Journal of the American Statistical Association

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

Simultaneous Decorrelation of Matrix Time Series

Yuefeng Han, Rong Chen, Cun-Hui Zhang & Qiwei Yao

To cite this article: Yuefeng Han, Rong Chen, Cun-Hui Zhang & Qiwei Yao (11 Jan 2023): Simultaneous Decorrelation of Matrix Time Series, Journal of the American Statistical Association, DOI: 10.1080/01621459.2022.2151448

To link to this article: https://doi.org/10.1080/01621459.2022.2151448

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2023, VOL. 00, NO. 0, 1–13: Theory and Methods
https://doi.org/10.1080/01621459.2022.2151448

Taylor & Francis
Taylor & Francis Group

Simultaneous Decorrelation of Matrix Time Series

Yuefeng Han [a], Rong Chen [b], Cun-Hui Zhang [b], and Qiwei Yao [c]

[a]Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN; [b]Department of Statistics, Rutgers University, Piscataway, NJ; [c]Department of Statistics, London School of Economics, London, UK

ABSTRACT

We propose a contemporaneous bilinear transformation for a $p \times q$ matrix time series to alleviate the difficulties in modeling and forecasting matrix time series when p and/or q are large. The resulting transformed matrix assumes a block structure consisting of several small matrices, and those small matrix series are uncorrelated across all times. Hence, an overall parsimonious model is achieved by modeling each of those small matrix series separately without the loss of information on the linear dynamics. Such a parsimonious model often has better forecasting performance, even when the underlying true dynamics deviates from the assumed uncorrelated block structure after transformation. The uniform convergence rates of the estimated transformation are derived, which vindicate an important virtue of the proposed bilinear transformation, that is, it is technically equivalent to the decorrelation of a vector time series of dimension $\max(p, q)$ instead of $p \times q$. The proposed method is illustrated numerically via both simulated and real data examples. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received May 2021
Accepted November 2022

KEYWORDS
Decorrelation transformation; Eigenanalysis; Forecasting; Matrix time series; Uniform convergence rates

1. Introduction

Let $X_t = (X_{t,i,j})$ be a $p \times q$ matrix time series, that is, there are $p \times q$ recorded values at each time from, for example, p individuals and over q indices or variables. Data recorded in this form are increasingly common in this information age, due to the demand to solve practical problems from, among others, signal processing, medical imaging, social networks, IT communication, genetic linkages, industry production and distribution, economic activities and financial markets. Extensive developments on statistical inference for matrix data under iid settings can be found in Negahban and Wainwright (2011), Rohde and Tsybakov (2011), Xia and Yuan (2019), and the references therein. Many matrix sequences are recorded over time, exhibiting significant serial dependence which is valuable for modeling and future prediction. The surge of development in analyzing matrix time series includes bilinear autoregressive models (Hoff 2015; Chen, Xiao, and Yang 2021; Xiao, Han, Chen 2022), factor models based on Tucker's decomposition for tensors (Wang, Liu, and Chen 2019; Chen, Tsay, and Chen 2020; Chen et al. 2020; Han et al. 2020; Chen, Yang, and Zhang 2022; Han, Chen, and Zhang 2022), and factor models based on the tensor CP decomposition (Chang et al. 2021; Han, Zhang, and Chen 2021; Han and Zhang 2022).

A common feature in the aforementioned approaches is dimension-reduction, as the so-called "curse-of-dimensionality" is more pronounced in modeling time series than iid observations, which is exemplified by the limited practical usefulness of vector ARMA (VARMA) models. Note that an unregularized VAR(1) model with dimension p involves at least p^2 parameters. Hence, finding an effective way to reduce the number of parameters is of fundamental importance in modeling and forecasting high dimensional time series. In the context of vector time series, most available approaches may be divided into three categories: (i) regularized VAR or VARMA methods incorporating LASSO or alternative penalties (Basu and Michailidis 2015; Guo, Wang, and Yao 2016; Lin and Michailidis 2017; Zhou and Raskutti 2018; Gao et al. 2019; Ghosh, Khare, and Michailidis 2019; Han and Tsay 2020; Han, Chen, and Wu 2020; Han, Tsay, and Wu 2023), (ii) factor models of various forms (Pena and Box 1987; Bai and Ng 2002; Forni et al. 2005; Lam and Yao 2012), (iii) various versions of independent component analysis (Back and Weigend 1997; Tiao and Tsay 1989; Matteson and Tsay 2011; Huang and Tsay 2014; Chang, Guo, and Yao 2018). Note that the literature in each of the three categories is large, it is impossible to list all the relevant references here.

In this article we propose a new parsimonious approach for analyzing matrix time series, which is in the spirit of independent component analysis. It transforms a $p \times q$ matrix series into the new matrix series of the same size by a contemporaneous bilinear transformation (i.e., the values at different time lags are not mixed together). The new matrix series is divided into several submatrix series and those submatrix series are uncorrelated across all time lags. Hence, an overall parsimonious model can be developed as those submatrix series can be modeled separately without the loss of information on the overall linear dynamics. This is particularly advantageous for forecasting future values. In general cross-serial correlations among different component series are valuable for future

CONTACT Qiwei Yao q.yao@lse.ac.uk Department of Statistics, London School of Economics, London, UK.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

prediction. However, the gain from incorporating those cross-serial correlations directly in a moderately high dimensional model is typically not enough to offset the error resulted from estimating the additional large number of parameters. This is why forecasting a large number of time series together based on a joint time series model can often be worse than that forecasting each component series separately by ignoring cross-serial correlations completely. The proposed transformation alleviates the problem by channeling all cross-serial correlations into the transformed submatrix series and those subseries are uncorrelated with each other across all time lags. Hence, the relevant information can be used more effectively in forecasting within each small models. Note that the transformation is one-to-one, therefore, the good forecasting performance of the transformed series can be easily transformed back to the forecasting for the original matrix series. Our empirical study in Section 4 also demonstrates that, even when the underlying model does not exactly follow the assumed uncorrelated block structure, the decorrelation transformation often produces superior out-of-sample prediction, as the transformation enhances the within block autocorrelations, and the cross-correlations between the blocks are weaker or too weak to be practically useful.

The basic idea of the proposed transformation is similar to the so-called principal component analysis for time series (TS-PCA) of Chang, Guo, and Yao (2018). Actually one might be tempted to stack a $p \times q$ matrix series into a $(pq) \times 1$ vector time series, and apply TS-PCA directly. This requires to search for a $(pq) \times (pq)$ transformation matrix. In contrast, the proposed bilinear transformation is facilitated by two matrices of size, respectively, $p \times p$ and $q \times q$. Indeed our asymptotic analysis indicates that the proposed new bilinear transformation is technically equivalent to a TS-PCA transformation with dimension $\max(p, q)$ instead of $(p \times q)$; see Remark 6 in Section 3. Furthermore the bilinear transformation does not mix rows and columns together, as they may represent radically different features in the data. See, for example, the example in Section 4.2 for which the bilinear transformation also outperforms the vectorized TS-PCA approach in a post-sample forecasting.

The rest of the article is organized as follows. The targeted bilinear decorrelation transformation is characterized in Section 2. We introduce a new normalization under which the required transformation consists of two orthogonal transformations. The orthogonality allows us to accumulate the information from different time lags without cancelation. The estimation of the bilinear transformation boils down to the eigenanalysis of two positive-definite matrices of sizes $p \times p$ and $q \times q$, respectively. Hence, the computation can be carried out on a laptop/PC with p and q equal to a few thousands. Theoretical properties of the proposed estimation are presented in Section 3. The non-asymptotic error bounds of the estimated transformation are derived based on some concentration inequalities (e.g., Rio 2000; Merlevède, Peligrad, Rio 2011); further leading to the uniform convergence rates of the estimation. Numerical illustration with both simulated and real data is reported in Section 4, which shows superior performance in forecasting over the methods without transformation and TS-PCA. Once again it reinforces the fact that the cross-serial correlations are important and useful information for future prediction for a large number of time series, and, however, it is necessary to adopt the proposed decorrelation transformation (or other dimension-reduction techniques) in order to use the information effectively. All technical proofs are relegated to a supplementary material.

2. Methodology

2.1. Decorrelation Transformations

Again, let $X_t = (X_{t,i,j})$ be a $p \times q$ matrix time series. We assume that X_t is weakly stationary in the sense that all the first two moments are finite and time-invariant. Our goal is to seek for a bilinear transformation such that the transformed matrix series admits the following segmentation structure:

$$X_t = BU_tA^\top, \qquad U_t = \begin{pmatrix} U_{t,1,1} & U_{t,1,2} & \cdots & U_{t,1,n_c} \\ U_{t,2,1} & U_{t,2,2} & \cdots & U_{t,2,n_c} \\ \vdots & \vdots & \vdots & \vdots \\ U_{t,n_r,1} & U_{t,n_r,2} & \cdots & U_{t,n_r,n_c} \end{pmatrix}, \tag{1}$$

where A and B are unknown, respectively, $q \times q$ and $p \times p$ invertible constant matrices, and random matrix U_t is unobservable, $U_{t,i,j}$ is a $p_i \times q_j$ matrix with unknown p_i and q_j, $\sum_{i=1}^{n_r} p_i = p$, and $\sum_{j=1}^{n_c} q_j = q$. Furthermore, $\mathrm{cov}\{\mathrm{vec}(U_{t+\tau,i_1,j_1}), \mathrm{vec}(U_{t,i_2,j_2})\} = 0$ for any $(i_1, j_1) \neq (i_2, j_2)$ and any integer τ, that is, all those submatrix series are uncorrelated with each other across all time lags.

Remark 1. (i) The decorrelation bilinear transformation (1) is in the same spirit as TS-PCA of Chang, Guo, and Yao (2018) which transforms linearly a vector time series to a new vector time series of the same dimension but segmented into several subvector series, and those subvectors are uncorrelated across all time lags. Thus, as far as the linear dynamics is concerned, one can model each of those subvector time series separately. It leads to appreciable improvement in future forecasting, as the transformed process encapsulates all the cross-serial correlations into the auto-correlations of those uncorrelated subvector processes.

(ii) In the matrix time series setting, one would be tempted to stack all the elements of X_t into a long vector and to apply TS-PCA of Chang, Guo, and Yao (2018) directly, though it destroys the original matrix structure. This requires to search for the $(pq) \times (pq)$ decorrelation transformation matrix Φ such that $\mathrm{vec}(X_t) = \Phi\mathrm{vec}(U_t)$. The proposed model (1) is to impose a low-dimensional structure $\Phi = A \otimes B$, where $\otimes$ denotes the matrix Kronecker product, such that the technical difficulty is reduced to that of estimating two transformation matrices of size $p \times p$ and $q \times q$, respectively, with significant faster convergence rate; see Remark 6 in Section 3. The empirical results in Section 4 also demonstrate the benefit of maintaining the matrix structure in out-sample prediction performance.

(iii) The segmentation structure in (1) may be too rigid and the division of the submatrices may not be that regular. In practice a small number of row blocks and/or column blocks may be resulted from applying the estimation method proposed in Section 2.3. Then applying the same method again to each submatrix series $U_{t,i,j}$ may lead to a finer segmentation with irregularly sized small blocks.

