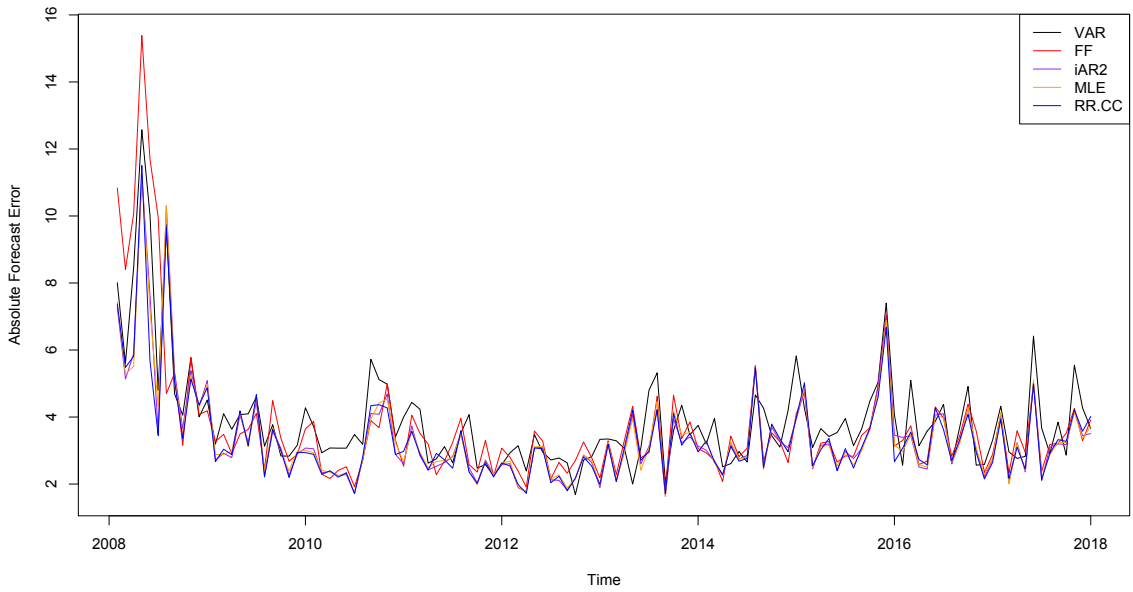

Figure 4: One-step rolling forecast errors of five selected methods.

(iv) **PROJ, LSE, MLE:** Fit the MAR(1) model (without rank constraint) to $\boldsymbol{X}_t$ using projection, least squares and MLE methods. See Chen et al. (2019a) for details.

(v) **RR.LS:** Reduced rank MAR(1) model, fitted by least squares.

(vi) **RR.CC:** Reduced rank MAR(1) model, fitted by MLE under the assumption (8).

From Table 5, it is seen that VAR(1) model involves a 100×100 coefficient matrix and significantly overfits the data, with the worst out-sample prediction performance. The MAR model (LSE and MLE), has similar performance to fitting each individual series separately (iAR(1) and iAR(2)). Comparing to the MAR model without rank constraint, the reduced rank models use less number of parameters and predict better. In particular, the RRMAR model estimated by MLE (RR.CC) has the smallest rolling forecast error. We also plot the forecast errors of five selected methods in Figure 4. From the figure, it is seen that the blue curve is in general below all other curves constantly, confirming the best performance of RR.CC shown in Table 5.

7 Conclusion

We introduce the reduced rank matrix autoregressive model, which relies on an autoregressive term involving bilinear coefficient matrices, and assumes rank deficiency of the coefficient matrices. An important feature of the model is that it is similar to the matrix factor model, but it is generative and the prediction can be done easily. Both LSE and MLE are studied, where the latter is considered under an additional assumption that the covariance tensor of the error matrix is separable. We propose to use extended BIC to select the ranks of the coefficient matrices. Our numerical analysis suggests that even if the separability assumption on the covariance tensor does not hold, MLE still has reasonable and almost equally good performance, comparing with LSE. On the other hand, MLE can perform much better when that assumption does stand. Therefore, we would recommend the use of MLE in practice.

There are a number of directions to extend the study of the reduced rank autoregressive model. For example the conditional mean can involve multiple terms of the form $\sum_{j=1}^{J} \boldsymbol{A}_{j1}\boldsymbol{X}_{t-1}\boldsymbol{A}_{j2}'$, and multiple lagged terms $\boldsymbol{X}_{t-1}, \ldots, \boldsymbol{X}_{t-p}$. The model can be extended for tesor time series as well. More importantly, the asymptotic analysis has been carried out for the fixed dimensional case in the current paper. It is interesting and important to study the model under the high dimensional paradigm. In particular, we would like to understand: (i) what are the convergence rates of $\hat{\boldsymbol{A}}_i$; and (ii) how to obtain initial estimates of $\boldsymbol{A}_i$ to start the alternating algorithm. To select the ranks, either the information criterion based procedure can be adapted to account for the high dimensionality, or the singular (eigen-)value based approach (Lam and Yao, 2012; Wang et al., 2019) can be employed. The relationship between the reduced rank tensor autoregressive model and the dynamic tensor factor model (Chen et al., 2019b) is also worth exploring.

References

Allen, G. I. and Tibshirani, R. (2010). Transposable regularized covariance models with an application to missing data imputation. *The Annals of Applied Statistics*, 4(2):764.

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832.

Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Ann. Math. Statistics*, 22:327–351.

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis.* Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, third edition.

Basu, S., Li, X., and Michailidis, G. (2019). Low rank and structured modeling of high-dimensional vector autoregressions. *IEEE Transactions on Signal Processing*, 67(5):1207–1222.

Basu, S., Michailidis, G., et al. (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567.

Bunea, F., She, Y., and Wegkamp, M. H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann. Statist.*, 39(2):1282–1309.

Chen, E. Y. and Chen, R. (2019). Modeling dynamic transport network with matrix factor models: with an application to international trade flow. *arXiv preprint arXiv:1901.00769*.

