Statistica Sinica Preprint No: SS-2023-0009	
Title	A Regularized Low Tubal-Rank Model for High-dimensional Time Series Data
Manuscript ID	SS-2023-0009
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0009
Complete List of Authors	Samrat Roy and George Michailidis
Corresponding Authors	Samrat Roy
E-mails	samratroy67@yahoo.com

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

A Regularized Low Tubal-Rank Model for High-dimensional Time Series Data

Samrat Roy* and George Michailidis

Indian Institute of Management Ahmedabad and University of Florida

Abstract:

High dimensional time series analysis has diverse applications in macroeconometrics and finance. Recent factor-type models employing tensor-based decompositions prove to be computationally involved due to the non-convex nature of the underlying optimization problem and also they do not capture the underlying temporal dependence of the latent factor structure. This work leverages the concept of tubal rank and develops a matrix-valued time series model, which first captures the temporal dependence in the data, and then the remainder signals across the time points are decomposed into two components: a low tubal rank tensor representing the baseline signals, and a sparse tensor capturing the additional idiosyncrasies in the signal. We address the issue of identifiability of various components in our model and subsequently develop a scalable Alternating Block Minimization algorithm to solve the convex regularized optimization problem for estimating the parameters. We provide finite sample error bounds under high dimensional scaling for the model parameters.

*Corresponding author. E-mail: samratroy67@yahoo.com

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

Key words and phrases: High Dimensional Time Series, Tensor, Tubal Rank.

1. Introduction

There has been a lot of interest in modeling high-dimensional time series due not only to traditional application areas in macroeconomics and finance (De Mol et al., 2008; Blanchard and Perotti, 2002; Bernanke et al., 2005), but also emerging ones, including dynamic traffic networks (Chen and Chen, 2022), functional genomics (Michailidis and d'Alché Buc, 2013) and neuroscience (Seth et al., 2015). To accommodate high dimensionality, both regularized versions of VAR models (Bańbura et al., 2010; Basu et al., 2015; Kock and Callot, 2015; Ghosh et al., 2019; Wang et al., 2024) and dynamic factor models (Bai and Wang (2015), Lam et al. (2012), Chang et al. (2018)) have been developed. Wang et al. (2019) and Chen et al. (2022) recently extended the aforementioned factor models to matrix and tensor valued time series by expressing the data into a low dimensional dynamic signal as $A_1 G_t A_2^T$, or $\mathcal{G}_t \times_1 A_1 \times_2 A_2 \times_3 \cdots \times_k A_k$ respectively, with A_i's being the loading matrices and G_t and $\mathcal{G}_t$ are the core factor matrix and tensors.

The two key points that we aim to address through our work are

(a) In both Wang et al. (2019) and Chen et al. (2022), the factor matrix

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

G_t and the factor tensor $\mathcal{G}_t$ are assumed to drive all the dynamics of the data. However, they capture the temporal dependence through lagged covariance matrix. On the other hand, as it will be discussed later, we aim to capture the same through our model in the form of a transition matrix B^*. For the ease of exposition, we only include lag-1 temporal dependence in our model. However, one can similarly add more lag-terms to capture higher order temporal dependence.

(b) In many applications, the data may have a more complex underlying low dimensional structure, which may not be justified by the aforementioned tucker decomposition-based factor-loading representation. For example, (**1**) matrix-valued product ratings data with the rows and the columns corresponding to the customer age groups and different product categories respectively. E-Commerce experts suggest (see Example 2 of Agarwal et al. (2012)) decomposing such underlying signals into two parts: the first part carrying similar baseline signals across different product categories, and the second part representing the additional signals that are active for only some specific combinations of the age groups and product categories (see Section 2 for more details). (**2**) *Macroeconomic matrix-valued quarterly data*, where at each time point a set of Macroeconomic indicators (GDP,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

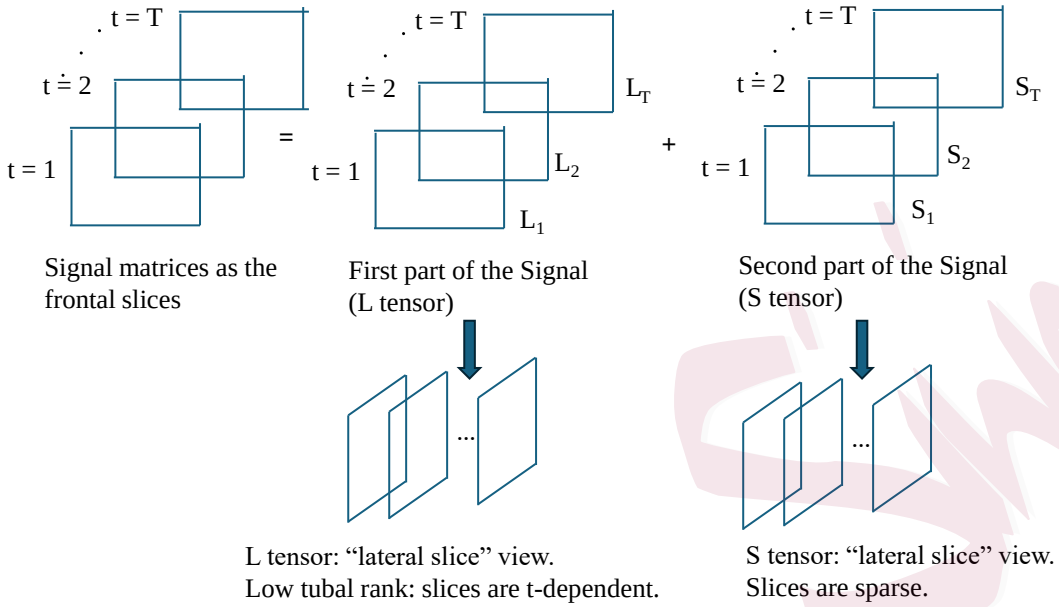

Figure 1: Low dimensional structure of Signal

Inflation index and so on) are available for different countries in European Union (see Section 5 for details). Since the countries under European Union follow harmonized economic policies with a unified objective, as in the previous example, it is meaningful to assume that one part of the signal will be similar or shared across the different countries. Whereas, the second component of the signal is devoted to capturing the additional idiosyncratic elements, that are specific to only some of the combinations of variables and countries, arising from an economic boom or financial stress.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