(iv) Similar to TS-PCA, the desired segmentation structure (1) may not exist in real applications. Then the estimated bilin-

ear transformation, by the method in Section 2.3, leads to an approximate segmentation which encapsulates most significant correlations across different component series into the segmented submatrices while ignoring some small correlations. With enhanced auto-correlations, those submatrix processes become more predictable while those ignored small correlations are typically practically negligible. Consequently the improvement in future prediction still prevails in spite of the lack of an exact segmentation structure. See Chang, Guo, and Yao (2018), and also a real data example in Section 4.2 in which the three versions of segmentation (therefore, at least two of them are approximations) uniformly outperform the various methods without the transformation. A simulation study in Section 4.1 also shows that when the underlying true model deviates from the desired segmentation structure by a moderate amount, the approximated segmentation model produces superior out-sample prediction performance than those without the transformation.

(v) The use of bilinear form in (1) is common in matrix and tensor data analysis. For example, it is used in regression with matrix-type covariates (Basu et al. 2012; Zhou, Li, Zhu 2013; Wang, Zhang, and Dunson 2019), matrix autoregressive model (Hoff 2015; Chen, Xiao, and Yang 2021) and matrix factor model (Wang, Liu, and Chen 2019; Chen, Tsay, and Chen 2020; Chen, Yang, and Zhang 2022).

(vi) For dealing with time series with structural changes, it is possible to allow A and B to be time varying, which is technically more demanding and will be explored elsewhere.

2.2. Normalization and Identification

To simplify the statements, let $\mathbb{E}(X_t) = 0$. This amounts to center the data by the sample mean first, which will not affect the asymptotic properties of the estimation under the stationarity assumption. Thus, $\mathbb{E}(U_t) = 0$.

The terms A, B, and U_t on the RHS of (1) are not uniquely defined, and any A, B, and U_t satisfying the required condition will serve the decorrelation purpose. Therefore, we can take the advantage from this lack of uniqueness to identify the "ideal" A and B to improve the estimation effectiveness. More precisely, by applying a new normalization, we identify A and B to be orthogonal, which facilitates the accumulation of the information from different time lags without cancelation. To this end, we first list some basic facts as follows.

1. $(B, U_t, A^\top)$ in (1) can be replaced by $(BD_1, D_1^{-1}U_tD_2^{-1}, D_2A^\top)$ for any invertible block diagonal matrices D_1 and D_2 with the block sizes, respectively, $(p_1, \ldots, p_{n_r})$ and $(q_1, \ldots, q_{n_c})$. In particular, D_1 and D_2 can be the permutation matrices which permute the columns and rows within each block. In addition, D_1 and D_2 can be the block matrices that permute n_r row blocks and n_c column blocks, respectively. Note both $(p_1, \ldots, p_{n_r})$ and $(q_1, \ldots, q_{n_c})$ are defined by (1) only up to any permutation.
2. The q columns of BU_t can be divided into n_c uncorrelated blocks of sizes $(q_1, \ldots, q_{n_c})$, that is, the columns of BU_t resemble the same pattern of the uncorrelated blocks as those of U_t. Thus, $\Sigma_1^{(u)} \equiv \mathbb{E}(U_t^\top B^\top BU_t)/p$ is a block diagonal matrix with the block sizes $(q_1, \ldots, q_{n_c})$. In the same vein, $\Sigma_2^{(u)} = \mathbb{E}(U_tA^\top AU_t^\top)/q$ is a block diagonal matrix with the block sizes $(p_1, \ldots, p_{n_r})$.
3. Put $B = (B_1, \ldots, B_{n_r})$ and $A = (A_1, \ldots, A_{n_c})$, where B_i is $p \times p_i$ and A_i is $q \times q_i$. Then linear spaces $\mathcal{M}(B_1), \ldots, \mathcal{M}(B_{n_r})$ and $\mathcal{M}(A_1), \ldots, \mathcal{M}(A_{n_c})$ are uniquely defined by (1) up to any permutation, though B and A are not, where $\mathcal{M}(G)$ denotes the linear space spanned by the columns of matrix G. Note that for any $q \times q$ positive definite matrix Σ_1 and $p \times p$ positive definite matrix Σ_2,

$$\begin{aligned}
\Sigma_1^{-1/2}A &= (\Sigma_1^{-1/2}A_1, \ldots, \Sigma_1^{-1/2}A_{n_c}),\\
\mathcal{M}(A_i) &= \Sigma_1^{1/2}\mathcal{M}(\Sigma_1^{-1/2}A_i) \quad i = 1, \ldots, n_c,\\
\Sigma_2^{-1/2}B &= (\Sigma_2^{-1/2}B_1, \ldots, \Sigma_2^{-1/2}B_{n_r}),\\
\mathcal{M}(B_i) &= \Sigma_2^{1/2}\mathcal{M}(\Sigma_2^{-1/2}B_i) \quad i = 1, \ldots, n_r.
\end{aligned}$$

Put $\Sigma_1^{(x)} = \mathbb{E}(X_t^\top X_t)/p$ and $\Sigma_2^{(x)} = \mathbb{E}(X_tX_t^\top)/q$. Proposition 1 indicates that if we replace X_t in (1) by a normalized version $\{\Sigma_2^{(x)}\}^{-1/2}X_t\{\Sigma_1^{(x)}\}^{-1/2}$, we can treat both A and B in (1) as orthogonal matrices. This orthogonality plays a key role in combining together the information from different time lags in our estimation (see the proof of Proposition 2 in the Appendix, supplementary materials).

Proposition 1. Let both $\Sigma_1^{(x)}$ and $\Sigma_2^{(x)}$ be invertible. Then it holds that

$$X_t = \{\Sigma_2^{(x)}\}^{1/2}B_*U_t^*A_*^\top\{\Sigma_1^{(x)}\}^{1/2}, \tag{2}$$

where U_t^* admits the same segmentation structure as U_t in (1), and A_* and B_* are, respectively, $q \times q$ and $p \times p$ orthogonal matrices. More precisely,

$$\begin{aligned}
U_t^* &= \{\Sigma_2^{(u)}\}^{-1/2}U_t\{\Sigma_1^{(u)}\}^{-1/2},\\
A_* &= \{\Sigma_1^{(x)}\}^{-1/2}A\{\Sigma_1^{(u)}\}^{1/2}, B_* = \{\Sigma_2^{(x)}\}^{-1/2}B\{\Sigma_2^{(u)}\}^{1/2}.
\end{aligned} \tag{3}$$

The proof of Proposition 1 is almost trivial. First note that both $\Sigma_1^{(u)}$, $\Sigma_2^{(u)}$ are invertible. Then (2) follows from (1) and (3) directly. The orthogonality of, for example, B_* follows from equality $\mathbb{E}[\{\Sigma_2^{(x)}\}^{-1/2}X_tX_t^\top\{\Sigma_2^{(x)}\}^{-1/2}]/q = I_p$, (2) and (3). Also note that $\{\Sigma_2^{(u)}\}^{-1/2}$ and $\{\Sigma_1^{(u)}\}^{-1/2}$ are of the same block diagonal structure as, respectively, $\Sigma_2^{(u)}$ and $\Sigma_1^{(u)}$. This implies that U_t^* admits the same segmentation structure as U_t.

Write $A_* = (A_{*1}, \ldots, A_{*n_c}), B_* = (B_{*1}, \ldots, B_{*n_r})$, where A_{*j} has q_j columns and B_{*i} has p_i columns. Then $U_{t,i,j}^* = A_{*j}\{\Sigma_2^{(x)}\}^{-1/2}X_t\{\Sigma_1^{(x)}\}^{-1/2}B_{*i}$. Note that $(B_*, U_t^*, A_*^\top)$ in (2) are (still) not uniquely defined, similar to the property (i) above. In fact, only the linear spaces $\mathcal{M}(B_{*1}), \ldots, \mathcal{M}(B_{*n_r})$ and $\mathcal{M}(A_{*1}), \ldots, \mathcal{M}(A_{*n_c})$ are uniquely defined. Proposition 2 shows that we can take the orthonormal eigenvectors of two properly defined positive definite matrices as the columns of A_* and B_*. With A_* and B_* specified, the segmented U_t^* can be solved from (2) directly. Let

$$\begin{aligned}
V_{\tau,i,j}^{(1)} &= \{\Sigma_1^{(x)}\}^{-1/2}\mathbb{E}(X_{t+\tau}^\top E_{i,j}X_t)\{\Sigma_1^{(x)}\}^{-1/2},\\
V_{\tau,i,j}^{(2)} &= \{\Sigma_2^{(x)}\}^{-1/2}\mathbb{E}(X_{t+\tau}E_{i,j}X_t^\top)\{\Sigma_2^{(x)}\}^{-1/2},
\end{aligned}$$

where $E_{i,j}$ is the unit matrix with 1 at position (i, j) and 0 elsewhere, and $E_{i,j}$ is $p \times p$ in the first equation, and $q \times q$ in the second equation. For a prespecified integer $\tau_0 \geq 1$, let

$$W^{(1)} = \sum_{\tau=-\tau_0}^{\tau_0} \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{V_{\tau,i,j}^{(1)} \big(V_{\tau,i,j}^{(1)}\big)^\top}{p^2},$$
$$W^{(2)} = \sum_{\tau=-\tau_0}^{\tau_0} \sum_{i=1}^{q} \sum_{j=1}^{q} \frac{V_{\tau,i,j}^{(2)} \big(V_{\tau,i,j}^{(2)}\big)^\top}{q^2}. \tag{4}$$

Proposition 2. Let both $\Sigma_1^{(x)}$ and $\Sigma_2^{(x)}$ be invertible, and all the eigenvalues of $W^{(i)}$ be distinct, $i = 1, 2$. Also let $\tau_0 \geq 1$. Then the q orthonormal eigenvectors of $W^{(1)}$ can be taken as the columns of A_*, and the p orthonormal eigenvectors of $W^{(2)}$ can be taken as the columns of B_*.

The proposition above does not give any indication on how to arrange the columns of A_* and B_*, which should be ordered according to the latent uncorrelated block structure (1). We address this issue in Section 2.3.

Remark 2. The representation (1) suffers from the indeterminacy due to the fact that the column blocks and row blocks can be permuted, and the columns and rows within each block can be rotated, see property (i) above. However, this indeterminacy does not have material impact on identifying the desired structure (1), as any one of such representations will serve the purpose. The developed asymptotic theory guarantees that the proposed estimator converges to a representation which fulfills the conditions imposed on (1).

Remark 3. In case that $W^{(1)}, W^{(2)}$ have tied eigenvalues and the corresponding eigenvectors across different blocks, Proposition 2 no longer holds. Then the blocks sharing the tied eigenvalues cannot be separated. To avoid the tied eigenvalues, we may use different values of τ_0, or different form of $W^{(i)}$. For example, we may replace $W^{(1)}$ by

$$W^{(1,f)} = \sum_{\tau=-\tau_0}^{\tau_0} \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{f\left(V_{\tau,i,j}^{(1)} V_{\tau,i,j}^{(1)\top}\right)}{p^2},$$

where $f(V)$ is a function of a symmetric matrix V in which $f(V) = \Gamma D^{(f)} \Gamma^\top$, $V = \Gamma D \Gamma^\top$ is the eigen-decomposition for V, and $D^{(f)} = \mathrm{diag}(f(d_1), \ldots, f(d_q))$ is the diagonal-element-wise transformation of the diagonal matrix D of the eigenvalues.