Chen, L. and Huang, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, 107(500):1533–1545.

Chen, R., Xiao, H., and Yang, D. (2019+a). Autoregressive models for matrix-valued time series. *Jounral of Econometrics*. To appear. arXiv:1812.08916.

Chen, R., Yang, D., and Zhang, C.-h. (2019b). Factor models for high-dimensional tensor time series. *arXiv preprint arXiv:1905.07530*.

Cichocki, A., Mandic, D., De Lathauwer, L., Zhou, G., Zhao, Q., Caiafa, C., and Phan, H. A. (2015). Tensor decompositions for signal processing applications: From two-way to multiway component analysis. *IEEE signal processing magazine*, 32(2):145–163.

Cichocki, A., Zdunek, R., Phan, A. H., and Amari, S.-i. (2009). *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation.* John Wiley & Sons.

Davis, R. A., Zang, P., and Zheng, T. (2016). Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics*, 25(4):1077–1096.

De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000a). A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4):1253–1278.

De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000b). On the best rank-1 and rank-(r 1, r 2,..., rn) approximation of higher-order tensors. *SIAM journal on Matrix Analysis and Applications*, 21(4):1324–1342.

De Silva, V. and Lim, L.-H. (2008). Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1084–1127.

Diebold, F. X., Li, C., and Yue, V. Z. (2008). Global yield curve dynamics and interactions: a dynamic nelson–siegel approach. *Journal of Econometrics*, 146(2):351–363.

Ding, S. and Dennis Cook, R. (2018). Matrix variate regressions and envelope models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(2):387–408.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3 – 56.

Gao, Z. and Tsay, R. S. (2020). A two-way transformed factor model for matrix-variate time series. *arXiv preprint arxiv:2011.09029*.

Ghosh, S., Khare, K., and Michailidis, G. (2019). High-dimensional posterior consistency in bayesian vector autoregressive models. *Journal of the American Statistical Association*, 114(526):735–748.

Ghosh, S., Khare, K., and Michailidis, G. (2020+). Strong selection consistency of bayesian vector autoregressive models based on a pseudo-likelihood approach. *The Annals of Statistics*.

Giannone, D., Reichlin, L., and Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4):665–676.

Hafner, C. M., Linton, O. B., and Tang, H. (2020). Estimation of a multiplicative correlation structure in the large dimensional case. *Journal of Econometrics*, 217(2):431–470.

Hall, E. C., Raskutti, G., and Willett, R. M. (2018). Learning high-dimensional generalized linear autoregressive models. *IEEE Transactions on Information Theory*, 65(4):2401–2422.

Han, F., Lu, H., and Liu, H. (2015). A direct estimation of high dimensional stationary vector autoregressions. *J. Mach. Learn. Res.*, 16:3115–3150.

Han, Y., Chen, L., and Wu, W. (2020+a). Sparse nonlinear vector autoregressive models. Technical report.

Han, Y., Chen, R., Yang, D., and Zhang, C.-h. (2020b). Tensor factor model estimation by iterative projection. *arXiv preprint arXiv:2006.02611*.

Han, Y., Zhang, C.-h., and Chen, R. (2020c). Rank determination in tensor factor model. *arXiv preprint arxiv:2011.07131*.

Hannan, E. J. (1970). *Multiple time series*. John Wiley and Sons, Inc., New York-London-Sydney.

Haughton, D. M. et al. (1988). On the choice of a model to fit data from an exponential family. *The Annals of Statistics*, 16(1):342–355.

Hoff, P. D. (2015). Multilinear tensor regression for longitudinal relational data. *The annals of applied statistics*, 9(3):1169.

Hoff, P. D. et al. (2011). Separable covariance arrays via the tucker product, with applications to multivariate relational data. *Bayesian Analysis*, 6(2):179–196.

Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.

Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *J. Multivariate Anal.*, 5:248–264.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200.

Kock, A. B. and Callot, L. (2015). Oracle inequalities for high dimensional vector autoregressions. *J. Econometrics*, 186(2):325–344.

Kohn, R. (1979). Asymptotic estimation and hypothesis testing results for vector linear time series models. *Econometrica: Journal of the Econometric Society*, pages 1005–1030.

Lam, C. and Yao, Q. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *Ann. Statist.*, 40(2):694–726.

Lin, J. and Michailidis, G. (2017). Regularized estimation and testing for high-dimensional multi-block vector-autoregressive models. *The Journal of Machine Learning Research*, 18(1):4188–4236.

Lin, J. and Michailidis, G. (2020). Regularized estimation of high-dimensional factor-augmented vector autoregressive (favar) models. *Journal of Machine Learning Research*, 21(117):1–51.

Linton, O. B. and Tang, H. (2019). Estimation of the kronecker covariance model by partial means and quadratic form. *arXiv preprint arXiv:1906.08908*.

Loh, P.-L. and Wainwright, M. J. (2011). High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *Advances in Neural Information Processing Systems*, pages 2726–2734.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer-Verlag, Berlin.

Melnyk, I. and Banerjee, A. (2016). Estimating structured vector autoregressive models. In *International Conference on Machine Learning*, pages 830–839.

Moench, E., Ng, S., and Potter, S. (2013). Dynamic hierarchical factor models. *Review of Economics and Statistics*, 95(5):1811–1817.

Mukherjee, A., Chen, K., Wang, N., and Zhu, J. (2015). On the degrees of freedom of reduced-rank estimators in multivariate regression. *Biometrika*, 102(2):457–477.

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

Nicholson, W. B., Matteson, D. S., and Bien, J. (2017). Varx-l: Structured regularization for large vector autoregressions with exogenous variables. *International Journal of Forecasting*, 3(33):627–651.

Raskutti, G. and Yuan, M. (2015). Convex Regularization for High-Dimensional Tensor Regression. *ArXiv e-prints*.