In this paper, we propose a matrix-valued time series model that aims to address the two key issues ((a) and (b)) discussed above. The model first captures the underlying temporal dependence through a sparse transition matrix B and then induces a low-dimensional structure in the remainder signal that is suitable for the previously mentioned examples in (b). To that end, instead of employing a tucker decomposition based factor-loading form, the signal is rather decomposed into two parts. When the first parts across different time points, denoted by L_t's, are arranged as the frontal slices along the time dimension (see Figure 1 for intuition), the resulting third-order tensor $\boldsymbol{\mathcal{L}}$ is assumed to exhibit *low tubal rank*, a concept introduced for tensor decompositions in Kilmer and Martin (2011) and Kilmer et al. (2013). As illustrated in Figure S8.10, for a third-order tensor, the horizontal slices, lateral slices and the tube fibers play the roles of rows, columns and elements of a matrix, respectively. The aforementioned papers introduced a novel concept of *t-product* [tensor counterpart of the product between a matrix column and a scalar; see Definition S4.1 in the supplement], *t-linear combination* [tensor counterpart of linear combination of the matrix columns; see Definition S4.7 in the supplement]. Finally, Kilmer and Martin (2011) define the tubal rank [tensor counterpart of the matrix rank, see Definition S4.9 in the supplement] and show that, just like the rank of a

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

matrix, the tubal rank of a third-order tensor is the number of lateral and horizontal slices that are *t-linearly independent* (see Figure S8.9, Section 2 and the discussion after Definition S4.7). Thus the low tubal rank tensor $\boldsymbol{\mathcal{L}}$ in Figure 1 captures the first part of the signal, where the similar baseline signals are shared across the lateral (or, horizontal) slices of $\boldsymbol{\mathcal{L}}$. Recalling the two examples discussed before, these slices will correspond to different product categories in the e-Commerce example, or different countries in the Macroeconomic example. In addition to this baseline signal, one can similarly define another third-order tensor $\boldsymbol{\mathcal{S}}$ (see Figure 1) that is sparse and captures the additional idiosyncrasies.

The key novel contribution of this paper is to develop the algorithm and technical tools to obtain estimates of the two components of the signal, along with the transition matrix that captures the temporal dependence and establish a non-asymptotic upper bound to the estimation error under both Gaussian and Sub-Exponential distributional assumption. This type of decomposition has been employed earlier in another line of research that deals with tensor recovery (see Liu et al. (2016)). However, to the best of our knowledge, this methodology and the subsequent theoretical analysis are new in the context of high-dimensional time series models.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

2. Low Tubal-Rank Time Series Model and Its Estimation

Suppose there are p variables of interest, for m entities, observed across T time periods, as discussed in the motivating examples. The data are then modeled as

$$X_t = M_t + E_t,\ t = 1, 2, \cdots, T, \tag{2.1}$$

where $X_t \in \mathbb{R}^{p\times m}$ is the matrix-valued time series observed at time point t and M_t and E_t are the signal and noise components associated with X_t. Further, we assume that $Cov\{Vec(M_t), Vec(E_{t'})\} = 0, \forall\ (t, t')$, that is, the error processes are uncorrelated with the signal across all time points, where the notation $Vec(\cdot)$ is used to denote the vectorized form of a matrix.

We introduce a sparse transition matrix $B^* \in \mathbb{R}^{pm\times pm}$ to capture the underlying temporal dependence in the dynamic signal and noise, by assuming $Vec(M_t) = B^*\ Vec(M_{t-1}) + Vec(F_t)$ and $Vec(E_t) = B^*\ Vec(E_{t-1}) + Vec(U_t)$, where both $Vec(F_t)$ and $Vec(U_t)$ are white noise processes with mean zero and $Cov\{Vec(F_t), Vec(U_{t'})\} = 0, \forall\ (t, t')$. The primary motivation of assuming sparsity on B^* is to deal with the excessive dimensionality of the same as compared to the available sample size. By assuming sparsity structure on B^*, our method selects and estimates only the significant parameters in B^*, and sets the insignificant ones to zero. This assump-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

tion has been widely used in the literature of high-dimensional time series models, including the SSVS approach of George et al. (2008), a Bayesian Non-parametric approach of Billio et al. (2019), and a Graphical VAR approach in Ahelegbey et al. (2016a,b). Basu et al. (2015) investigate the theoretical properties of ℓ_1-regularized estimate of the sparse VAR transition matrix. Asymptotic properties of lasso for high-dimensional time series have also been considered by Loh and Wainwright (2012), and Basu et al. (2015) provide detailed comparisons with the above studies. Ghosh et al. (2019) and Chakraborty et al. (2023) also assume sparsity structure on the VAR transition matrix, and formally extend the idea of Basu et al. (2015) in the Bayesian context and for the mixed frequency time series data respectively.

It is worth mentioning that, here we are assuming temporal dependence of the noise process, which is in line with the approach of Bai (2003) and Chen and Fan (2023). On the other hand, Wang et al. (2019) assume no temporal dependence of the noise process, and the factor process captures all the temporal dependence of the data. Liu and Zhang (2022) provide a detailed description on the differences between these two settings.

With the sparse transition matrix B^*, equation (2.1) translates into the following form,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

$$Vec(X_t) = Vec(F_t) + B^* \; Vec(X_{t-1}) + Vec(U_t), \; t = 1, 2, \cdots, T \qquad (2.2)$$

Once the temporal dependence is captured, we aim to employ the low-dimensional structure on the remainder signal F_t, that we discussed in Section 1 (see the examples discussed in (b) of Section 1). As argued there, one may come across time series data, for which the underlying low dimensional structure of the associated signal may not be fully justified by Tucker decomposition based factor-loading representation used in Wang et al. (2019) and Chen et al. (2022). Consider the earlier example from e-Commerce, where the matrix valued data on product ratings are available with customer age-group in one dimension and the product categories in the other dimension. As e-Commerce experts suggest (see Agarwal et al. (2012)), the signal corresponding to these ratings will be more or less similar or shared across the different product categories, say "Books", "Electronics" and "Clothing Accessories". However, in addition to this similar baseline part, there will be a few additional idiosyncratic elements in the signal, which are active for only some specific combinations of age-group and product categories. For example, the customers of age-group "16 years-30 years" are more inclined towards online shopping of clothing accessories and thus this combination demands additional signal component over and above the baseline signal. Hence, in the baseline part, there should be low-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

rankness along the product-categories and on the other hand, the second part should be a sparse component capturing the idiosyncrasies (see Figure S8.8 in the supplement). Similar justification follows for the other example from Macroeconomics too, with the baseline part capturing low-rankness across the countries and country-specific financial crisis giving rise to the additional idiosyncratic signal (see Figure S8.8 and Section 1).