In fact the condition that all eigenvalues of $W^{(i)}$ are different can be relaxed. For example, we only require any two eigenvalues of $W^{(i)}$ corresponding two different blocks to be different while the eigenvalues corresponding to the same black can be the same. See also a similar condition in the eigen-gap Δ in (14) in Section 3.

Remark 4. In practice, we need to specify τ_0 in (4). In principle any $\tau_0 \geq 1$ can be used for the purpose of segmentation. A larger τ_0 may capture more lag-dependence, but may also risk adding more "noise" when the dependence decays fast. Since autocorrelation is typically at its strongest at small lags, a relatively small τ_0, such as $\tau_0 \leq 5$, is often sufficient (Lam, Yao, and Bathia 2011; Chang, Guo, and Yao 2018; Wang, Liu, and Chen 2019; Chen, Yang, and Zhang 2022).

2.3. Estimation

The estimation for A_* and B_*, defined in (3), is based on the eigenanalysis of the sample versions of matrices defined in (4). To this end, let

$$\widehat{\Sigma}_1^{(x)} = \frac{1}{Tp} \sum_{t=1}^{T} X_t^\top X_t, \quad \widehat{\Sigma}_2^{(x)} = \frac{1}{Tq} \sum_{t=1}^{T} X_t X_t^\top, \tag{5}$$

$$\widehat{V}_{\tau,i,j}^{(1)} = \{\widehat{\Sigma}_1^{(x)}\}^{-1/2} \sum_{t=1}^{T-\tau} \frac{X_{t+\tau}^\top E_{i,j} X_t}{T-\tau} \{\widehat{\Sigma}_1^{(x)}\}^{-1/2},$$
$$\widehat{V}_{\tau,i,j}^{(2)} = \{\widehat{\Sigma}_2^{(x)}\}^{-1/2} \sum_{t=1}^{T-\tau} \frac{X_{t+\tau} E_{i,j} X_t^\top}{T-\tau} \{\widehat{\Sigma}_2^{(x)}\}^{-1/2}, \tag{6}$$

$$\widehat{W}^{(1)} = \frac{1}{p^2} \sum_{\tau=-\tau_0}^{\tau_0} \sum_{i=1}^{p} \sum_{j=1}^{p} \widehat{V}_{\tau,i,j}^{(1)} \big(\widehat{V}_{\tau,i,j}^{(1)}\big)^\top,$$

$$\widehat{W}^{(2)} = \frac{1}{q^2} \sum_{\tau=-\tau_0}^{\tau_0} \sum_{i=1}^{q} \sum_{j=1}^{q} \widehat{V}_{\tau,i,j}^{(2)} \big(\widehat{V}_{\tau,i,j}^{(2)}\big)^\top. \tag{7}$$

Performing the eigenanalysis for $\widehat{W}^{(1)}$, and arranging the order of the resulting q orthonormal eigenvectors $\widehat{\gamma}_1, \ldots, \widehat{\gamma}_q$ by the algorithm below, we take the reordered eigenvectors as the columns of $\widehat{A}_*$. The estimator $\widehat{B}_*$ is obtained in the same manner via the eigenanalysis for $\widehat{W}^{(2)}$. Then by (2), we obtain the transformed matrix series

$$\widehat{U}_t^* = \widehat{B}_*^\top \{\widehat{\Sigma}_2^{(x)}\}^{-1/2} X_t \{\widehat{\Sigma}_1^{(x)}\}^{-1/2} \widehat{A}_*. \tag{8}$$

Note that the estimation for A_* and that for B_* are carried out separately. They do not interfere with each other.

Now we present an algorithm to determine the order of the columns for $\widehat{A}_*$. By (2), the columns of $Z_t \equiv X_t \{\Sigma_1^{(x)}\}^{-1/2} A_*$ are divided into the n_c uncorrelated blocks. Define

$$\widehat{Z}_t \equiv (\widehat{z}_{t,i,j}) = X_t \{\widehat{\Sigma}_1^{(x)}\}^{-1/2} (\widehat{\gamma}_1, \ldots, \widehat{\gamma}_q). \tag{9}$$

We divide the columns of $\widehat{Z}_t$ into uncorrelated blocks according to the pairwise maximum cross-correlations between the columns. Specifically, the maximum cross-correlation between the kth and ℓth columns is defined as

$$\widehat{\rho}_{k,\ell} = \max_{1 \leq i,j \leq p, |\tau| \leq \tau_1} \left| \widehat{\mathrm{corr}}(\widehat{z}_{t+\tau,i,k}, \widehat{z}_{t,j,\ell}) \right|$$
$$= \max_{1 \leq i,j \leq p, |\tau| \leq \tau_1} \frac{\left| \widehat{\gamma}_k^\top \widehat{V}_{\tau,i,j}^{(1)} \widehat{\gamma}_\ell \right|}{\left\{ \widehat{\gamma}_k^\top \widehat{V}_{0,i,i}^{(1)} \widehat{\gamma}_k \, \widehat{\gamma}_\ell^\top \widehat{V}_{0,j,j}^{(1)} \widehat{\gamma}_\ell \right\}^{1/2}} \tag{10}$$

The second equality above follows from (6), (7), and (9). In the above expression, $\tau_1 \geq 1$ is a user-defined tuning parameter (See Remark 5). To determine all significantly correlated pairs of variable, rearrange $\widehat{\rho}_{k,\ell}$, $1 \leq k < \ell \leq q$, in the descending order: $\widehat{\rho}_{(1)} \geq \cdots \geq \widehat{\rho}_{(q_0)}$ and define

$$\widehat{r} = \arg \max_{1 \leq j \leq q_0} \frac{\widehat{\rho}_{(j)} + \delta_T}{\widehat{\rho}_{(j+1)} + \delta_T}, \tag{11}$$

where $\delta_T > 0$ is a small constant. We take the $\widehat{r}$ pair of columns corresponding to $\widehat{\rho}_{(1)}, \ldots, \widehat{\rho}_{(\widehat{r})}$ as correlated pairs, and treat the rest as uncorrelated pairs. The intuition is that $\rho_{(r)}/\rho_{(r+1)}$ is ∞ if $\rho_{(r)} > 0$ but $\rho_{(r+1)} = 0$. The use of δ_T is to smooth out the large variation of $\widehat{\rho}_{(j)}/\widehat{\rho}_{(j+1)}$ when both $\rho_{(j)}$ and $\rho_{(j+1)}$ are small. Similar ideas have also been used in determining the number of factors in Lam and Yao (2012), Ahn and Horenstein (2013), and Han, Chen, and Zhang (2022).

With the $\widehat{r}$ identified significantly correlated pairs, a connection graph is built, with the column indexes as the vertices and the edges between the identified correlated column pairs. The number of unconnected sub-graphs is the estimated number of blocks $\widehat{n}_c$, and each of maximum connected sub-graphs forms the estimated groups $\widehat{G}_i$, $i = 1, \ldots \hat{n}_c$. The corresponding $\widehat{n}_c$ groups of $\widehat{\gamma}_1, \ldots, \widehat{\gamma}_q$ are taken as the columns of $\widehat{A}_{*i}$, $i = 1, \ldots, \hat{n}_c$. Algorithmically, start with q groups with one column in each group; then iteratively check all pairs of groups and merge two groups together if there exists at least one pair of columns (one in each group) are significantly correlated; and stop when no groups can be merged.

The finite sample performance of the above algorithm can be improved by prewhitening each column time series of $\widehat{Z}_t$. This makes $\widehat{\rho}_{k,\ell}$, for different (k, ℓ), more comparable. See Remark 2(iii) of Chang, Guo, and Yao (2018). In practice the prewhitening can be carried out by fitting each column time series a VAR model with the order between 0 and 5 determined by AIC. The resulting residual series is taken as a prewhitened series.

Remark 5. (i) The ratio estimator in (11) picks the $\widehat{r}$ most correlated pairs of columns, and ignores the other small correlations in constructing the segmentation structure. Theorem 3 in Section 3 shows that the partition of $\widehat{A}_*$ into $\{\widehat{A}_{*1}, \ldots, \widehat{A}_{*\widehat{n}_c}\}$, determined by the above algorithm, provides a consistent estimation for the column segmentation of A_*.

(ii) There are two tuning parameters used in the procedure, the maximum lag τ_0 used in constructing $W^{(1)}$ and $W^{(2)}$ in (4) and the maximum lag τ_1 used in measuring the dependency between two columns (rows) in (10). Lag τ_1 is usually a sufficiently large integer, for example, between 10 and 20, in the spirit of the rule of thumb of Box and Jenkins (1970). Different values of τ_0 and τ_1 may, or may not, lead to different segmentation. However, the impact on, for example, future prediction is minimum, as the most information on linear dynamics is encapsulated in the most correlated pairs, as indicated by numerical examples in the Appendix, supplementary materials.

The ordering for columns of $\widehat{B}_*$ is arranged, in the same manner as above, by examining the pairwise correlations among the columns of

$$X_t^\top \{\widehat{\Sigma}_2^{(x)}\}^{-1/2} (\widehat{\gamma}_1^{(2)}, \ldots, \widehat{\gamma}_p^{(2)}),$$

where $\widehat{\gamma}_1^{(2)}, \ldots, \widehat{\gamma}_p^{(2)}$ are now the p orthonormal eigenvectors of $\widehat{W}^{(2)}$.

3. Theoretical Properties

To gain more appreciation of the methodology, we will show the consistency of the proposed detection method of uncorrelated components. We mainly focus on the estimation error and the ordering of $\widehat{A}_*$, as those for $\widehat{B}_*$ are similar. Denote by $X_{t,i,\cdot}$ the ith row of X_t, and $X_{t,i,j}$ the (i, j)th element of X_t. For matrix $V = (V_{ij})$, let $\|V\|_{\text{op}}$ denote the spectrum norm, and $\|V\|_{\max} = \max_{i,j} |V_{ij}|$. We introduce some regularity conditions first.

Assumption 1. Assume exponentially decay coefficients of the strong α-mixing condition, $\alpha(k) \le \exp\big(-c_0 k^{r_1}\big)$ for some constant $c_0 > 0$ and $0 < r_1 \le 1$, where

$$\alpha(k) = \sup_t \Big\{ \Big| \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) - \mathbb{P}(\mathcal{E}_1)\mathbb{P}(\mathcal{E}_2) \Big| : \mathcal{E}_1 \in \sigma(X_s, s \le t), \mathcal{E}_2 \in \sigma(X_s, s \ge t + k) \Big\}. \quad (12)$$

Here for any random variable/vector/matrix X, $\sigma(X)$ is understood to be the σ-field generated by X. Assumption 1 allows a very general class of time series models, including causal ARMA processes with continuously distributed innovations; see Bradley (2002), and also Section 2.6 of Fan and Yao (2003). The restriction $r_1 \le 1$ is introduced only for simple presentation.