Reinsel, G. C. and Velu, R. P. (1998). *Multivariate reduced-rank regression*, volume 136 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

Rohde, A. and Tsybakov, A. B. (2011). Estimation of high-dimensional low-rank matrices. *Ann. Statist.*, 39(2):887–930.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Sidiropoulos, N. D., De Lathauwer, L., Fu, X., Huang, K., Papalexakis, E. E., and Faloutsos, C. (2017). Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 65(13):3551–3582.

Tiao, G. C. and Box, G. E. P. (1981). Modeling multiple time series with applications. *J. Amer. Statist. Assoc.*, 76(376):802–816.

Tiao, G. C. and Tsay, R. S. (1989). Model specification in multivariate time series. *Journal of the Royal Statistical Society: Series B (Methodological)*, 51(2):157–195.

Tsay, R. S. (2014). *Multivariate time series analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ.

Tsay, R. S. and Tiao, G. C. (1985). Use of canonical analysis in time series model identification. *Biometrika*, 72(2):299–315.

Tsiligkaridis, T. and Hero, A. O. (2013). Covariance estimation in high dimensions via kronecker product expansions. *IEEE Transactions on Signal Processing*, 61(21):5347–5360.

Velu, R. and Reinsel, G. C. (2013). *Multivariate reduced-rank regression: theory and applications*, volume 136. Springer Science & Business Media.

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

Yuan, M. (2016). Degrees of freedom in low rank matrix estimation. *Science China Mathematics*, 59(12):2485–2502.

Yuan, M., Ekici, A., Lu, Z., and Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 69(3):329–346.

Zhao, J. and Leng, C. (2014). Structured lasso for regression with matrix covariates. *Statist. Sinica*, 24:799–814.

Zhou, H., Li, L., and Zhu, H. (2013). Tensor Regression with Applications in Neuroimaging Data Analysis. *J. Amer. Statist. Assoc.*, 108(502):540–552.

Zhou, S. (2014). Gemini: Graph estimation with matrix variate normal instances. *The Annals of Statistics*, 42(2):532–562.

Appendix

The proof of Theorem 1 is similar to that of Theorem 2, and is much simpler since it does not involve the matrices Σ_i and $\hat{\Sigma}_i$. Therefore, we will only present the proof of Theorem 2 and skip the proof of Theorem 1.

Proof of Theorem 2. Let $\hat{\boldsymbol{A}}_i^{\text{cc}}$ and $\hat{\Sigma}_i$ be the MLE under the model (3) and (8). First, using the arguments of the proof of Theorem 4 in Chen et al. (2019a), we have that $\hat{\boldsymbol{A}}_i^{\text{cc}} = \boldsymbol{A}_i + O_P(T^{-1/2})$, and $\hat{\Sigma}_i = \Sigma_i + o_P(1)$. For the rest of the proof, we will drop the superscript $^{\text{cc}}$ to simplify the notation. The gradient condition for $\hat{\boldsymbol{A}}_1$ is given by:

$$\hat{\boldsymbol{A}}_1 \hat{\boldsymbol{S}}_{1xx} = \hat{\Sigma}_1^{1/2} \hat{\boldsymbol{U}}_1 \hat{\boldsymbol{U}}_1' \hat{\Sigma}_1^{-1/2} \hat{\boldsymbol{S}}_{1yx}, \tag{14}$$

where

$$\begin{aligned}
\hat{\boldsymbol{S}}_{1xx} &= \sum_t \boldsymbol{X}_{t-1} \hat{\boldsymbol{A}}_2' \hat{\Sigma}_2^{-1} \hat{\boldsymbol{A}}_2 \boldsymbol{X}_{t-1}', \\
\hat{\boldsymbol{S}}_{1yx} &= \sum_t \boldsymbol{X}_t \hat{\Sigma}_2^{-1} \hat{\boldsymbol{A}}_2 \boldsymbol{X}_{t-1}', \\
\hat{\Sigma}_1 &= \frac{1}{T-1} \sum_t \left(\boldsymbol{X}_t - \hat{\boldsymbol{A}}_1 \boldsymbol{X}_{t-1} \hat{\boldsymbol{A}}_2' \right) \hat{\Sigma}_2^{-1} \left(\boldsymbol{X}_t - \hat{\boldsymbol{A}}_1 \boldsymbol{X}_{t-1} \hat{\boldsymbol{A}}_2' \right)',
\end{aligned}$$

and $\hat{\boldsymbol{U}}_1$ is the $d_1 \times k_1$ matrix consisting of the first k_1 leading eigenvectors (all normalized to have unit length) of $\hat{\Sigma}_1^{-1/2} \hat{\boldsymbol{S}}_{1yx} \hat{\boldsymbol{S}}_{1xx}^{-1} \hat{\boldsymbol{S}}_{1xy} \hat{\Sigma}_1^{-1/2}$. With similarly defined quantities (by swapping $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$, Σ_1 and Σ_2, and $\boldsymbol{X}_t$ and $\boldsymbol{X}_t'$ respectively), we have

$$\hat{\boldsymbol{A}}_2 \hat{\boldsymbol{S}}_{2xx} = \hat{\Sigma}_2^{1/2} \hat{\boldsymbol{U}}_2 \hat{\boldsymbol{U}}_2' \hat{\Sigma}_2^{-1/2} \hat{\boldsymbol{S}}_{2yx}.$$