To formulate this typical low dimensional structure, we first place the signal matrices F_t's as the frontal slices (see Figure 1), create a third-order tensor $\boldsymbol{\mathcal{F}} \in \mathbb{R}^{p\times m\times T}$ and assume that $\boldsymbol{\mathcal{F}} = \boldsymbol{\mathcal{L}}^* + \boldsymbol{\mathcal{S}}^*$, where $\boldsymbol{\mathcal{L}}^*$ and $\boldsymbol{\mathcal{S}}^*$ are two complementary types of low dimensional structure, as the aforementioned examples demand.

The $\boldsymbol{\mathcal{L}}^*$ component corresponds to a low tubal-rank tensor. As shown in Kilmer et al. (2013) and discussed in Section 1, the tubal-rank of a third-order tensor is analogous (in the appropriate algebra, see Section S4 in the supplement) to the rank of a matrix. Hence, analogously to the fact that low rankness of a matrix implies linear dependence among its columns and rows, a low value of tubal-rank characterizes a similar type of dependence, namely *t-linear* dependence (see Definition S4.7 and the ensuing discussion), among the lateral and horizontal slices. For the posited model, we select the dimension across the lateral slices to impose low-rankness. However, one can

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

always reorient the tensor and thus impose low-rankness assumption across any of the dimensions. The aforementioned dependence is governed by the concept of the t-product and the t-linear combination introduced in Kilmer et al. (2013). The formal definitions are deferred to the supplement and Figure S8.9 illustrates the key ideas. As it depicts, a lateral slice is said to be t-dependent on the other, if the former can be expressed as the t-product between the latter and a suitable tube. Using the definition of t-product, Figure S8.9 also provides an alternative representation of t-dependence in terms of Block-Circulant matrix [See notation S4.2 in the supplement]. Figure S8.1 in the supplement provides some numerical examples to show the relation between tubal-rank and the t-dependence among the lateral slices. Thus, based on this brief discussion, the purpose of the first component $\boldsymbol{\mathcal{L}}^*$ is to capture similar baseline part of the signal. The second component $\boldsymbol{\mathcal{S}}^*$ represents the additional idiosyncrasies and can be formally considered as a third order tensor, whose frontal slices are elementwise sparse.

Following the above discussion, we rewrite our model in equation (2.2) in more explicit form as follows:

$$Vec(X_t) = Vec(L_t^*) + Vec(S_t^*) + B^*\ Vec(X_{t-1}) + Vec(U_t),\ t = 1, 2, \cdots, T \tag{2.3}$$

where $B^* \in \mathbb{R}^{pm \times pm}$ is an elementwise sparse transition matrix. For $t =$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

$1, 2, \cdots, T$, L_t^* and S_t^* are the t^{th} frontal slices of the tensors $\boldsymbol{\mathcal{L}}^*$ and $\boldsymbol{\mathcal{S}}^*$ respectively, where $\boldsymbol{\mathcal{L}}^* \in \mathbb{R}^{p\times m\times T}$ is a low tubal-rank tensor and $\boldsymbol{\mathcal{S}}^* \in \mathbb{R}^{p\times m\times T}$ tensor whose frontal slices are elementwise sparse. $Vec(U_t)$ is a zero mean White Noise process. Later, while we have theoretical discussion in Section 3, we take different distributional assumptions on the processes $Vec(U_t)$, $Vec(L_t^*)$ and $Vec(S_t^*)$ and present our results under those distributional assumptions. Finally, a tensor counterpart of equation (2.3) can be written as:

$$\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{L}}^* + \boldsymbol{\mathcal{S}}^* + B^* \times \boldsymbol{\mathcal{Z}} + \boldsymbol{\mathcal{U}} \tag{2.4}$$

where, $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{p\times m\times T}$, $\boldsymbol{\mathcal{Z}} \in \mathbb{R}^{p\times m\times T}$ and $\boldsymbol{\mathcal{U}} \in \mathbb{R}^{p\times m\times T}$ are the third-order tensors whose frontal slices are $\{X_t\}_{t=1}^T$, $\{X_{t-1}\}_{t=1}^T$ and $\{U_t\}_{t=1}^T$ respectively, and the notation $B^* \times \boldsymbol{\mathcal{Z}}$ is used to denote a third-order tensor of size $p \times m \times T$, whose t^{th} frontal slice is the matricized version of the pm dimensional vector $B^* Vec(X_{t-1})$ in equation (2.3). Given the model in equation (2.4), our objective is to estimate the low tubal-rank signal component $\boldsymbol{\mathcal{L}}^*$, sparse signal component $\boldsymbol{\mathcal{S}}^*$ and sparse transition matrix B^* based on the available data. Note that, some additional constraints are required in order to make the model identifiable, which have been discussed in Section 2.1.

Note that, in the e-Commerce example (and similarly in the macroeconomics example), one can not simply use the Tucker decomposition in

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

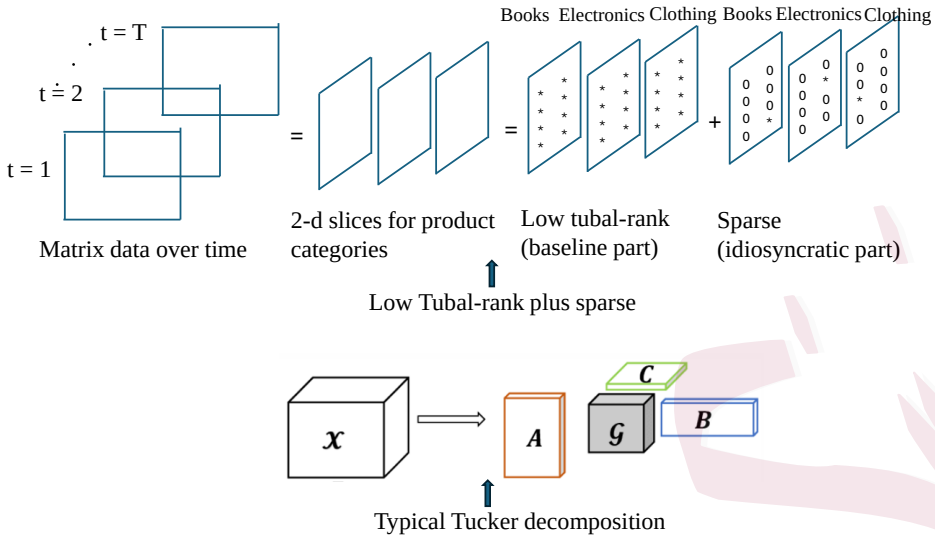

Figure 2: Pictorial illustration of the difference between Tucker decomposition and low Tubal-rank plus sparse approach

order to impose our desired low-dimensional structure. As explained in Kolda and Bader (2009), Chen et al. (2022), and depicted in Figure 2, a typical Tucker decomposition uses a core tensor, and then applies loading matrices along all three different directions of the core tensor. So, though it helps in reducing the dimension substantially, the reduction of dimension occurs arbitrarily from all three directions. However, in our examples, we have matrix-valued time series, with inherent low-dimensional pattern decomposed into two parts: baseline and additional idiosyncratic parts. In the baseline part, the signal will be similar or shared across the three product