Assumption 2. Assume $\mathbb{E}X_t = 0$. There exist certain finite constants c_*, c^*, such that

$$\sup_{i \le p, \|u\|_2 = 1} \text{var}\Big(X_{t,i,\cdot} u\Big) \le c^*, \quad \inf_{\|u\|_2 = 1} \mathbb{E}\|X_t u\|_2^2 / p \ge c_*. \quad (13)$$

Note that $\sup_{\|u\|_2=1} \text{var}\Big(X_{t,i,\cdot}^\top u\Big) = \|\mathbb{E} X_{t,i\cdot}^\top X_{t,i\cdot}\|_{\text{op}}$ and $\mathbb{E}\|X_t u\|_2^2/p = u^\top \Sigma_1^{(x)} u$. Assumption 2 on the eigenvalues is a common assumption in the high dimensional setting, for instance, Bickel and Levina (2008a, 2008b) and Xia, Cai, and Cai (2018).

Assumption 3. For any $x > 0$, $\max_{1 \le i \le p, 1 \le j \le q} \mathbb{P}\left(|X_{t,i,j}| \ge x\right) \le c_1 \exp\left(-c_2 x^{r_2}\right)$, for some constant $c_1, c_2 > 0$ and $0 < r_2 \le 2$.

Assumption 3 requires that the tail probability of each individual series of X_t decay exponentially fast. In particular, when $r_2 = 2$, each $X_{t,i,j}$ is sub-Gaussian.

Write the eigenvalue-eigenvector decomposition of $W^{(1)}$ as

$$W^{(1)} = \Gamma^{(1)} D \big(\Gamma^{(1)}\big)^\top,$$

where $D = \text{diag}(\lambda_1, \ldots, \lambda_q)$ with $\lambda_1 \ge \cdots \ge \lambda_q$, and $\Gamma^{(1)} = (\gamma_1, \ldots, \gamma_q)$. By the structure assumption in (1), the index set $\{1, \ldots, q\}$ is partitioned into n_c subsets $G_1, \ldots, G_{n_c}$ such that the columns of different sub-matrices $X_t \Big(\Sigma_1^{(1)}\Big)^{-1/2} \Gamma_{G_i}^{(1)}$, $i = 1, \ldots, n_c$, are uncorrelated with each other across all time lags. Define eigen-gap

$$\Delta = \min_{1 \le i < j \le n_c} \min_{k \in G_i, \ell \in G_j} |\lambda_k - \lambda_\ell|. \quad (14)$$

The following theorem provides the nonasymptotic bounds for the estimators of $V_{\tau,i,j}^{(1)}$, $W^{(1)}$ and A_{*j}, $1 \le i, j \le q$ under both exponential decay α mixing condition and exponential tail condition of X_t.

Theorem 1. Suppose Assumptions 1–3 hold with constants c_*, c^*, r_1, r_2, and τ_0 is a finite constant. Let $1/\beta_1 = 1/r_1 + 2/r_2$ and $1/\beta_2 = 1/r_1 + 1/r_2$. Then,

$$\left\|\widehat{V}^{(1)}_{\tau,i,j} - V^{(1)}_{\tau,i,j}\right\|_{\text{op}} \le C_1\eta_{T,p,q},$$

$$\forall 1 \le \tau \le \tau_0, 1 \le i,j \le p, \tag{15}$$

$$\left\|\widehat{W}^{(1)} - W^{(1)}\right\|_{\text{op}} \le C_1\eta_{T,p,q}, \tag{16}$$

in an event Ω_T with probability at least $1-\epsilon_T$, where

$$\eta_{T,p,q} = q\left(\sqrt{\frac{\log(pq/\epsilon_T)}{T}} + \frac{[\log(Tpq/\epsilon_T)]^{1/\beta_1}}{T} + \frac{[\log(Tpq/\epsilon_T)]^{2/\beta_2}}{T^2}\right), \tag{17}$$

C_1 is a constant depending on c_*, c^*, r_1, r_2 only. Moreover, there exists $\widetilde{A}_* \equiv (\widetilde{A}_{*1}, \ldots, \widetilde{A}_{*n_c})$ of which the columns are a permutation of $(\widehat{\gamma}_1, \ldots, \widehat{\gamma}_q)$ such that

$$\left\|\widetilde{A}_{*j}\widetilde{A}_{*j}^\top - A_{*j}A_{*j}^\top\right\|_{\text{op}} \le C_2\eta_{T,p,q} \quad \text{for} \quad 1 \le j \le n_c \tag{18}$$

in the same event Ω_T provided $C_1\eta_{T,p,q} \le \Delta/2$, where C_2 is a constant depending on c_*, c^*, r_1, r_2 only.

Remark 6. (i) Let P_G be the projection operator onto the column space of G, which can be written as $P_G = G(G^\top G)^{-1}G^\top$. Under the conditions of Theorem 1, we can obtain $\|P_{\widetilde{A}_{*j}\otimes\widetilde{B}_{*i}} - P_{A_{*j}\otimes B_{*i}}\|_{\text{op}} = O_{\mathbb{P}}(\max\{p,q\}[(\log(pq)/T)^{1/2} + \log(Tpq)^{1/\beta_1}/T])$ for each sub-group $1 \le i \le n_r, 1 \le j \le n_c$, where the columns of $\widetilde{B}_*$ are a permutation of the orthonormal eigenvectors of $\widehat{W}_{(2)}$. If we stack all the elements of X_t into a long vector such that $\text{vec}(X_t) = \Phi\text{vec}(U_t)$ and apply TS-PCA of Chang, Guo, and Yao (2018) directly, the estimation of the decorrelation transformation matrix would satisfy $\|P_{\widetilde{\Phi}_k} - P_{\Phi_k}\|_{\text{op}} = O_{\mathbb{P}}(pq[(\log(pq)/T)^{1/2} + \log(Tpq)^{1/\beta_1}/T])$ for $1 \le k \le n_r n_c$ and $\Phi = (\Phi_1, \ldots, \Phi_{n_r n_c})$. Obviously, $\|P_{\widetilde{A}_{*j}\otimes\widetilde{B}_{*i}} - P_{A_{*j}\otimes B_{*i}}\|_{\text{op}}$ has much sharper rate which is equivalent to that for the TS-PCA estimation with dimension $\max(p,q)$.

(ii) The consistency of $\widehat{W}^{(1)}$ requires $\max(p,q) = o(\sqrt{T})$. When A and B, or equivalently $W^{(1)}$ and $W^{(2)}$, have certain sparsity structures, this condition can be further relaxed. In TS-PCA, threshold estimator of Bickel and Levina (2008a) is employed to construct a sparse $\widehat{W}^{(1)}$, and then the convergence rates are improved to allow the dimension to be much larger than T (Chang, Guo, and Yao 2018). The similar extension can be established in our setting.

When the decays of the α-mixing coefficients and the tail probabilities of X_t are slower than Assumptions 1 and 3, we impose the following Assumptions 4 and 5 instead. These conditions ensure the Fuk-Nagaev-type inequalities for α-mixing processes; see also Rio (2000), Liu, Xiao, and Wu (2013), Wu and Wu (2016), and Zhang and Wu (2017).

Assumption 4. Let the α-mixing coefficients satisfy the condition $\alpha(k) \le c_0 k^{-r_1}$, where $\alpha(k)$ is defined in (12), $c_0 > 0$ and $r_1 > 1$.

Assumption 5. For any $x > 0$, $\max_{1\le i\le p, 1\le j\le q}\mathbb{P}\left(|X_{t,i,j}| \ge x\right) \le c_1 x^{-2r_2}$, for some constant $c_1 > 0$ and $r_2 > 1$.

Theorem 2 presents the uniform convergence rate for $\widehat{V}^{(1)}_{\tau,i,j}$ and $\widehat{W}^{(1)}$, $1 \le i,j \le q$. It is intuitively clear that the rate is much slower than that in Theorem 1.

Theorem 2. Suppose Assumptions 2, 4, and 5 hold with constants c_*, c^*, r_1, r_2, and τ_0 is a finite constant. Let $\beta_3 = r_2(r_1+1)/(r_1+r_2) > 1$. Then, (15) and (16) hold in an event Ω_T with probability at least $1-\epsilon_T$, where

$$\eta_{T,p,q} = q\left(\frac{(Tpq/\epsilon_T)^{1/\beta_3}}{T} + \sqrt{\frac{\log(pq/\epsilon_T)}{T}}\right), \tag{19}$$

and C_1 depends on c_*, c^*, r_1, r_2 only. Similarly, if $C_1\eta_{T,p,q} \le \Delta/2$, then there exists $\widetilde{A}_* \equiv (\widetilde{A}_{*1}, \ldots, \widetilde{A}_{*n_c})$ of which the columns are a permutation of $(\widehat{\gamma}_1, \ldots, \widehat{\gamma}_q)$ such that (18) holds in the same event Ω_T for some positive constant C_2 depending on c_*, c^*, r_1, r_2 only.

Theorem 3 implies that the partition of $\widehat{A}_*$ into $\{\widehat{A}_{1*}, \ldots, \widehat{A}_{*\widehat{n}_c}\}$ in Section 2.3 provides a consistent estimation for the column segmentation of $A_* = (A_{*1}, \ldots, A_{*n_c})$. To this end, let

$$\rho_{k,\ell} = \max_{1\le i,j\le p, |\tau|\le\tau_1} \frac{\left|\gamma_k^\top V^{(1)}_{\tau,i,j}\gamma_\ell\right|}{\left\{\gamma_k^\top V^{(1)}_{0,i,i}\gamma_k\, \gamma_\ell^\top V^{(1)}_{0,j,j}\gamma_\ell\right\}^{1/2}}. \tag{20}$$

See also (10). By (2), the columns of $Z_t \equiv X_t\{\Sigma_1^{(x)}\}^{-1/2}A_*$ are divided into the n_c uncorrelated blocks. We assume that those n_c blocks can also be obtained by the algorithm in Section 2.3 with $\widehat{\rho}_{k,\ell}$ replaced by $\rho_{k,\ell}\cdot I(\rho_{k,\ell} \ge \rho)$ for some constant $\rho > 0$. Denote by $G_1, \ldots, G_{n_c}$, respectively, the indices of the components of Z_t in each of those n_c blocks. Denote by $\widehat{G}_1, \ldots, \widehat{G}_{\widehat{n}_c}$, respectively, the indices of the components of $\widehat{Z}_t$ in each of the $\widehat{n}_c$ uncorrelated blocks identified by the algorithm in Section 2.3. Rearrange of the order of $\widehat{G}_1, \ldots, \widehat{G}_{\widehat{n}_c}$ if necessary. Then the theorem below implies that $\mathbb{P}(\hat{n}_c = n_c) \to 1$ and $\mathbb{P}(\widehat{G}_i = G_i | \hat{n}_c = n_c) \to 1$ for $1 \le i \le n_c$.