Note that $\hat{\boldsymbol{U}}_1$ can also be viewed as the first k_1 leading left singular vectors (all normalized to have unit length) of $\hat{\boldsymbol{M}}_1 := \hat{\Sigma}_1^{-1/2} \hat{\boldsymbol{S}}_{1yx} \hat{\boldsymbol{S}}_{1xx}^{-1/2}$. Let $\tilde{\boldsymbol{M}}_1 := \hat{\Sigma}_1^{-1/2} \boldsymbol{A}_1 \left(\sum_t \boldsymbol{X}_{t-1} \boldsymbol{A}_2' \hat{\Sigma}_2^{-1} \hat{\boldsymbol{A}}_2 \boldsymbol{X}_{t-1}' \right) \hat{S}_{1xx}^{-1/2}$, and $\tilde{\boldsymbol{U}}_1$ be the orthogonal matrix consisting of the normalized left singular vectors of $\tilde{\boldsymbol{M}}_1$. Let $\hat{\boldsymbol{M}}_1 = \hat{\boldsymbol{U}}_1 \hat{\boldsymbol{D}}_1 \hat{\boldsymbol{V}}_1'$ and $\tilde{\boldsymbol{M}}_1 = \tilde{\boldsymbol{U}}_1 \tilde{\boldsymbol{D}}_1 \tilde{\boldsymbol{V}}_1'$ be the SVD of $\hat{\boldsymbol{M}}_1$ and $\tilde{\boldsymbol{M}}_1$ respectively. The convergence rates of $\hat{\boldsymbol{A}}_i$ imply that $\hat{\boldsymbol{M}}_1 = \tilde{\boldsymbol{M}}_1 + O_P(1/\sqrt{T})$, and it follows that $\hat{\boldsymbol{U}}_1 = \tilde{\boldsymbol{U}}_1 + O_P(1/\sqrt{T})$, $\hat{\boldsymbol{V}}_1 = \tilde{\boldsymbol{V}}_1 + O_P(1/\sqrt{T})$ and $\hat{\boldsymbol{D}}_1 = \tilde{\boldsymbol{D}}_1 + O_P(1/\sqrt{T})$. Therefore, it holds that

$$\hat{\boldsymbol{M}}_1 - \tilde{\boldsymbol{M}}_1 = \tilde{\boldsymbol{U}}_1 \tilde{\boldsymbol{D}}_1 (\hat{\boldsymbol{V}}_1 - \tilde{\boldsymbol{V}}_1)' + \tilde{\boldsymbol{U}}_1 (\hat{\boldsymbol{D}}_1 - \tilde{\boldsymbol{D}}_1) \tilde{\boldsymbol{V}}_1' + (\hat{\boldsymbol{U}}_1 - \tilde{\boldsymbol{U}}_1) \tilde{\boldsymbol{D}}_1 \tilde{\boldsymbol{V}}_1' + O_P(1/\sqrt{T}),$$

and

$$(\boldsymbol{I}-\tilde{\boldsymbol{U}}_1\tilde{\boldsymbol{U}}_1')(\hat{\boldsymbol{U}}_1-\tilde{\boldsymbol{U}}_1)\tilde{\boldsymbol{U}}'=(\boldsymbol{I}-\tilde{\boldsymbol{U}}_1\tilde{\boldsymbol{U}}_1')(\hat{\boldsymbol{M}}_1-\tilde{\boldsymbol{M}}_1)\tilde{\boldsymbol{V}}_1\tilde{\boldsymbol{D}}_1^{-1}\tilde{\boldsymbol{U}}_1'+O_P(1/\sqrt{T}). \tag{15}$$

Note that $\tilde{\boldsymbol{U}}_1'\tilde{\boldsymbol{U}}_1=\hat{\boldsymbol{U}}_1'\hat{\boldsymbol{U}}_1=\boldsymbol{I}_{k_1}$, hence

$$\tilde{\boldsymbol{U}}_1'(\hat{\boldsymbol{U}}_1-\tilde{\boldsymbol{U}}_1)+(\hat{\boldsymbol{U}}_1-\tilde{\boldsymbol{U}}_1)'\tilde{\boldsymbol{U}}_1=O_P(1/\sqrt{T}).$$

Using the preceding equation, we have

$$\begin{aligned}\hat{\boldsymbol{U}}_1\hat{\boldsymbol{U}}_1'&=\tilde{\boldsymbol{U}}_1\tilde{\boldsymbol{U}}_1'+\tilde{\boldsymbol{U}}_1(\hat{\boldsymbol{U}}_1-\tilde{\boldsymbol{U}}_1)'+(\hat{\boldsymbol{U}}_1-\tilde{\boldsymbol{U}}_1)\tilde{\boldsymbol{U}}_1'+O_P(1/\sqrt{T})\\&=\tilde{\boldsymbol{U}}_1\tilde{\boldsymbol{U}}_1'+\tilde{\boldsymbol{U}}_1(\hat{\boldsymbol{U}}_1-\tilde{\boldsymbol{U}}_1)'(\tilde{\boldsymbol{U}}_1\tilde{\boldsymbol{U}}_1'+\boldsymbol{I}-\tilde{\boldsymbol{U}}_1\tilde{\boldsymbol{U}}_1')+(\hat{\boldsymbol{U}}_1-\tilde{\boldsymbol{U}}_1)\tilde{\boldsymbol{U}}_1'+O_P(1/\sqrt{T})\\&=\tilde{\boldsymbol{U}}_1\tilde{\boldsymbol{U}}_1'+\tilde{\boldsymbol{U}}_1(\hat{\boldsymbol{U}}_1-\tilde{\boldsymbol{U}}_1)'(\boldsymbol{I}-\tilde{\boldsymbol{U}}_1\tilde{\boldsymbol{U}}_1')+(\boldsymbol{I}-\tilde{\boldsymbol{U}}_1\tilde{\boldsymbol{U}}_1')(\hat{\boldsymbol{U}}_1-\tilde{\boldsymbol{U}}_1)\tilde{\boldsymbol{U}}_1'+O_P(1/\sqrt{T}).\end{aligned} \tag{16}$$

Combining (15) and (16) leads to

$$\begin{aligned}\hat{\boldsymbol{U}}_1\hat{\boldsymbol{U}}_1'=\tilde{\boldsymbol{U}}_1\tilde{\boldsymbol{U}}_1'&+(\boldsymbol{I}-\tilde{\boldsymbol{U}}_1\tilde{\boldsymbol{U}}_1')(\hat{\boldsymbol{M}}_1-\tilde{\boldsymbol{M}}_1)\tilde{\boldsymbol{V}}_1\tilde{\boldsymbol{D}}_1^{-1}\tilde{\boldsymbol{U}}_1'\\&+\tilde{\boldsymbol{U}}_1\tilde{\boldsymbol{D}}_1^{-1}\tilde{\boldsymbol{V}}_1'(\hat{\boldsymbol{M}}_1-\tilde{\boldsymbol{M}}_1)'(\boldsymbol{I}-\tilde{\boldsymbol{U}}_1\tilde{\boldsymbol{U}}_1')+O_P(1/\sqrt{T}).\end{aligned} \tag{17}$$