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

categories: books, electronics and clothing accessories. Thus, to induce this shared pattern, it behooves us to impose low-rankness along the product category direction. To that end, as explained earlier and depicted in Figure 2 and Figure 1, we first create two-dimensional lateral slices corresponding to each product category (with time in one dimension and age-group in the other dimension), and assume low tubal-rank among these slices. As discussed earlier in this section, and formally presented in Section S4 of the supplementary material, just like low-rank and linear dependence between the matrix columns, low tubal-rank induces t-linear dependence (see Figure S8.9 and the discussion after Definition S4.7 in the supplementary material) among the above lateral slices, and that dependence will capture the shared baseline signal across the product categories. In addition to this baseline part, the idiosyncratic part is captured by the sparse tensor as shown in Figure 2. Hence, the low-dimensional structure induced by the Tucker decomposition is quite different than our low tubal-rank plus sparse approach, and the former is not suitable for our desired pattern.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

2.1 Estimation of the Signal components and the Transition Matrix

The tubal rank and sparsity constraints in equation (2.4) are non-convex and as common with other high-dimensional models, a convex relaxation is introduced to compute the model parameters. Next, we introduce some necessary notation employed in the sequel. Define matrix $Y \in \mathbb{R}^{T\times pm}$ as $Y = \{Vec(X_1), Vec(X_2), \cdots, Vec(X_T)\}^T$. Thus, the t^{th} row of Y corresponds to the t^{th} time point, wherein the observed matrix X_t is arranged in vectorized form. Similarly, define $T \times pm$ dimensional matrices L^*, S^*, Z and U associated with $Vec(L_t^*), Vec(S_t^*), Vec(X_{t-1})$ and $Vec(U_t)$ respectively. Thus, the model in equation (2.4) (and in equation (2.3)) can be alternatively expressed in the following form:

$$Y = L^* + S^* + ZB^{*T} + U \tag{2.5}$$

It can be seen from equation (S4.1) and the related discussion in the supplement that $\frac{1}{T}\left\|Circ(\boldsymbol{\mathcal{L}}^*)\right\|_*$ corresponds to a suitable convex relaxation for the low tubal-rank constraint, where $Circ(\boldsymbol{\mathcal{L}}^*)$ is the *block-circulant matrix* associated with the tensor $\boldsymbol{\mathcal{L}}^*$ (see notation S4.2). Indeed, it is an alternative form of the Tensor Nuclear Norm defined in Lu et al. (2018) (see Definition 7 in Lu et al. (2018) and equation 12 in Lu et al. (2016)) and

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

imposing a constraint on this norm translates into an analogous constraint on the tubal-rank. Further, since $\boldsymbol{\mathcal{S}}^*$ consists of element-wise sparse frontal slices, one can simply treat each frontal slice as a matrix and employ the usual element-wise ℓ_1 norm for each of them. Hence, we consider a regularizer $\sum_{t=1}^{T} \|S_t^*\|_1$, or equivalently $\|S^*\|_1$, for the sparse component. Similarly, we employ an element-wise ℓ_1 norm for the sparse transition matrix B^*.

The objective function for estimating the parameters $\boldsymbol{\mathcal{L}}^*$ and $\boldsymbol{\mathcal{S}}^*$ and B^* becomes:

$$\min_{L,S,B} \{ \frac{1}{2T} \left\| Y - L - S - ZB^T \right\|_F^2 + \lambda_L \frac{1}{T} \|Circ(\boldsymbol{\mathcal{L}})\|_* + \lambda_S \|S\|_1 + \lambda_B \|B\|_1 \} \tag{2.6}$$

wherein λ_L, λ_S and λ_B are non-negative regularization parameters corresponding to the low tubal-rank signal component, sparse signal component and sparse transition matrix, respectively.

Remark 2.1. *While penalizing $Circ(\boldsymbol{\mathcal{L}}^*)$, the factor $\frac{1}{T}$ can be interpreted as follows: for any third-order tensor in $\mathbb{R}^{p \times m \times T}$, the associated block-circulant matrix consists of m blocks. Each block corresponds to a particular lateral slice and contains all of the T possible block-circulant arrangements of that lateral slice. Hence, in addition to exhibiting inter-slice t-dependence, the penalty on $\|Circ(\boldsymbol{\mathcal{L}})\|_*$ also induces intra-slice dependence among the T block-circulant arrangements of a slice. The factor $\frac{1}{T}$ thus adjusts for this*

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

additional penalization.

Before presenting an algorithm for estimating the model parameters $(\boldsymbol{\mathcal{L}}^*, \boldsymbol{\mathcal{S}}^*, B^*)$, we address the issue of their identifiability next. There are two sources of non-identifiability in our posited model. First, in order to correctly identify $\boldsymbol{\mathcal{L}}^*$ and $\boldsymbol{\mathcal{S}}^*$, one should be able to disentangle the two components from each other. In other words, the low tubal-rank component $\boldsymbol{\mathcal{L}}^*$ should be "*incoherent*" with the sparse component $\boldsymbol{\mathcal{S}}^*$, so that one can identify and recover the two components $\boldsymbol{\mathcal{L}}^*$ and $\boldsymbol{\mathcal{S}}^*$ separately. In addition to that, the overall signal component $F = L^* + S^*$ in equation (2.5), should also be incoherent with the third component ZB^{*T}. In the case of a multivariate response regression model, an incoherence condition between the low rank and the sparse component of the regression coefficient matrix, is usually operationalized through conditions on the singular vectors of the low rank component obtained from the singular value decomposition (SVD); for details, see, Candès et al. (2011). In our case, we adapt the approach employed in Agarwal et al. (2012) to the current setting and require

$$\|Circ(\boldsymbol{\mathcal{L}}^*)\|_\infty \leq \frac{\alpha_1}{\sqrt{pmT}} \text{ and } \|F\|_\infty \leq \frac{\alpha_2}{\sqrt{pmT}\,\|Z\|_{sp}}$$

wherein $\alpha_1 > 0$ and $\alpha_2 > 0$ are some fixed parameters. Note that based on Proposition S3.1 in the supplement, a low tubal rank for $\boldsymbol{\mathcal{L}}^*$ translates