Theorem 3. Suppose conditions of Theorems 1 or 2 hold and τ_1 is a finite constant. Suppose

$$\kappa_1\eta_{T,p,q} < \Delta, \qquad \kappa_2\eta_{T,p,q} < \rho^*, \tag{21}$$

where κ_1, κ_2 are certain positive constants depending on c_*, c^*, r_1, r_2 only, $\eta_{T,p,q}$ is defined in (17) or (19) and depends on ϵ_T. Then in an event Ω_T with probability at least $1-\epsilon_T$, we have

$$\widehat{n}_c = n_c \text{ and } \widehat{G}_i = G_i, \quad 1 \le i \le n_c.$$

Remark 7. The first inequality in (21) requires that the minimum eigen-gap Δ between different uncorrelated groups is sufficiently larger than the estimation error $\eta_{T,p,q}$, such that all the groups are identifiable. The second inequality in (21) ensures that there are no cross-group edges among $G_1, \ldots, G_{n_c}$. The constants κ_1 and κ_2 are specified in the proof of the theorem.

Table 1. Example 1—Means and standard errors (SE) of $D(\widehat{A}_*, A_*)$ and $D(\widehat{B}_*, B_*)$ in a simulation with 1000 replications.

		$D(\widehat{A}_*, A_*)$		$D(\widehat{B}_*, B_*)$			$D(\widehat{A}_*, A_*)$		$D(\widehat{B}_*, B_*)$	
T	(q, p)	Mean	SE	Mean	SE	(q, p)	Mean	SE	Mean	SE
100	(4, 4)	0.135	0.103	0.116	0.094	(4, 8)	0.124	0.096	0.169	0.056
500		0.081	0.080	0.052	0.065		0.074	0.078	0.089	0.047
1000		0.048	0.062	0.050	0.063		0.057	0.066	0.056	0.037
5000		0.021	0.042	0.020	0.042		0.019	0.039	0.032	0.027
100	(8, 8)	0.172	0.057	0.172	0.055	(16, 16)	0.204	0.031	0.206	0.032
500		0.098	0.048	0.086	0.044		0.132	0.032	0.131	0.031
1000		0.074	0.045	0.070	0.041		0.097	0.027	0.093	0.026
5000		0.034	0.029	0.035	0.030		0.041	0.017	0.042	0.018
100	(32, 32)	0.224	0.018	0.225	0.018	(100, 16)	0.251	0.006	0.187	0.032
500		0.162	0.019	0.161	0.020		0.212	0.008	0.127	0.031
1000		0.127	0.018	0.125	0.018		0.185	0.009	0.099	0.028
5000		0.058	0.013	0.058	0.013		0.106	0.008	0.044	0.019

4. Numerical Properties

4.1. Simulation

We illustrate the proposed decorrelation method with simulated examples. We always set $\tau_0 = 5$ in (7), and $\tau_1 = 15$ in (10). The prewhitening, as stated at the end of Section 2.3, is always applied in determining the groups of the columns of $\widehat{A}_*$ and $\widehat{B}_*$. As the true transformation matrices A_* and B_* defined in (3) cannot be computed easily even for simulated examples, we use the following proxies instead

$$A_* = \{\widehat{\Sigma}_1^{(x)}\}^{-1/2} A \{\widehat{\Sigma}_1^{(u)}\}^{1/2}, \qquad B_* = \{\widehat{\Sigma}_2^{(x)}\}^{-1/2} B \{\widehat{\Sigma}_2^{(u)}\}^{1/2},$$

where $\widehat{\Sigma}_1^{(x)}$, $\widehat{\Sigma}_2^{(x)}$ are given in (5), and

$$\widehat{\Sigma}_1^{(u)} = \frac{1}{Tp} \sum_{t=1}^{T} U_t^\top B^\top B U_t, \qquad \widehat{\Sigma}_2^{(u)} = \frac{1}{Tq} \sum_{t=1}^{T} U_t A A^\top U_t^\top.$$

To abuse the notation, we still use A_*, B_* to denote their proxies in this section since the true A_*, B_* never occur in our computation. For each setting, we replicate the simulation 1000 times.

Example 1. We consider model (1) in which all the component series of U_t are independent and ARMA(1, 2) of the form:

$$z_t = b z_{t-1} + \epsilon_t + a_1 \epsilon_{t-1} + a_2 \epsilon_{t-2}, \tag{22}$$

where b is drawn independently from the uniform distribution on $(-0.98, -0.5) \cup (0.5, 0.98)$, a_1, a_2 are drawn independently from the uniform distribution on $(-0.98, -0.3) \cup (0.3, 0.98)$, and ϵ_t are independent and $N(0, 1)$. The elements of A and B are drawn independently from $U(-1, 1)$. For this example, we assume that we know $n_r = p$ and $n_c = q$ in (1). The focus is to investigate the impact of p and q on the estimation for B and A.

Performing the eigenanalysis for $\widehat{W}^{(1)}$ and $\widehat{W}^{(2)}$ defined in (7), we take the resulting two sets of orthonormal eigenvectors as the columns of $\widehat{A}_*$ and $\widehat{B}_*$, respectively. To measure the estimation error, we define

$$D(\widehat{A}_*, A_*) = \frac{1}{2q(\sqrt{q} - 1)} \sum_{j=1}^{q} \Big(\frac{1}{\max_i |d_{i,j}|} + \frac{1}{\max_i |d_{j,i}|} - 2 \Big),$$

where $d_{i,j}$ is the (i, j)-h element of matrix $\widehat{A}_*^\top A_*$. Note that $D(\widehat{A}_*, A_*)$ is always between 0 and 1, and $D(\widehat{A}_*, A_*) = 0$ if $\widehat{A}_*$ is a column permutation of A_*.

We set $T = 100, 500, 1000, 5000$, and $p, q = 4, 8, 16, 32, 100$. The means and standard errors of $D(\widehat{A}_*, A_*)$ and $D(\widehat{B}_*, B_*)$ over the 1000 replications are reported in Table 1. As expected, the estimation errors decrease as T increases. Furthermore the error in estimating A_* increases as q increases, and that in estimating B_* increases as p increases. Also noticeable is the fact that the quality of the estimation for A_* depends on q only and that for B_* depends on p only, as $D(\widehat{A}_*, A_*)$ is about the same for $(q, p) = (4, 4)$ and $(q, p) = (4, 8)$, and $D(\widehat{B}_*, B_*)$ is about the same for $(q, p) = (4, 8)$ and $(q, p) = (8, 8)$. When $(q, p) = (32, 32)$, each A and B contains 1024 unknown parameters. The quality of their estimation with $T = 100$ is not ideal, but not substantially worse than that for $(q, p) = (16, 16)$, and the estimation is very accurate with $T = 5000$. For $(q, p) = (100, 16)$, matrix time series X_t contains 1600 component series. The estimation for A_* is not accurate enough as $q = 100$, while the estimation for B_* remains as good as that for $(q, p) = (16, 16)$, and is clearly better than that for $(q, p) = (32, 32)$.

Example 2. To examine the performance of the proposed method for identifying uncorrelated blocks, we consider now a only column-segmentation model

$$X_t = U_t A^\top,$$

where X_t, U_t are $p \times q$ matrix time series. Note that adding the row transformation matrix B in the model entails the application of the same method twice without adding any new insights on the performance of the method, as the estimation for A_* and that for B_* are performed separately. See Section 2.3.

The transformation matrix A in the above model is generated in the same way as in Example 1. All the element time series of U_t, except those in columns 2, 3, and 5, are simulated independently from ARMA(1,2) model (22). Denoted by $U_{t,j}$ the jth column of U_t, we let

$$U_{t,2} = U_{t+1,1}, \quad U_{t,3} = U_{t+2,1}, \quad U_{t,5} = U_{t+1,4}.$$

Hence, the first three columns of U_t forms one block of size 3, columns 4 and 5 forms a block of size 2, and the rest of the columns are uncorrelated with each other, and are uncorrelated to the first two blocks.

Table 2. Example 2—Relative frequencies correct and incorrect segmentation in a simulation with 1000 replications.

T	(q, p)	Correct	Merging	Splitting	Other	(q, p)	Correct	Merging	Splitting	Other
100	(6, 3)	0.369	0.274	0.299	0.058	(6, 6)	0.371	0.207	0.208	0.214
500		0.659	0.237	0.056	0.048		0.639	0.218	0.071	0.072
1000		0.721	0.220	0.031	0.028		0.706	0.219	0.030	0.045
5000		0.837	0.142	0.008	0.013		0.815	0.156	0.011	0.018
100	(10, 6)	0.103	0.087	0.270	0.540	(10, 10)	0.116	0.091	0.249	0.544
500		0.323	0.148	0.141	0.388		0.299	0.195	0.116	0.390
1000		0.369	0.227	0.086	0.318		0.363	0.218	0.067	0.352
5000		0.542	0.248	0.014	0.196		0.495	0.269	0.019	0.217

We set $T = 100, 500, 1000, 5000$. For each sample size, we set $(p, q) = (3, 6)$, $(6, 6)$, $(6, 10)$ and $(10, 10)$. Hence, the number of groups is $n_c = 3, 3, 7$, and 7, respectively. For each setting, we perform the eigenanalysis for $\widehat{W}^{(1)}$ defined as (7). We then apply the algorithm in Section 2.3 to arrange the orders of the q resulting orthonormal eigenvectors to form estimator $\widehat{A}_*$, for which the number of the correlated pairs of columns is determined by (11). Table 2 reports the relative frequencies of the correct specification of all the n_c uncorrelated blocks. We also report the relative frequencies of two types of wrong segmentation: (a) Merging with $\widehat{n}_c = n_c - 1$, that is, $n_c - 2$ blocks are correctly specified, and the remaining two blocks are put into one; (b) Splitting with $\widehat{n}_c = n_c + 1$, that is, $n_c - 1$ blocks are correctly specified, and the remaining block incorrectly splits into two. Table 2 indicates that the relative frequency of the correct specification increases as the sample size T increases, and it decreases as q increases while the performance is much less sensitive to the increase of p. It is noticeable that the probability of the event $\{\widehat{n}_c = n_c + 1\}$ is very small especially when T is large. With $q = 6$, the probability for the correct specification is not smaller than 64% with $T = 500$, and is greater than 70% with $T = 1000$. Furthermore the probability for $\widehat{n}_c = n_c$ or $n_c - 1$ is over 92% for $T \geq 500$. Note that when $\widehat{n}_c = n_c - 1$, an effective dimension reduction is still achieved in spite of missing a split.

For the instances with the correct specification, we also calculate the estimation errors for the n_c subspaces. More precisely, put $A_* = (A_1, \ldots, A_{n_c})$ and $\widehat{A}_* = (\widehat{A}_1, \ldots, \widehat{A}_{n_c})$. We measure the estimation error by

$$D_1(\widehat{A}_*, A_*) = \frac{1}{n_c} \sum_{j=1}^{n_c} \Big\{1 - \frac{1}{\text{rank}(A_j)} \text{trace}\big(A_j A_j^\top \widehat{A}_j \widehat{A}_j^\top\big)\Big\}.$$

Note that $\{\text{rank}(A_j)\}^{-1}\text{trace}\big(A_j A_j^\top \widehat{A}_j \widehat{A}_j^\top\big)$ is always between 0 and 1. Furthermore it is equal to 1 if $\mathcal{M}(A_j) = \mathcal{M}(\widehat{A}_j)$, and 0 if the two spaces are perpendicular with each other; see, for example, Stewart and Sun (1990) and Pan and Yao (2008). The boxplots of $D_1(\widehat{A}_*, A_*)$ presented in Figure 1 indicate that the estimation errors decrease when T increases, and the errors with large q are greater than those with small q.