Let $\boldsymbol{M}_1:=\Sigma_1^{-1/2}\boldsymbol{A}_1\Gamma_2^{1/2}$ and $\boldsymbol{M}_1=\boldsymbol{U}_1\boldsymbol{D}_1\boldsymbol{V}_1'$ be its SVD. The consistencies of of $\hat{\boldsymbol{A}}_i$ and $\hat{\Sigma}_i$ also imply that $\tilde{\boldsymbol{M}}_1$, $\tilde{\boldsymbol{U}}_1$, $\tilde{\boldsymbol{V}}_1$ are consistent for $\boldsymbol{M}_1$, $\boldsymbol{U}_1$ and $\boldsymbol{V}_1$ respectively. This fact, combined with (17), yields

$$\begin{aligned}\hat{\boldsymbol{U}}_1\hat{\boldsymbol{U}}_1'=\tilde{\boldsymbol{U}}_1\tilde{\boldsymbol{U}}_1'&+(\boldsymbol{I}-\boldsymbol{U}_1\boldsymbol{U}_1')\Sigma_1^{-1/2}\left(\sum_t\boldsymbol{E}_t\Sigma_2^{-1}\boldsymbol{A}_2\boldsymbol{X}_{t-1}'\right)\Gamma_2^{-1/2}\left(\Sigma_1^{-1/2}\boldsymbol{A}_1\Gamma_2^{1/2}\right)^+\\&+\left(\Gamma_2^{1/2}\boldsymbol{A}_1'\Sigma_1^{-1/2}\right)^+\Gamma_2^{-1/2}\left(\sum_t\boldsymbol{E}_t\Sigma_2^{-1}\boldsymbol{A}_2\boldsymbol{X}_{t-1}'\right)'\Sigma_1^{-1/2}(\boldsymbol{I}-\boldsymbol{U}_1\boldsymbol{U}_1')+o_P(1/\sqrt{T}).\end{aligned}$$

Note that $\boldsymbol{U}_1\boldsymbol{U}_1'=\mathbb{P}_1$. Let $\boldsymbol{P}_i=\Sigma_i^{1/2}\mathbb{P}_i\Sigma_i^{-1/2}$. Plugging in the preceding equation into (14), and using the facts that $(\boldsymbol{I}-\boldsymbol{U}_1\boldsymbol{U}_1')\Sigma_1^{-1/2}\boldsymbol{A}_1\Gamma_2=\boldsymbol{0}$ and $\hat{\Sigma}_1^{1/2}\tilde{\boldsymbol{U}}_1\tilde{\boldsymbol{U}}_1'\hat{\Sigma}_1^{-1/2}\boldsymbol{A}_1\left(\sum_t\boldsymbol{X}_{t-1}\boldsymbol{A}_2'\hat{\Sigma}_2^{-1}\hat{\boldsymbol{A}}_2\boldsymbol{X}_{t-1}'\right)=\boldsymbol{A}_1\left(\sum_t\boldsymbol{X}_{t-1}\boldsymbol{A}_2'\hat{\Sigma}_2^{-1}\hat{\boldsymbol{A}}_2\boldsymbol{X}_{t-1}'\right)$, we get

$$\begin{aligned}&\hat{\boldsymbol{A}}_1\left(\sum_t\boldsymbol{X}_{t-1}\hat{\boldsymbol{A}}_2'\hat{\Sigma}_2^{-1}\hat{\boldsymbol{A}}_2\boldsymbol{X}_{t-1}'\right)-\boldsymbol{A}_1\left(\sum_t\boldsymbol{X}_{t-1}\boldsymbol{A}_2'\hat{\Sigma}_2^{-1}\hat{\boldsymbol{A}}_2\boldsymbol{X}_{t-1}'\right)\\=&(\boldsymbol{I}-\boldsymbol{P}_1)\left(\sum_t\boldsymbol{E}_t\Sigma_2^{-1}\boldsymbol{A}_2\boldsymbol{X}_{t-1}'\right)\Gamma_2^{-1/2}\left(\Sigma_1^{-1/2}\boldsymbol{A}_1\Gamma_2^{1/2}\right)^+\Sigma_1^{-1/2}\boldsymbol{A}_1\Gamma_2\\&+\boldsymbol{P}_1\left(\sum_t\boldsymbol{E}_t\Sigma_2^{-1}\boldsymbol{A}_2\boldsymbol{X}_{t-1}'\right)+o_P(\sqrt{T})\\=&(\boldsymbol{I}-\boldsymbol{P}_1)\left(\sum_t\boldsymbol{E}_t\Sigma_2^{-1}\boldsymbol{A}_2\boldsymbol{X}_{t-1}'\right)\boldsymbol{A}_1'(\boldsymbol{A}_1\Gamma_2\boldsymbol{A}_1')^+\boldsymbol{A}_1\Gamma_2+\boldsymbol{P}_1\left(\sum_t\boldsymbol{E}_t\Sigma_2^{-1}\boldsymbol{A}_2\boldsymbol{X}_{t-1}'\right)+o_P(\sqrt{T}),\end{aligned}$$

and

$$
\begin{aligned}
&(\hat{\boldsymbol{A}}_1 - \boldsymbol{A}_1)\left(\sum_t \boldsymbol{X}_{t-1}\boldsymbol{A}_2'\Sigma_2^{-1}\boldsymbol{A}_2\boldsymbol{X}_{t-1}'\right) + \boldsymbol{A}_1\left[\sum_t \boldsymbol{X}_{t-1}(\hat{\boldsymbol{A}}_2 - \boldsymbol{A}_2)'\Sigma_2^{-1}\boldsymbol{A}_2\boldsymbol{X}_{t-1}'\right] \\
=&(\boldsymbol{I} - \boldsymbol{P}_1)\left(\sum_t \boldsymbol{E}_t\Sigma_2^{-1}\boldsymbol{A}_2\boldsymbol{X}_{t-1}'\right)\boldsymbol{A}_1'(\boldsymbol{A}_1\Gamma_2\boldsymbol{A}_1')^+\boldsymbol{A}_1\Gamma_2 + \boldsymbol{P}_1\left(\sum_t \boldsymbol{E}_t\Sigma_2^{-1}\boldsymbol{A}_2\boldsymbol{X}_{t-1}'\right) + o_P(\sqrt{T}),
\end{aligned} \tag{18}
$$