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

to a small matrix rank for $Circ(\boldsymbol{\mathcal{L}}^*)$ and vice versa. Hence, the nature of the posited incoherence constraint follows from that for the matrix case. Specifically, by restricting the "spikiness" -a matrix with high "spikiness" will have very few non-zero elements and all the remaining elements will be zeros- of the elements of $Circ(\boldsymbol{\mathcal{L}}^*)$, one can ensure a sufficient number of non-zero elements in $Circ(\boldsymbol{\mathcal{L}}^*)$ and thus the same to hold for each of the frontal slices of $\boldsymbol{\mathcal{L}}^*$. However, each of the last $m(T-1)$ columns of $Circ(\boldsymbol{\mathcal{L}}^*)$ can be written by rearranging elements of any one of its first m columns. Hence, by restricting the "spikiness" of pmT elements of only the first m columns, one can essentially control the 'spikiness' of all the elements of $Circ(\boldsymbol{\mathcal{L}}^*)$, which leads to the first incoherence condition posited above. Similarly, the second incoherence condition controls the spikiness of the overall signal component F with respect to the sparse transition matrix B^*, where the upper bound of the spikiness is now suitably scaled by $\|Z\|_{sp}$.

Note that the objective function (2.6), denoted by $f(L, S, B)$, is *jointly convex* in its arguments and hence the following alternating block minimization procedure summarized in Algorithm 1, will obtain the desired minimizer. The details of Steps $1-3$, that update L, S and B, respectively, are presented in Section S5 of the supplementary materials.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

Algorithm 1 Alternating Block Minimization for minimizing $f(L, S, B)$

Input: data $Y \in \mathbb{R}^{T \times pm}$, $\lambda_L, \lambda_S, \lambda_B$
Initialize: $L^{(0)}, S^{(0)} \in \mathbb{R}^{T \times pm}$ and $B^{(0)} \in \mathbb{R}^{pm \times pm}$
repeat
    Step 1: Update $L^{(t+1)} = \arg\min_{L} f(L, S^{(t)}, B^{(t)})$, given $S^{(t)}$ and $B^{(t)}$
    Step 2: Update $S^{(t+1)} = \arg\min_{S} f(L^{(t+1)}, S, B^{(t)})$, given $L^{(t+1)}$ and $B^{(t)}$
    Step 3: Update $B^{(t+1)} = \arg\min_{B} f(L^{(t+1)}, S^{(t+1)}, B)$, given $L^{(t+1)}$ and $S^{(t+1)}$
until $f(L^{(t+1)}, S^{(t+1)}, B^{(t+1)})$ converges

3. Theoretical Results

We start by defining the estimation error $e^2(\hat{L}, \hat{S}, \hat{B})$ as follows:

$$e^2(\hat{L}, \hat{S}, \hat{B}) = \frac{1}{T}\left\|\hat{L} - L^*\right\|_F^2 + \frac{1}{T}\left\|\hat{S} - S^*\right\|_F^2 + \left\|\hat{B} - B^*\right\|_F^2 \tag{3.1}$$

The factor $\frac{1}{T}$ in the first two parts of the above definition can be interpreted as follows: from equation (2.5) it can be seen that our posited model includes $L^* \in \mathbb{R}^{T \times pm}$ and $S^* \in \mathbb{R}^{T \times pm}$ as the two components of the signal and the transition matrix $B^* \in \mathbb{R}^{pm \times pm}$ captures the autoregressive temporal dependence in the data. Thus, while the size of B^* does not grow with T, the same of L^* and S^* increase as T increases. The factor $\frac{1}{T}$ in our definition, thus adjusts for this increasing size of L^* and S^*.

We assume that $\boldsymbol{\mathcal{L}}^*$ has tubal rank $r \ll \min\{p, m\}$. The rank of the associated block-circulant matrix is denoted by R and it is bounded above by $r \times T$ (see Proposition S3.1 in the supplement). We then assume that the matrix S^* has $s_1 \ll pmT$ non-zero elements. Similarly, we assume

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

that the matrix B^* has $s_2 \ll p^2m^2$ non-zero elements. The roadmap of the technical developments is as follows: We first summarize and describe our assumptions. For deterministic realizations and under certain regularity conditions, Lemma 3.1 establishes the bound on the estimation error $e^2(\hat{L}, \hat{S}, \hat{B})$. Theorem 3.2 and Theorem S6.1 extend the result to random realizations under Gaussian and Sub-Exponential distribution respectively. The proofs of the results are delegated to Section S3 of the supplement.

Assumption 1. *The loss function satisfies Restricted Strong Convexity with curvature* $\gamma > 0$ *(and tolerance* $\tau_L = 0$*) over the set, characterized by Lemma S1.1 and Lemma S1.2 in the supplement. In other words, there exists a positive constant* $\gamma > 0$ *such that* $\frac{1}{2}\left\|Z\Delta_B^T\right\|_F^2 \geq \frac{\gamma}{2}\left\|\Delta_B\right\|_F^2$,

for all Δ_B satisfying equation (S1.2) in the supplement.