Example 3. In this example we examine the gain in post-sample forecasting due to the decorrelation transformation. Now in model (1) all the elements of A and B are drawn independently from $U(-1, 1)$, with (p, q) equal to (6, 6), (6, 8), (8, 8), and (10, 10). For each (p, q) combination, the first three rows (or columns) form one block, the next two rows (or columns) form a second block, and all the other row (or column) is independent of the rest of the rows (or columns). Table 3 shows the configuration of U_t for $(p, q) = (8, 8)$.

Each univariate block of U_t (e.g., $U_{t,6,6}$) is an AR(1) process with the autoregressive coefficient drawn from the $U[0.1, 0.3]$ and independent and $N(0, 1)$ innovations. For each block of size (3, 1), (1, 3), (2, 1), and (1, 2), a vector AR(1) model is used. All elements of the coefficient matrix is drawn independently from $U[0, 1]$, then normalized so its singular values are within $[0.1, 0.3]$. The vector innovations consist of independent and N(0, 1) random variables. Each of the remaining four blocks follow a matrix AR(1) (i.e., MAR(1)) model of Chen, Xiao, and Yang (2021), that is,

$$U_{t,i,j} = \lambda_{i,j} \Phi_{1,i,j} U_{t-1,i,j} \Phi_{2,i,j}^\top + E_{t,i,j}, \quad i, j = 1, 2, \tag{23}$$

where $\Phi_{1,i,j}$ and $\Phi_{2,i,j}$ are the coefficient matrices with unit spectral norm, all elements of $E_{t,i,j}$ are independent and $N(0, 1)$. Matrices $\Phi_{1,i,j}$ and $\Phi_{2,i,j}$ are constructed such that their singular values are all one and the left and right singular matrices are orthonormal, and the scalar coefficient $\lambda_{i,j}$ controls the level of auto-correlation and its values for the four blocks are marked in Table 3.

We set sample sizes $T = 500, 1000, 2000, 5000$, and $(p, q) = (6, 6), (8, 6), (8, 8), (10, 10)$. For each setting, we simulate a matrix time series of length $T + M$ with $M = 10$. To perform one-step ahead post-sample forecasting, for each $i = 0, 1, \ldots, M-1$, we use the first T observations for identifying the uncorrelated blocks and for computing $\widehat{A}_*, \widehat{B}_*$ and $\widehat{U}_t^*$ defined in (8). Depending on its shape/size, we fit each identified block in $\widehat{U}_t^*$ with an MAR(1), VAR(1), or AR(1) model. The fitted models are used to predict U_{T+i+1}^*. The predicted value for X_{T+i+1} is then obtained by

$$\widehat{X}_{T+i+1} = \big(\widehat{\Sigma}_2^{(x)}\big)^{1/2} \widehat{B}_* \widehat{U}_{T+i+1}^* \widehat{A}_*^\top \big(\widehat{\Sigma}_1^{(x)}\big)^{1/2}, \quad i = 0, 1, \ldots, M-1. \tag{24}$$

We then compute the mean squared error:

$$\text{MSE} = \frac{1}{pq} \sum_{i=1}^{p} \sum_{j=1}^{q} (\widehat{X}_{s,i,j} - \mu_{s,i,j})^2, \tag{25}$$

where $\widehat{X}_{s,i,j}$ denote the one-step ahead predicted value for the (i, j)th element $X_{s,i,j}$ of X_s, and $\mu_{s,i,j} = \mathbb{E}(X_{s,i,j} | X_{s-1}, X_{s-2}, \ldots)$. We compare $\widehat{X}_{s,i,j}$ with $\mu_{s,i,j}$, instead of $X_{s,i,j}$, to remove the impact of the noise in the observed $X_{s,i,j}$. We report overall the mean and the standard deviation of MSE over 1000 replications.

For the comparison purpose, we also include several other models in our simulation: (i) VAR(1) model, that is, X_t is stacked

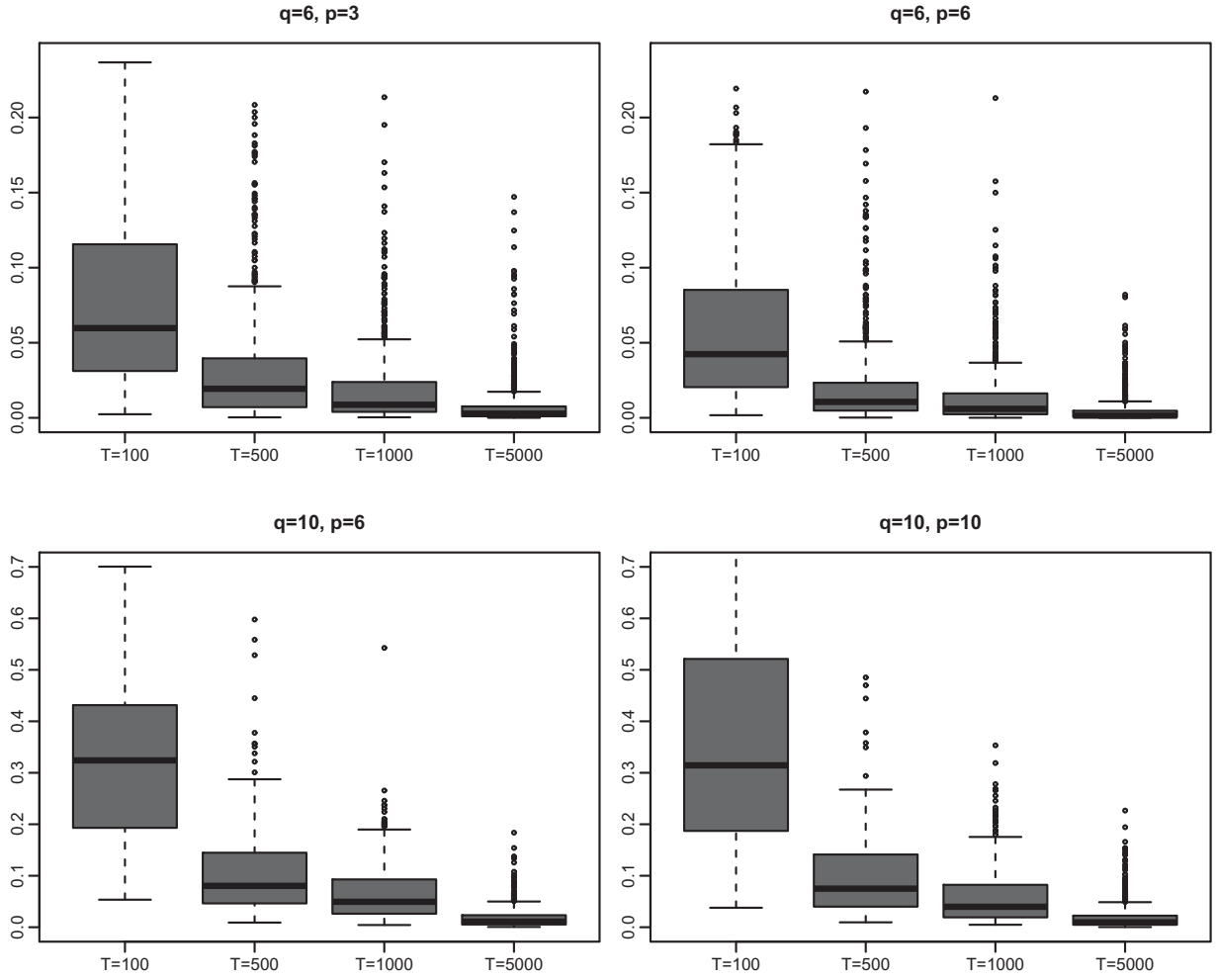

Figure 1. Example 2—boxplots of $D_1(\widehat{A}_*, A_*)$ in a simulation with 1000 replications.

Table 3. Example 3: Block configuration of U_t for the case of $(p, q) = (8, 8)$.

	1	2	3	4	5	6	7	8
1								
2		0.81			0.64			
3								
4								
5		0.25			0.25			
6								
7								
8								

into a vector of length pq and a VAR(1) is fitted to the vector series directly, (ii) MAR(1) model, that is, MAR(1) of Chen, Xiao, and Yang (2021), is fitted directly to X_t, and (iii) TS-PCA, that is, all the elements of X_t are stacked into a long vector and the segmentation method of Chang, Guo, and Yao (2018) is applied to the vector series. In addition, we also include two oracle models: (O1) the true segmentation blocks are used to model $\widehat{U}_t^*$; and (O2) True $\Sigma_1^{(x)}, \Sigma_2^{(x)}$ and true A_*, B_* are used and the true segmentation blocks are used to model $\widehat{U}_t^*$.

The mean and standard deviations (in bracket) of the MSEs are listed in Table 4. It is clear that the forecasting based on the proposed decorrelation transformation is more accurate than those based on MAR(1), VAR(1), and TS-PCA directly. The improvement is often substantial. The only exception is the case of $(p, q) = (6, 6)$ and $T = 5000$: the 36-dimension VAR(1) performs marginally better (i.e., the decrease of 0.002 in MSE) than the transformation based method. On the other hand, the improvement of the transformation based forecast over VAR(1) for sample size $T \leq 2000$ is more significant. The relative poor performance of MAR(1) is due to the substantial discrepancy between MAR(1) model and the data generating mechanism used in this example. Note that there are $(p+q)^2$ free parameters in VAR(1) while there are merely $(p^2 + q^2)$ parameters in MAR(1). With large T, VAR(1) is more flexible than MAR(1). TS-PCA performs poorly, due to the large error in estimating the large transformation matrix of size $pq \times pq$. See also Remark 6. In addition, TS-PCA also requires to fit VAR models for several large blocks of size 9, 6, 6 and etc.

Table 4 indicates that the oracle model O1 performs only slightly better than segmentation when sample size $T = 500$ or 1000. Oracle Model O2 only deviates from the true model in terms of the estimation error of the coefficients in the AR models, hence, performs the best. The MSE of O2 is very close to 0, since the only source of error in MSE is the estimation error of the AR models for each blocks of U_t, which goes to zero quickly as T increases since these models are of small dimensions. When T is large, the VAR(1), proposed segmentation and O1 all perform about the same, which is worse than O2.

Table 4. Example 3—One-step ahead post-sample forecasting: means and standard deviations (in bracket) of MSEs in a simulation with 1000 replications.