A similar formula holds for $\hat{\boldsymbol{A}}_2$:

$$
\begin{aligned}
&\left(\sum_t \boldsymbol{X}_{t-1}'\boldsymbol{A}_1'\Sigma_1^{-1}\boldsymbol{A}_1\boldsymbol{X}_{t-1}\right)(\hat{\boldsymbol{A}}_2 - \boldsymbol{A}_2)' + \left[\sum_t \boldsymbol{X}_{t-1}'\boldsymbol{A}_1'\Sigma_1^{-1}(\hat{\boldsymbol{A}}_1 - \boldsymbol{A}_1)\boldsymbol{X}_{t-1}\right]\boldsymbol{A}_2' \\
=&\Gamma_1\boldsymbol{A}_2'(\boldsymbol{A}_2\Gamma_1\boldsymbol{A}_2')^+\boldsymbol{A}_2\left(\sum_t \boldsymbol{X}_{t-1}'\boldsymbol{A}_1'\Sigma_1^{-1}\boldsymbol{E}_t\right)(\boldsymbol{I} - \boldsymbol{P}_2) + \left(\sum_t \boldsymbol{X}_{t-1}'\boldsymbol{A}_1'\Sigma_1^{-1}\boldsymbol{E}_t\right)\boldsymbol{P}_2 + o_P(\sqrt{T}).
\end{aligned} \tag{19}
$$

Combining (18) and (19), it holds that after vectorization

$$
\begin{aligned}
&\sum_t \begin{pmatrix} (\boldsymbol{X}_{t-1}\boldsymbol{A}_2'\Sigma_2^{-1}\boldsymbol{A}_2\boldsymbol{X}_{t-1}') \otimes \boldsymbol{I} & (\boldsymbol{X}_{t-1}\boldsymbol{A}_2'\Sigma_2^{-1}) \otimes (\boldsymbol{A}_1\boldsymbol{X}_{t-1}) \\ (\boldsymbol{A}_2\boldsymbol{X}_{t-1}') \otimes (\boldsymbol{X}_{t-1}'\boldsymbol{A}_1'\Sigma_1^{-1}) & \boldsymbol{I} \otimes (\boldsymbol{X}_{t-1}'\boldsymbol{A}_1'\Sigma_1^{-1}\boldsymbol{A}_1\boldsymbol{X}_{t-1}) \end{pmatrix}\begin{pmatrix} \mathrm{vec}\left(\hat{\boldsymbol{A}}_1 - \boldsymbol{A}_1\right) \\ \mathrm{vec}\left(\hat{\boldsymbol{A}}_2' - \boldsymbol{A}_2'\right) \end{pmatrix} \\
=&\sum_t \begin{pmatrix} \boldsymbol{X}_{t-1}\boldsymbol{A}_2'\Sigma_2^{-1} \otimes \boldsymbol{P}_1 + [\Gamma_2\boldsymbol{A}_1'(\boldsymbol{A}_1\Gamma_2\boldsymbol{A}_1')^+\boldsymbol{A}_1\boldsymbol{X}_{t-1}\boldsymbol{A}_2']\Sigma_2^{-1} \otimes (\boldsymbol{I} - \boldsymbol{P}_1) \\ \boldsymbol{P}_2 \otimes \boldsymbol{X}_{t-1}'\boldsymbol{A}_1'\Sigma_1^{-1} + (\boldsymbol{I} - \boldsymbol{P}_2) \otimes [\Gamma_1\boldsymbol{A}_2'(\boldsymbol{A}_2\Gamma_1\boldsymbol{A}_2')^+\boldsymbol{A}_2\boldsymbol{X}_{t-1}'\boldsymbol{A}_1'\Sigma_1^{-1}] \end{pmatrix}\mathrm{vec}(\boldsymbol{E}_t) + o_P(\sqrt{T}).
\end{aligned} \tag{20}
$$

Note that $\Sigma_e = \Sigma_2 \otimes \Sigma_1$. Multiplying both sides of (20) by the matrix

$$
\begin{pmatrix} \boldsymbol{I} \otimes \Sigma_1^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \Sigma_2^{-1} \otimes \boldsymbol{I} \end{pmatrix},
$$

then (20) becomes

$$
\sum_t (\boldsymbol{W}_t\Sigma_e^{-1}\boldsymbol{W}_t')\begin{pmatrix} \mathrm{vec}\left(\hat{\boldsymbol{A}}_1 - \boldsymbol{A}_1\right) \\ \mathrm{vec}\left(\hat{\boldsymbol{A}}_2' - \boldsymbol{A}_2'\right) \end{pmatrix} = \sum_t \boldsymbol{Q}_{t-1}\Sigma_e^{-1}\mathrm{vec}(\boldsymbol{E}_t) + o_P(\sqrt{T}). \tag{21}
$$