Assumption 2. $\left\|Circ(\boldsymbol{\mathcal{L}}^*)\right\|_\infty \leq \frac{\alpha_1}{\sqrt{pmT}}$ *and* $\left\|F\right\|_\infty \leq \frac{\alpha_2}{\sqrt{pmT}\left\|Z\right\|_{sp}}$, *where* α_1 *and* α_2 *are some fixed parameters*

Assumption 3. *Under the deterministic setup, the regularizer parameters* $(\lambda_L, \lambda_S, \lambda_B)$ *satisfy* $\lambda_L \geq 4\frac{1}{T}\left\|Circ(\boldsymbol{\mathcal{U}})\right\|_{sp}, \lambda_S \geq 8\left\|\frac{U}{\sqrt{T}}\right\|_\infty + \frac{4\alpha_1}{\sqrt{pmT}}$ *and* $\lambda_B \geq 8\left\|\frac{Z'U}{T}\right\|_\infty + \frac{4\alpha_2}{\sqrt{pmT}}$

Assumption 1 ensures that the loss function exhibits strong convexity over some restricted set of interest, as defined in equation (S1.2) in the

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

supplement. This is a fairly standard assumption in the high-dimensional literature Agarwal et al. (2012). As discussed in Section 2.1, Assumption 2 is aimed to ensure that the low tubal-rank component $\boldsymbol{\mathcal{L}}^*$ is incoherent with the sparse component $\boldsymbol{\mathcal{S}}^*$ and the overall signal, say F, which is the sum of L^* and S^* in Equation (2.5), is incoherent with the third component ZB^{*T}. It is worth recalling that this assumption is a straightforward application of the "spikiness" restriction on the elements of the low-rank matrix, as introduced in Agarwal et al. (2012). We directly impose that restriction on $Circ(\boldsymbol{\mathcal{L}}^*)$, which is a reasonable low rank matrix-counterpart of our low tubal-rank component $\boldsymbol{\mathcal{L}}^*$. The reader may revisit Section 2.1 for more details. This assumption is milder than the other Tensor Incoherence conditions (see Zhang and Aeron (2016)), which involve the components of the t-SVD. Assumption 3 imposes a certain lower bound to the three regularizer parameters, a common requirement in the high-dimensional literature.

The following lemma establishes an upper bound to $e^2(\hat{L}, \hat{S}, \hat{B})$ in the case of deterministic realizations.

Lemma 3.1. *In the deterministic case and under the Assumptions (1), (2) and (3), the estimation error $e^2(\hat{L}, \hat{S}, \hat{B})$ satisfies the following:*

$$e^2(\hat{L}, \hat{S}, \hat{B}) \preceq \lambda_L^2 \; r + \lambda_S^2 \; s_1 + \lambda_B^2 \; s_2 \tag{3.2}$$

where the notation '$\preceq$' denotes an upper bound, ignoring all the constant

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

factor.

Note that the result is broadly in line with Theorem 1 of Agarwal et al. (2012); specifically, when the loss function satisfies Restricted Strong Convexity with tolerance $\tau_L = 0$ and the parameters of interest are exactly (not approximately) Low Rank and Sparse, a similar form of error bound is obtained. In the current setting, the tubal-rank (instead of matrix rank) enters the bound, as well the sparsity s_1 and s_2 reflecting the nature of S^* and B^* respectively.

Next, the above result is extended under different distributional assumptions. We start with the Gaussian case. Recalling the model posited in equation (2.2), we assume that $Vec(U_t)$ are i.i.d. $N(0, \sigma^2 I_{pm})$ and $Vec(F_t)$ are i.i.d. $N(0, \Sigma_F)$ for $t = 1, 2, \cdots, T$ with $Cov(Vec(U_t), Vec(F_{t'})) = 0 \forall (t, t')$. For any discrete time, centered, covariance-stationary process $\{p_t\}$, we use the notation $\Gamma_p(h)$ to denote the corresponding autocovariance function. In other words, $\Gamma_p(h) = Cov(p_t, p_{t+h})$ where $t, h \in \mathbb{Z}$. To avoid complex notations, we use p_{1t} and p_{2t} to denote $Vec(X_{t-1})$ and $Vec(U_t)$ respectively for $t = 1, 2, \cdots, T$. Thus $\{p_{1t}\}$ and $\{p_{2t}\}$ are two centered, stationary Gaussian processes and also $Cov(p_{1t}, p_{2t}) = 0 \forall t$. As in Basu et al. (2015), we first define the spectral density corresponding to the process $\{p_{1t}\}$ as follows $f_{p_1}(\theta) = \frac{1}{2\pi} \sum_{\ell=-\infty}^{\infty} \Gamma_{p_1}(\ell) e^{-i\ell\theta}, \theta \in [-\pi, \pi]$, and assume that it exists with

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

its maximum eigen value being bounded a.e. on $[-\pi, \pi]$. In terms of notation, this implies that $\mathcal{M}(f_{p_1}) = \underset{\theta\in[-\pi,\pi]}{\text{ess sup}}\Lambda_{\max}(f_{p_1}(\theta)) < \infty$. Similarly, the maximum eigen value of the spectral density corresponding to the process $\{p_{2t}\}$ is denoted by $\mathcal{M}(f_{p_2})$ and we assume that $\mathcal{M}(f_{p_2}) < \infty$. Finally, as in Basu et al. (2015), we define the cross spectral density of the two processes $\{p_{1t}\}$ and $\{p_{2t}\}$ as follows: $f_{p_1,p_2}(\theta) = \frac{1}{2\pi}\sum_{\ell=-\infty}^{\infty}\Gamma_{p_1,p_2}(\ell)e^{-i\ell\theta}, \theta \in [-\pi, \pi]$ where $\Gamma_{p_1,p_2}(h) = Cov(p_{1t}, p_{2\,\overline{t+h}})$, $t, h \in \mathbb{Z}$. We assume that the above cross spectral density exists and its maximum eigen value is bounded a.e. on $[-\pi, \pi]$. In terms of the notation, $\mathcal{M}(f_{p_1,p_2}) = \underset{\theta\in[-\pi,\pi]}{\text{ess sup}}\sqrt{\Lambda_{\max}(f^*_{p_1,p_2}(\theta)f_{p_1,p_2}(\theta))}$ $< \infty$. These assumptions will be used in supplement S3 while we prove the next theorem.

Theorem 3.2. *Suppose the signals and the errors follow Gaussian distribution as discussed above and the above assumptions on the spectral and cross-spectral densities are satisfied. Then with probability greater than* $1 - 6\ exp\{-c\ (2\log(pm))\}$ *we will have,*

$$e^2(\hat{L}, \hat{S}, \hat{B}) \leq c_1\ \sigma^2\frac{r(p+m)}{T} + c_2\ [\sigma^2\frac{s_1\log(pmT)}{T} + \frac{\alpha_1^2 s_1}{pmT}]+ \\ c_3\ [\mathbb{Q}^2(B^*, \sigma^2, \Sigma_F)\frac{s_2\ 2\log(pm)}{T} + \frac{\alpha_2^2 s_2}{pmT}] \tag{3.3}$$

where, $\mathbb{Q}(B^*, \sigma^2, \Sigma_F) = \mathcal{M}(f_{p_1}) + \mathcal{M}(f_{p_2}) + \mathcal{M}(f_{p_1,p_2})$

The bound is analogous to the matrix case; for the latter, with m_1 rows, m_2 columns and rank w, the bound involves the expression $\sigma^2\frac{w(m_1+m_2)}{n}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

This comprises two parts: $w(m_1 + m_2)$ corresponds to the degrees of freedom, which is in the order of the number of free elements and a multiplicative factor $\frac{\sigma^2}{n}$ corresponding to the error variance. Analogously, the term rp (or, rm) corresponds to the r t-independent lateral slices (or, horizontal slices) and estimation of the p (or, m) tubes in that slice. The multiplicative factor now becomes $\frac{\sigma^2}{T}$.