		MSEs and standard deviations			
		$T = 500$	$T = 1000$	$T = 2000$	$T = 5000$
$(p, q) = (6, 6)$	MAR(1)	0.515(0.333)	0.501(0.304)	0.513(0.305)	0.482(0.280)
	VAR(1)	0.369(0.142)	0.231(0.065)	0.137(0.037)	0.044(0.014)
	TS-PCA	0.492(0.232)	0.278(0.094)	0.149(0.043)	0.048(0.018)
	segmentation	0.257(0.180)	0.144(0.098)	0.087(0.094)	0.046(0.051)
	model O1	0.207(0.173)	0.098(0.079)	0.046(0.033)	0.018(0.012)
	model O2	0.018(0.010)	0.009(0.005)	0.004(0.003)	0.002(0.001)
$(p, q) = (6, 8)$	MAR(1)	0.981(0.496)	0.977(0.491)	0.947(0.488)	0.942(0.485)
	VAR(1)	0.549(0.235)	0.323(0.111)	0.179(0.058)	0.095(0.022)
	TS-PCA	0.787(0.359)	0.442(0.168)	0.197(0.068)	0.096(0.023)
	segmentation	0.329(0.262)	0.166(0.125)	0.097(0.073)	0.061(0.049)
	model O1	0.233(0.120)	0.127(0.064)	0.078(0.041)	0.051(0.035)
	model O2	0.032(0.015)	0.016(0.007)	0.008(0.004)	0.003(0.001)
$(p, q) = (8, 8)$	MAR(1)	0.971(0.434)	0.924(0.363)	0.915(0.391)	0.908(0.384)
	VAR(1)	0.861(0.304)	0.447(0.148)	0.275(0.079)	0.121(0.033)
	TS-PCA	0.965(0.420)	0.564(0.215)	0.290(0.106)	0.129(0.036)
	segmentation	0.499(0.276)	0.271(0.158)	0.176(0.131)	0.091(0.092)
	model O1	0.399(0.189)	0.252(0.154)	0.193(0.137)	0.120(0.133)
	model O2	0.040(0.015)	0.019(0.007)	0.010(0.004)	0.004(0.001)
$(p, q) = (10, 10)$	MAR(1)	1.244(0.512)	1.214(0.521)	1.207(0.481)	1.218(0.509)
	VAR(1)	1.920(0.654)	1.151(0.353)	0.646(0.191)	0.354(0.078)
	TS-PCA	1.822(0.711)	1.216(0.413)	0.726(0.211)	0.365(0.128)
	segmentation	0.967(0.505)	0.664(0.326)	0.465(0.240)	0.312(0.224)
	model O1	0.852(0.403)	0.617(0.307)	0.446(0.235)	0.341(0.233)
	model O2	0.062(0.022)	0.030(0.010)	0.016(0.005)	0.006(0.002)

Table 5. Sensitivity of the Kronecker product structure using different N and c.

		MSEs and standard deviations				
		$c = 0$	$c = 0.05$	$c = 0.15$	$c = 0.2$	$c = 0.3$
$N = 10$	VAR(1)	0.861(0.304)	0.877(0.327)	0.872(0.314)	0.874(0.291)	0.869(0.306)
	segmentation	0.499(0.276)	0.490(0.241)	0.620(0.303)	0.732(0.399)	0.859(0.504)
$N = 100$	VAR(1)	0.861(0.304)	0.856(0.295)	0.860(0.303)	0.874(0.326)	0.878(0.317)
	segmentation	0.499(0.276)	0.570(0.314)	0.721(0.458)	0.856(0.530)	0.976(0.689)

NOTE: The results are the means and standard deviations (in bracket) of MSEs of one-step ahead post-sample forecasting based on 1000 replications.

Example 4. Next, we assess the sensitivity of the proposed matrix segmentation method with respective to the Kronecker product structure ($\text{vec}(X_t) = (A \otimes B)\text{vec}(U_t)$). We use the same model setting in Example 3 and choose $(p, q) = (8, 8), T = 500$. Instead of model (1), we consider $\text{vec}(X_t) = (A \otimes B + \Psi)\text{vec}(U_t)$, where Ψ is a perturbation matrix with N none-zero elements. We also set each nonzero element of Ψ to be $c\|A \otimes B\|_F/(pq) \times \omega$, $\omega \overset{\text{iid}}{\sim}$ Rademacher distribution. Table 5 shows MSEs and standard deviations of VAR(1) and the proposed matrix segmentation method over 1000 replications, for several combinations of N and c. Again, N controls the number of nonzero perturbations and c controls the level of the perturbations, comparing to the average size of the elements in $A \otimes B$. When $c = 0$, it becomes the original case in Table 4. From Table 5, it is seen that the segmentation is still useful in prediction when the model is deviated from the assumed block structure within certain perturbation. The prediction performance of the VAR model is not affected by the perturbation since it does not assume any block structure. It is seen that the segmentation outperforms the VAR model when $c < 0.3$ when $N = 10$ and $c < 0.2$ when $N = 100$. This feature demonstrates the benefit of dimension reduction through segmentation, even when the underlying model does not have the exact Kronecker product structure.

4.2. Real Data Analysis

We now illustrate the proposed decorrelation method with a real data example. The data concerned is a 17×6 matrix time series consisting of the logarithmic daily averages of the six air pollutant concentration readings (i.e., $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, O_3) from the 17 monitoring stations in Beijing and its surrounding areas (i.e., Tianjin and Hebei Province) in China. The sampling period is January 1, 2015–December 31, 2016 for a total of $T = 731$ days. The readings of different pollutants at different locations are crossly correlated.

We first remove the seasonal mean of the 17×6 matrix time series. Setting $\tau_0 = 5$ in (7), we apply the proposed bilinear transformation to discover the uncorrelated block structure. The finding is stable: with $5 \le \tau_1 \le 30$ in (10), the transformed matrix series admits $\widehat{n}_c = 4$ uncorrelated column blocks with sizes 3, 1, 1, 1, respectively, and $\widehat{n}_r = 15$ uncorrelated row blocks with two blocks of size 2 and the rest of size 1. Overall there are 15×4 blocks, among which there are 2 blocks of size 2×3, 13 blocks of size 1×3, 6 blocks of size 2×1 and 39 blocks of size 1×1.

To check the post-sample forecasting performance, we calculate the rolling one-step, and two-step ahead out-sample forecasts for each of the daily readings in the last three months in

Table 6. One-step and two-step ahead post-sample forecasting for the air pollution data: means and standard deviations (in bracket) of MSPE_s, regrets and adjusted ratios of the various methods.

Method	One-step forecast			Two-step forecast		
	Raw forecast	Regret	Adjust ratio	Raw forecast	Regret	Adjust ratio
MAR(1)	0.351 (0.208)	0.154	1.305	0.516 (0.303)	0.319	1.340
VAR(1)	0.400 (0.272)	0.203	1.720	0.717 (0.406)	0.520	2.185
Univariate AR	0.346 (0.205)	0.149	1.263	0.501 (0.286)	0.304	1.277
TS-PCA	0.332 (0.188)	0.135	1.144	0.450 (0.247)	0.253	1.063
Segmentation with 15×4 blocks	0.315 (0.182)	0.118	1.000	0.435 (0.246)	0.238	1.000
Segmentation with 14×3 blocks	0.319 (0.185)	0.122	1.034	0.446 (0.256)	0.249	1.046
Segmentation with 16×5 blocks	0.318 (0.183)	0.121	1.025	0.443 (0.247)	0.246	1.034

2016 (total 92 days). The methods included in the comparison are (i) the forecasting based on the segmentation derived from the proposed decorrelation transformation by fitting an MAR(1), VAR(1) or AR(1) to each identified block according to its size and shape; (ii) MAR(1) for the original 17×6 matrix series; (iii) VAR(1) for the vectorized original series, (iv) univariate AR(1) for each of the 17×6 original series, and (v) the TS-PCA (Chang, Guo, and Yao 2018) for the vectorized original series. The TS-PCA leads to the segmentation consisting of one group of size 3, three groups of size 2, and 93 groups of size 1. For each group, a VAR(1) model is used for prediction.

For the two segmentation approaches, we fix the needed transformation obtained using data up to September 30, 2016, and the corresponding segmentation structure. The time series model for each individual block is updated throughout the rolling forecasting period.

In addition to the identified segmentation with 15×4 blocks, we also compute the forecasts based on two other segmentations: one with 14×3 blocks, and another with 16×5 blocks. The former is obtained by merging two most correlated single row blocks in the discovered segmentation into one block and two most correlated single column blocks into one block. The latter is obtained by splitting one of the two blocks with two rows in the discovered segmentation into two single row blocks and splitting the block with three columns into two blocks. This will reveal the impact of slightly "wrong" segmentations on forecasting performance.

Let the mean squared forecast error for sth observation be

$$\text{MSPE}_s = \frac{1}{pq}\sum_{i=1}^{p}\sum_{j=1}^{q}(\widehat{X}_{s,i,j} - X_{s,i,j})^2, \tag{26}$$

where $\widehat{X}_{s,i,j}$ denote the one-step ahead predicted value for the (i, j)th element $X_{s,i,j}$ of X_s. Two-step ahead MSPE_s is defined similarly. The means and the standard deviations of MSPE_s of one-step and two-step ahead post-sample forecasts for the pollution readings in the last 92 days are listed in Table 6. The prediction based on the bilinear decorrelation transformation (even using "wrong" segmentations) is clearly more accurate than those without transformation as well as that of the TS-PCA. Using the "wrong" segmentation deteriorates the performance only slightly, indicating that a partial (instead of total) decorrelation still leads to significant gain in prediction. Applying TS-PCA to the vectorized series requires to estimate a 102×102 transformation matrix (instead of the 17×17 and 6×6 matrices by using the matrix time series structure). It leads to

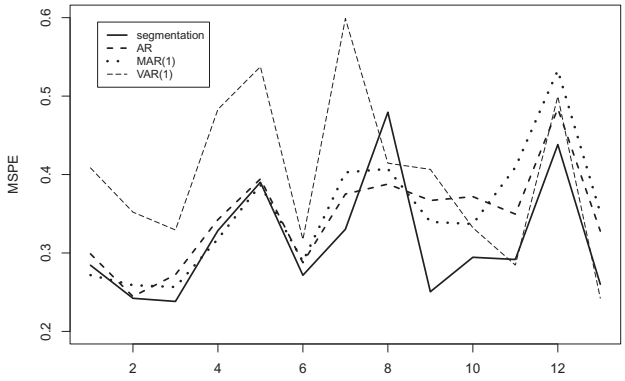

Figure 2. Weakly averaged one-step forecast errors of the univariate AR(1) models for each component series (AR), the matrix AR(1) model (MAR(1)), the vector AR(1) for vectorized series (VAR(1)), the method based on the bilinear decorrelation transformation (segmentation).

larger estimation errors (see Remark 6 in Section 3). Nevertheless it still provides more accurate predicts than those without transformation. Also note that fitting each of the original 17×6 time series with a univariate AR(1) separately leads to a better performance than those from fitting MAR(1) and VAR(1) to the original matrix series jointly. This indicates that the cross-serial correlation is useful and important information for future forecasting. However, to make efficient use of the information, it is necessary to adopt some effective independent component analysis or dimension-reduction techniques such as the proposed decorrelation transformation which pushes correlations across different series into the autocorrelations of some transformed series.