Since $\|\boldsymbol{A}_1\|_F = \|\hat{\boldsymbol{A}}_1\|_F = 1$, it holds that $\boldsymbol{\alpha}'\mathrm{vec}(\hat{\boldsymbol{A}}_1 - \boldsymbol{A}_1) = O_P(1/T)$. By the law of large numbers, (21) implies that

$$
\boldsymbol{H}\begin{pmatrix} \mathrm{vec}\left(\hat{\boldsymbol{A}}_1 - \boldsymbol{A}_1\right) \\ \mathrm{vec}\left(\hat{\boldsymbol{A}}_2' - \boldsymbol{A}_2'\right) \end{pmatrix} = \frac{1}{T}\sum_t \boldsymbol{Q}_{t-1}\Sigma_e^{-1}\mathrm{vec}(\boldsymbol{E}_t) + o_P(1/\sqrt{T}),
$$

and the proof is completed by an application of the martingale central limit theorem. □

The proof of Corollary 3 is a direct application of the following lemma and Theorems 1, 2. Lemma 1 is regarding the central limit theorems of singular vectors under the fixed dimensional setting. Although some cases are available in the literature, we have not seen any formulation that is exactly the same. Therefore, we provide Lemma 1 and a proof here for the completeness. The proof essentially relies on the matrix perturbation theory.

Lemma 1. *Suppose $\boldsymbol{M}$ is a $p \times p$ symmetric matrix of rank $r \leqslant p$, and let $\boldsymbol{M} = \boldsymbol{U}\Lambda\boldsymbol{U}'$ be its spectral decomposition, where $\boldsymbol{U}$ is a $p \times r$ ortho-normal matrix and Λ is a $r \times r$ diagonal matrix. Denote the j-th diagonal element of Λ by λ_j, and define $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_r)'$. Assume that the λ_j's are distinct. Suppose $\{\hat{\boldsymbol{M}}_n\}$ is a sequence of random matrices and $\{a_n\}$ is a sequence of diverging positive numbers such that*

$$a_n \mathrm{vec}(\hat{\boldsymbol{M}}_n - \boldsymbol{M}) \Rightarrow N(\boldsymbol{0}, \Theta).$$

Let $\hat{\boldsymbol{M}}_n = \hat{\boldsymbol{U}}_n \hat{\Lambda}_n \hat{\boldsymbol{U}}_n'$ be the spectral decomposition of $\hat{\boldsymbol{M}}$ corresponding to the r leading eigenvalues. Define the matrix $\boldsymbol{R}$ as

$$\boldsymbol{R} = (\boldsymbol{I}_r \otimes \boldsymbol{U}, \boldsymbol{I}_r \otimes \boldsymbol{U}^{\perp}) \begin{pmatrix} (\Lambda \otimes \boldsymbol{I}_r - \boldsymbol{I}_r \otimes \Lambda + \boldsymbol{L}_r \boldsymbol{L}_r')^{-1} (\boldsymbol{I}_{r^2} - \boldsymbol{L}_r \boldsymbol{L}_r')(\boldsymbol{U}' \otimes \boldsymbol{U}') \\ (\Lambda^{-1}\boldsymbol{U}') \otimes (\boldsymbol{U}^{\perp})' \end{pmatrix}.$$

Then

$$a_n \mathrm{vec}(\hat{\boldsymbol{U}}_n - \boldsymbol{U}) \Rightarrow N(\boldsymbol{0}, \boldsymbol{R}\Theta\boldsymbol{R}'),$$

and

$$a_n(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}) \Rightarrow N\left[\boldsymbol{0}, \boldsymbol{L}_r'(\boldsymbol{U}' \otimes \boldsymbol{U}')\Theta(\boldsymbol{U} \otimes \boldsymbol{U})\boldsymbol{L}_r\right].$$

Proof. We will show that $\hat{\boldsymbol{U}} = \boldsymbol{U} + O_P(1/a_n)$ and $\hat{\boldsymbol{\lambda}} = \boldsymbol{\lambda} + O_P(1/a_n)$ later.

Expand $\hat{\boldsymbol{M}}\hat{\boldsymbol{U}} = \hat{\boldsymbol{U}}\hat{\Lambda}$ around the true values and omit small order terms, we have

$$(\hat{\boldsymbol{U}} - \boldsymbol{U})\Lambda + \boldsymbol{U}(\hat{\Lambda} - \Lambda) - \boldsymbol{M}(\hat{\boldsymbol{U}} - \boldsymbol{U}) = (\hat{\boldsymbol{M}} - \boldsymbol{M})\boldsymbol{U} + o_P(1/a_n). \tag{22}$$

Multiplying both sides of (22) by $\boldsymbol{U}'$ leads to

$$\boldsymbol{U}'(\hat{\boldsymbol{U}} - \boldsymbol{U})\Lambda - \Lambda\boldsymbol{U}'(\hat{\boldsymbol{U}} - \boldsymbol{U}) + (\hat{\Lambda} - \Lambda) = \boldsymbol{U}'(\hat{\boldsymbol{M}} - \boldsymbol{M})\boldsymbol{U} + o_P(1/a_n).$$

It follows that

$$\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda} = \boldsymbol{L}_r'(\boldsymbol{U}' \otimes \boldsymbol{U}')\mathrm{vec}(\hat{\boldsymbol{M}} - \boldsymbol{M}), \tag{23}$$

and

$$\begin{aligned}&\left(\Lambda\otimes\boldsymbol{I}_r-\boldsymbol{I}_r\otimes\Lambda+\boldsymbol{L}_r\boldsymbol{L}_r'\right)\mathrm{vec}\left[\boldsymbol{U}'(\hat{\boldsymbol{U}}-\boldsymbol{U})\right]\\&\qquad=(\boldsymbol{I}_{r^2}-\boldsymbol{L}_r\boldsymbol{L}_r')(\boldsymbol{U}'\otimes\boldsymbol{U}')\mathrm{vec}(\hat{\boldsymbol{M}}-\boldsymbol{M})+o_P(1/a_n).\end{aligned}\tag{24}$$

The asymptotic distribution of $\hat{\boldsymbol{\lambda}}$ follows (23) immediately. In deriving (24), we have implicitly used the fact that

$$\boldsymbol{U}'(\hat{\boldsymbol{U}}-\boldsymbol{U})+(\hat{\boldsymbol{U}}-\boldsymbol{U})'\boldsymbol{U}=o_P(1/a_n).$$