The second part of the error bound is related to the sparse component and can be interpreted as follows: the first term $\sigma^2 \frac{s_1 \log(pmT)}{T}$ arises as a result of estimating s_1 non-zero elements of $T \times pm$ dimensional matrix S^*. Note that there are $\binom{pmT}{s_1}$ possible subsets of size s_1 and thus the numerator includes the corresponding term with the scaling $\log(\binom{pmT}{s_1}) \approx s_1 \log(pmT)$. The next term $\frac{\alpha_1^2 s_1}{pmT}$ appears due to the non-identifiability of the low tubal-rank and sparse components.

Finally, the first term of the last part is in line with the Proposition 4.1 of Basu et al. (2015). In the proof of Theorem 3.2, we arrive at the same choice of the regularizer parameter λ_B as the one given in Proposition 4.1 of Basu et al. (2015). This choice of λ_B leads to the first term of the last part. On the other hand, the second term arises due to the second incoherence condition between the signal F and the component ZB^{*T}, as discussed in Section 2.1.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

4. Performance Evaluation

In this section, we illustrate the performance of our estimation procedure described in Section 2.1, based on synthetic data under different settings. The true data generating process for the simulation is described in Section S2 of the supplement. Given the simulated data, we employ the Algorithm 1, discussed in Section 2.1, to obtain $\hat{\boldsymbol{\mathcal{L}}}$, $\hat{\boldsymbol{\mathcal{S}}}$ and $\hat{B}$. The regularization parameters λ_L, λ_S and λ_B are selected by a three-dimensional grid search method. We run the algorithm and obtain the estimates for different grids of the triplets $(\lambda_L, \lambda_S, \lambda_B)$ and select that triplet for which the tubal-rank of $\hat{\boldsymbol{\mathcal{L}}}$ and the positions of the non-zero elements in $\hat{\boldsymbol{\mathcal{S}}}$ and $\hat{B}$ are as close as possible to the tubal-rank of $\boldsymbol{\mathcal{L}}^*$ and the positions of the non-zero elements in $\boldsymbol{\mathcal{S}}^*$ and B^* respectively. It is worth mentioning that, later we develop an AIC criteria in order to select the optimum values of the regularization parameters, when the true tubal-rank and sparsity levels are unknown to us.

Performance Evaluation: We use Relative Error, Tubal-rank of $\hat{\boldsymbol{\mathcal{L}}}$ and Sensitivity and Specificity of both $\boldsymbol{\mathcal{S}}^*$ and B^* estimation as the criteria of evaluation. Small values of relative error, along with the closeness of tubal-rank of $\hat{\boldsymbol{\mathcal{L}}}$ and tubal-rank of $\boldsymbol{\mathcal{L}}^*$ characterize the quality of the estimation. In addition to that, sensitivity and specificity together as-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

sess the ability of support recovery. Considering the definition of Estimation Error provided in equation (3.1), the Relative Error (RE) is defined as $\frac{\frac{1}{T}\left\|\hat{L}-L^*\right\|_F^2+\frac{1}{T}\left\|\hat{S}-S^*\right\|_F^2+\left\|\hat{B}-B^*\right\|_F^2}{\frac{1}{T}\|L^*\|_F^2+\frac{1}{T}\|S^*\|_F^2+\|B^*\|_F^2}$. Specificity of $\boldsymbol{\mathcal{S}}^*$ estimation (SP_S) is defined as 1− False Positive Rate (FPR), where, FPR is defined as $\frac{\text{number of non-zero elements in } \hat{\boldsymbol{\mathcal{S}}}, \text{ which are actually zero in } \boldsymbol{\mathcal{S}}^*}{\text{number of elements that are zero in } \boldsymbol{\mathcal{S}}^*}$. Sensitivity of $\boldsymbol{\mathcal{S}}^*$ estimation (SN_S) is defined as $\frac{\text{number of non-zero elements in } \hat{\boldsymbol{\mathcal{S}}}, \text{ which are actually non-zero in } \boldsymbol{\mathcal{S}}^*}{\text{number of non-zero elements in } \boldsymbol{\mathcal{S}}^*}$. Specificity and sensitivity of B^* estimation can be defined in a similar way.

Using the above-mentioned criteria we evaluate the performance of our method under two different scenarios. Each scenario corresponds to some specific values of the pair (p, m), where the estimates are obtained for different values of sample size T. The objective is to show that larger sample sizes lead to better estimation results. Furthermore, within each scenario, we obtain the estimates under three different sub-cases. Now we first describe all the scenarios and the sub-cases and then summarize the results under all these cases. The results reported in the following tables are based on 100 replicates.

Scenario 1: $p = 10, m = 10$, Scenario 2: $p = 20, m = 20$. Sub-case 1: We fix the tubal-rank $r = 3$ and edge-density of $B^* = 0.04$, but vary the number of non-zero elements in $\boldsymbol{\mathcal{S}}^*$ (2 non-zero elements per slice and 4 non-zero elements per slice). Sub-case 2: Here we fix the tubal-rank

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

$r = 3$ and the number of non-zero elements in each slice of $\mathcal{S}^* = 2$, but vary the edge-density of B^* (edge-density of B^* is taken as 0.04 and 0.06 respectively). Sub-case 3: Finally, in this sub-case, we fix the number of non-zero elements in each slice of $\mathcal{S}^* = 2$ and the edge density of $B^* = 0.04$, but vary the tubal-rank of $\mathcal{L}^*$ (tubal-rank is taken as 2 and 3 respectively).