Also included in the table are "regret" defined as the difference between MSPE and the in-sample residual variance of the fitted VAR(1) model for the vectorized data (i.e., a vector time series of dimension $17 \times 6 = 102$), and "adjusted ratio" defined as the ratio of the regret to the regret of the best prediction model (i.e., the segmentation with 15×4 blocks). The regret tries to measure the forecasting error on the predictable signal, removing the impact from unpredictable noise in the model. Since the fitted VAR(1) model uses more than 10,000 parameters, its sample residual variance underestimates the noise variance in the model, and is taken as a proxy for the latter. The adjusted ratios shows that the "regret" of TS-PCA is 14.4% worse than the 15×4 segmentation for one-step ahead forecast, and 6.4% worse than the 15×4 segmentation for two-step ahead forecast.

To further evaluate the forecasting stability, Figure 2 shows weekly average of one-step rolling forecast MSPE within the forecasting period under various models. It is seen that the

proposed segmentation outperforms the other three methods in most of the periods in terms of prediction.

In this example, the identified segmentation of $(\widehat{n}_r, \widehat{n}_c) = (15, 4)$ for the original 17×6 matrix series is more likely to be an approximation of the underlying dependence structure rather than a true model. The fact that the two "wrong" segmentation models (though quite close to the discovered one) as well as the aggressive segmentation via vectorized TS-PCA transformation provide comparable, though inferior, post-sample forecasting performance lends further support to the claim that the proposed decorrelation method makes the transformed series more predictable than the original ones, regardless model (1) holds or not. See Remark 1(iv) in Section 2.1.

Supplementary Materials

The supplemental material contain additional simulation results, additional information on the air pollutant example and all proofs.

Acknowledgements

The authors sincerely thank the joint editor, associate editor, and referees for their valuable comments that helped improve the article substantially.

Funding

Han was supported in part by National Science Foundation grant IIS-1741390. Chen was supported in part by National Science Foundation grants DMS-1503409, DMS-1737857, and IIS-1741390. Zhang was supported in part by NSF grants DMS-2052949, DMS-2210850, IIS-1741390 and CCF-1934924. Yao was supported in part by the U.K. Engineering and Physical Sciences Research Council grant EP/V007556/1, EP/X002195/1.

ORCID

Yuefeng Han https://orcid.org/0000-0003-2311-3035
Rong Chen https://orcid.org/0000-0003-0793-4546
Qiwei Yao https://orcid.org/0000-0003-2065-8486

References

Ahn, S. C., and Horenstein, A. R. (2013), "Eigenvalue Ratio Test for the Number of Factors," *Econometrica*, 81, 1203–1227. [5]

Back, A. D., and Weigend, A. S. (1997), "A First Application of Independent Component Analysis to Extracting Structure from Stock Returns," *International Journal of Neural Systems*, 8, 473–484. [1]

Bai, J., and Ng, S. (2002), "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221. [1]

Basu, S., Dunagan, J., Duh, K., and Muniswamy-Reddy, K.-K. (2012), "Blrd: Applying Bilinear Logistic Regression to Factored Diagnosis Problems," *ACM SIGOPS Operating Systems Review*, 45, 31–38. [3]

Basu, S., and Michailidis, G. (2015), "Regularized Estimation in Sparse High-Dimensional Time Series Models," *The Annals of Statistics*, 43, 1535–1567. [1]

Bickel, P. J., and Levina, E. (2008a), "Covariance Regularization by Thresholding," *The Annals of Statistics*, 36, 2577–2604. [5,6]

——— (2008b), "Regularized Estimation of Large Covariance Matrices," *The Annals of Statistics*, 36, 199–227. [5]

Box, G. E. P., and Jenkins, G. M. (1970), *Times Series Analysis. Forecasting and Control*, San Francisco, CA.-London-Amsterdam: Holden-Day. [5]

Bradley, R. C. (2005), "Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions," *Probability Surveys*, 2, 107–144. [5]

Chang, J., Guo, B., and Yao, Q. (2018), "Principal Component Analysis for Second-Order Stationary Vector Time Series," *The Annals of Statistics*, 46, 2094–2124. [1,2,3,4,5,6,9,11]

Chang, J., He, J., Yang, L., and Yao, Q. (2021), "Modelling Matrix Time Series via a Tensor CP-Decomposition," arXiv:2112.15423. [1]

Chen, E. Y., Tsay, R. S., and Chen, R. (2020), "Constrained Factor Models for High-Dimensional Matrix-Variate Time Series," *Journal of the American Statistical Association*, 115, 775–793. [1,3]

Chen, E. Y., Yun, X., Chen, R., and Yao, Q. (2020), "Modeling Multivariate Spatial-Temporal Data with Latent Low-Dimensional Dynamics," arXiv preprint arXiv:2002.01305. [1]

Chen, R., Xiao, H., and Yang, D. (2021), "Autoregressive Models for Matrix-Valued Time Series," *Journal of Econometrics*, 222, 539–560. [1,3,8,9]

Chen, R., Yang, D., and Zhang, C.-H. (2022), "Factor Models for High-Dimensional Tensor Time Series," *Journal of the American Statistical Association*, 117, 94–116. [1,3,4]

Fan, J., and Yao, Q. (2003), *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer Series in Statistics. New York: Springer. [5]

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2005), "The Generalized Dynamic Factor Model: One-Sided Estimation and Forecasting," *Journal of the American Statistical Association*, 100, 830–840. [1]

Gao, Z., Ma, Y., Wang, H., and Yao, Q. (2019), "Banded Spatio-Temporal Autoregressions," *Journal of Econometrics*, 208, 211–230. [1]

Ghosh, S., Khare, K., and Michailidis, G. (2019), "High-Dimensional Posterior Consistency in Bayesian Vector Autoregressive Models," *Journal of the American Statistical Association*, 114, 735–748. [1]

Guo, S., Wang, Y., and Yao, Q. (2016), "High-Dimensional and Banded Vector Autoregressions," *Biometrika*, 103, 889–903. [1]

Han, Y., Chen, L., and Wu, W. (2020), "Sparse Nonlinear Vector Autoregressive Models," Technical Report. [1]

Han, Y., Chen, R., Yang, D., and Zhang, C.-H. (2020), "Tensor Factor Model Estimation by Iterative Projection," arXiv preprint arXiv:2006.02611. [1]

Han, Y., Chen, R., and Zhang, C.-H. (2022), "Rank Determination in Tensor Factor Model," *Electronic Journal of Statistics*, 16, 1726–1803. [1,5]

Han, Y., and Tsay, R. S. (2020), "High-Dimensional Linear Regression for Dependent Observations with Application to Nowcasting," *Statistica Sinica*, 30, 1797–1827. [1]

Han, Y., Tsay, R. S., and Wu, W. B. (2023), "High Dimensional Generalized Linear Models for Temporal Dependent Data," *Bernoulli*, 29, 105–131. [1]

Han, Y., and Zhang, C.-H. (2022), "Tensor Principal Component Analysis in High Dimensional CP Models," *IEEE Transactions on Information Theory*, forthcoming. [1]

Han, Y., Zhang, C.-H., and Chen, R. (2021), "CP Factor Model for Dynamic Tensors," arXiv:2110.15517. [1]

Hoff, P. D. (2015), "Multilinear Tensor Regression for Longitudinal Relational Data," *Annals of Applied Statistics*, 9, 1169–1193. [1,3]

Huang, D., and Tsay, R. (2014), "A Refined Scalar Component Approach to Multivariate Time Series Modeling," Manuscript. [1]

Lam, C., and Yao, Q. (2012), "Factor Modeling for High-Dimensional Time Series: Inference for the Number of Factors," *The Annals of Statistics*, 40, 694–726. [1,5]

Lam, C., Yao, Q., and Bathia, N. (2011), "Estimation of Latent Factors for High-Dimensional Time Series," *Biometrika*, 98, 901–918. [4]

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, 4188–4236. [1]

Liu, W., Xiao, H., and Wu, W. B. (2013), "Probability and Moment Inequalities under Dependence," *Statistica Sinica*, 23, 1257–1272. [6]

Matteson, D. S., and Tsay, R. S. (2011), "Dynamic Orthogonal Components for Multivariate Time Series," *Journal of the American Statistical Association*, 106, 1450–1463. [1]

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, 435–474. [2]

Negahban, S., and Wainwright, M. J. (2011), "Estimation of (near) Low-Rank Matrices with Noise and High-Dimensional Scaling," *The Annals of Statistics*, 39, 1069–1097. [1]

Pan, J., and Yao, Q. (2008), "Modelling Multiple Time Series via Common Factors," *Biometrika*, 95, 365–379. [8]

Pena, D., and Box, G. E. (1987), "Identifying a Simplifying Structure in Time Series," *Journal of the American statistical Association*, 82, 836–843. [1]

Rio, E. (2000), "*Théorie asymptotique des processus aléatoires faiblement dépendants*, volume 31 of Mathématiques & Applications (Berlin) [Mathematics & Applications]. Berlin: Springer. [2,6]

Rohde, A., and Tsybakov, A. B. (2011), "Estimation of High-Dimensional Low-Rank Matrices," *The Annals of Statistics*, 39, 887–930. [1]

Stewart, G. W., and Sun, J. G. (1990), *Matrix Perturbation Theory*. Computer Science and Scientific Computing. Boston, MA: Academic Press, Inc.. [8]

Tiao, G. C., and Tsay, R. S. (1989), "Model Specification in Multivariate Time Series," *Journal of the Royal Statistical Society*, Series B, 51, 157–195. [1]

Wang, D., Liu, X., and Chen, R. (2019), "Factor Models for Matrix-Valued High-Dimensional Time Series," *Journal of Econometrics*, 208, 231–248. [1,3,4]

Wang, L., Zhang, Z., and Dunson, D. (2019), "Symmetric Bilinear Regression for Signal Subgraph Estimation," *IEEE Transactions on Signal Processing*, 67, 1929–1940. [3]

Wu, W.-B., and Wu, Y. N. (2016), "Performance Bounds for Parameter Estimates of High-Dimensional Linear Models with Correlated Errors," *Electronic Journal of Statistics*, 10, 352–379. [6]

Xia, D., and Yuan, M. (2019), "Statistical Inferences of Linear Forms for Noisy Matrix Completion," arXiv preprint arXiv:1909.00116. [1]

Xia, Y., Cai, T., and Cai, T. T. (2018), "Multiple Testing of Submatrices of a Precision Matrix with Applications to Identification of between Pathway Interactions," *Journal of the American Statistical Association*, 113, 328–339. [5]

Xiao, H., Han, Y., and Chen, R. (2022), "Reduced Rank Autoregressive Models for Matrix Time Series," *Journal of Business & Economic Statistics*, forthcoming. [1]

Zhang, D., and Wu, W. B. (2017), "Gaussian Approximation for High Dimensional Time Series," *The Annals of Statistics*, 45, 1895–1919. [6]

Zhou, H., Li, L., and Zhu, H. (2013), "Tensor Regression with Applications in Neuroimaging Data Analysis," *Journal of the American Statistical Association*, 108, 540–552. [3]

Zhou, H. H., and Raskutti, G. (2018), "Non-parametric Sparse Additive Auto-Regressive Network Models," *IEEE Transactions on Information Theory*, 65, 1473–1492. [1]