From (24) we deduce that

$$\begin{aligned}\mathrm{vec}\left[\boldsymbol{U}'(\hat{\boldsymbol{U}}-\boldsymbol{U})\right]&=(\boldsymbol{I}_r\otimes\boldsymbol{U}')\mathrm{vec}(\hat{\boldsymbol{U}}-\boldsymbol{U})\\&=\left(\Lambda\otimes\boldsymbol{I}_r-\boldsymbol{I}_r\otimes\Lambda+\boldsymbol{L}_r\boldsymbol{L}_r'\right)^{-1}(\boldsymbol{I}_{r^2}-\boldsymbol{L}_r\boldsymbol{L}_r')(\boldsymbol{U}'\otimes\boldsymbol{U}')\mathrm{vec}(\hat{\boldsymbol{M}}-\boldsymbol{M})+o_P(1/a_n).\end{aligned}\tag{25}$$

Multiplying both sides of (22) by $(\boldsymbol{U}^\perp)'$ gives

$$(\boldsymbol{U}^\perp)'(\hat{\boldsymbol{U}}-\boldsymbol{U})\Lambda=(\boldsymbol{U}^\perp)'(\hat{\boldsymbol{M}}-\boldsymbol{M})\boldsymbol{U}+o_P(1/a_n),$$

and therefore,

$$(\boldsymbol{U}^\perp)'(\hat{\boldsymbol{U}}-\boldsymbol{U})=(\boldsymbol{U}^\perp)'(\hat{\boldsymbol{M}}-\boldsymbol{M})\boldsymbol{U}\Lambda^{-1}+o_P(1/a_n),$$

and

$$\begin{aligned}\mathrm{vec}\left[(\boldsymbol{U}^\perp)'(\hat{\boldsymbol{U}}-\boldsymbol{U})\right]&=\left[\boldsymbol{I}_r\otimes(\boldsymbol{U}^\perp)'\right]\mathrm{vec}(\hat{\boldsymbol{U}}-\boldsymbol{U})\\&=\left[(\Lambda^{-1}\boldsymbol{U}')\otimes(\boldsymbol{U}^\perp)'\right]\mathrm{vec}(\hat{\boldsymbol{M}}-\boldsymbol{M})+o_P(1/a_n).\end{aligned}\tag{26}$$

Combining (25) and (26), it holds that

$$\begin{aligned}&\begin{pmatrix}\boldsymbol{I}_r\otimes\boldsymbol{U}'\\\boldsymbol{I}_r\otimes(\boldsymbol{U}^\perp)'\end{pmatrix}\mathrm{vec}(\hat{\boldsymbol{U}}-\boldsymbol{U})\\&\quad=\begin{pmatrix}(\Lambda\otimes\boldsymbol{I}_r-\boldsymbol{I}_r\otimes\Lambda+\boldsymbol{L}_r\boldsymbol{L}_r')^{-1}(\boldsymbol{I}_{r^2}-\boldsymbol{L}_r\boldsymbol{L}_r')(\boldsymbol{U}'\otimes\boldsymbol{U}')\\(\Lambda^{-1}\boldsymbol{U}')\otimes(\boldsymbol{U}^\perp)'\end{pmatrix}\mathrm{vec}(\hat{\boldsymbol{M}}-\boldsymbol{M})+o_P(1/a_n).\end{aligned}$$

Since

$$\begin{pmatrix}\boldsymbol{I}_r\otimes\boldsymbol{U}'\\\boldsymbol{I}_r\otimes(\boldsymbol{U}^\perp)'\end{pmatrix}^{-1}=(\boldsymbol{I}_r\otimes\boldsymbol{U},\boldsymbol{I}_r\otimes\boldsymbol{U}^\perp),$$

it follows that

$$\mathrm{vec}(\hat{\boldsymbol{U}}-\boldsymbol{U})=\boldsymbol{R}\,\mathrm{vec}(\hat{\boldsymbol{M}}-\boldsymbol{M})+o_P(1/a_n),$$

and the proof is complete. □

Proof of Theorem 4. We give the proof for the joint EBIC(r_1, r_2). The proof for separate EBIC follows similar arguments, and will be skipped.

Let $\sigma_0^2 := \mathbb{E}\|\boldsymbol{E}_t\|_F^2/(d_1 d_2)$. It is straightforward to show that when $(r_1, r_2) = (k_1, k_2)$

$$\frac{1}{Td_1d_2}\sum_{t=2}^{T}\|\boldsymbol{X}_t - \hat{\boldsymbol{A}}_1^{\text{ls}}(r_1, r_2)\boldsymbol{X}_{t-1}(\hat{\boldsymbol{A}}_2^{\text{ls}}(r_1, r_2))'\|_F^2 \xrightarrow{p} \sigma_0^2;$$

when $r_1 < k_1$ or $r_2 < k_2$,

$$\frac{1}{Td_1d_2}\sum_{t=2}^{T}\|\boldsymbol{X}_t - \hat{\boldsymbol{A}}_1^{\text{ls}}(r_1, r_2)\boldsymbol{X}_{t-1}(\hat{\boldsymbol{A}}_2^{\text{ls}}(r_1, r_2))'\|_F^2 \xrightarrow{p} \sigma_1^2 > \sigma_0^2;$$

and when $r_1 \geqslant k_1$, $r_2 \geqslant k_2$, and at least one of the inequalities is strict,

$$\frac{1}{Td_1d_2}\sum_{t=2}^{T}\|\boldsymbol{X}_t - \hat{\boldsymbol{A}}_1^{\text{ls}}(r_1, r_2)\boldsymbol{X}_{t-1}(\hat{\boldsymbol{A}}_2^{\text{ls}}(r_1, r_2))'\|_F^2 = \sigma_0^2 + O_p(1/T).$$

Then a direct calculation shows that the joint EBIC does not under select or over select the ranks with probability approaching one. □