As depicted in Table S8.1 in the supplement, the relative error decrease as the sample size increases. Moreover, as the estimation is equipped with more and more samples, it becomes easier to achieve the target tubal-rank of the true $\mathcal{L}^*$. Finally, the values of the specificity and sensitivity approach to 1, with increase in the sample size. The first part of Table S8.1 (third and fourth column: Sub-case 1) fixes the value of true tubal-rank as 3 and the edge-density of B^* as 0.04 and then increases the number of non-zero elements in each slice of $\mathcal{S}^*$ from 2 to 4. This leads to an increase in the relative error, which is in accordance with the theoretical finding in Lemma 3.1. For example, with sample size 80 when one increases the number of non-zero elements in each slice of $\mathcal{S}^*$ from 2 to 4, the relative error increases from 0.30 to 0.37. The same argument follows for the other two parts of Table S8.1 (Sub-case 2 and Sub-case 3) as well. The second part (Sub-case 2) fixes the true tubal-rank and the number of non-zero elements in $\mathcal{S}^*$ and increases the edge-density of B^*. Finally, in the third part (Sub-case 3),

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

the sparsity level of $\boldsymbol{\mathcal{S}}^*$ and B^* are fixed and we increase the tubal-rank of $\boldsymbol{\mathcal{L}}^*$ from 2 to 3. However, in both the cases, the Relative Error increases as we increase the tubal-rank or the number of non-zero elements.

Table S8.2 in the supplement displays the similar results under Scenario 2, where both p and m are now increased to 20. As expected, more samples are required to achieve good performance while we increase the number of parameters. However, in all these cases, as in Scenario 1, both estimation and support recovery performance become stronger with increase in the sample size. Also, the Relative Error increases as we increase the tubal-rank (see Sub-case 1 of Table S8.2) or the number of non-zero elements in $\boldsymbol{\mathcal{S}}^*$ or in B^* (see Sub-case 2 and Sub-case 3 of Table S8.2 respectively). In Section S7 of the supplementary materials, we define AIC which is used to choose the optimum balues of λ_L, λ_S and λ_B.

Finally, we compare the predictive performance of our proposed model with regularized VAR model in Basu et al. (2015). We first fix a forecast horizon, h. Then, for each $t' \in \{T-10, T-9, \cdots, T-h\}$, we consider the data up to time point t', and use it to estimate the model parameters. Using the estimated parameters, we obtain the predicted value of $X_{t'+h}$, that is $\hat{X}_{t'+h}$, and compute the Root Mean Squared Error (RMSE) (Ghosh et al. (2019), Chakraborty et al. (2023)) as $\sqrt{\frac{1}{10-h+1}\sum_{t'=T-10}^{T-h}\frac{\left\|X_{t'+h}-\hat{X}_{t'+h}\right\|_F^2}{pm}}$. In

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

this expression, the first average, that is $\frac{\left\|X_{t'+h}-\hat{X}_{t'+h}\right\|_F^2}{pm}$, characterizes the average squared error, where the average is taken over the pm elements of the error $X_{t'+h}-\hat{X}_{t'+h}\in\mathbb{R}^{pm}$. The second average is taken over $10-h+1$ estimation samples. To assess the predictive performance of our model, we obtain the RMSE values with $h=1,2,$ and 3 for our simulated data with $p=10$, $m=10$, $T=80$, true tubal-rank of $\boldsymbol{\mathcal{L}}^*=3$, number of non-zero elements in each slice of $\boldsymbol{\mathcal{S}}^*=2$, and the edge-density of $B^*=0.04$. Then we compare them with the RMSE values corresponding to the regularized VAR model Basu et al. (2015). As depicted in Table 1, our model outperforms the other method in terms of predictive ability.

Forecast horizon (h)	Low tubal-rank plus sparse	Regularized VAR model
1	0.20	0.62
2	0.21	0.63
3	0.24	0.65

Table 1: Predictive performance using RMSE values: our model outperforms the regularized VAR model Basu et al. (2015)

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

5. Application to Macroeconomic Data

We illustrate the proposed model to a data set of $p = 16$ key macroeconomic indicators for $m = 11$ Eurozone countries - Austria, Belgium, Finland, France, Germany, Greece, Ireland, Italy, Netherlands, Portugal and Spain. The data comprise of matrix-valued time series observed quarterly, starting from 2002-Q2 to 2019-Q4. The macroeconomic variables under consideration are summarized in Table S8.3 in the supplement. To address the issue of non-stationarity, the variables were processed by applying necessary transformations, as suggested in McCracken and Ng (2020) and Stock and Watson (2005). Table S8.3 also displays the transformation for each variable, along with their sources. Recalling equation (2.2), we use the notation $X_t \in \mathbb{R}^{p\times m}$ to denote the matrix-valued observation, whose $(i,j)^{th}$ element is the value of the i^{th} variable for j^{th} country at t^{th} quarter, $i = 1, 2, \cdots, p = 16; j = 1, 2, \cdots, m = 11$ and $t = 1, 2, \cdots, T = 71$.

The estimated $\boldsymbol{\mathcal{L}}^*$, denoted by $\hat{\boldsymbol{\mathcal{L}}}$ has tubal-rank 7. This implies that out of 11 countries, there are 7 countries for which the lateral slices in $\hat{\boldsymbol{\mathcal{L}}}$ are t-linearly independent. The slices corresponding to the remaining 4 countries can be expressed as t-linear combinations of the aforementioned 7 slices. Further investigation reveals that the set of t-linearly independent slices is essentially $\{\hat{\boldsymbol{\mathcal{L}}}_{\text{Austria}}, \hat{\boldsymbol{\mathcal{L}}}_{\text{Belgium}}, \hat{\boldsymbol{\mathcal{L}}}_{\text{Finland}}, \hat{\boldsymbol{\mathcal{L}}}_{\text{France}}, \hat{\boldsymbol{\mathcal{L}}}_{\text{Germany}}, \hat{\boldsymbol{\mathcal{L}}}_{\text{Greece}}, \hat{\boldsymbol{\mathcal{L}}}_{\text{Ireland}}\}$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

and any slice in the remaining set $\{\hat{\mathcal{L}}_{\text{Italy}}, \hat{\mathcal{L}}_{\text{Netherlands}}, \hat{\mathcal{L}}_{\text{Portugal}}, \hat{\mathcal{L}}_{\text{Spain}}\}$ are t-linearly dependent on the previous set. Figure S8.2 in the supplement partially depicts the slices of $\hat{\mathcal{L}}$. The t-independent slices are marked in green, whereas the ones that are t-dependent, are marked in yellow. All these countries are the members of European Union (EU) and a part of Euro-Area as they share the same currency, Euro. Furthermore, the countries under Euro-Area harmonize their economic and fiscal policies in order to meet some common objectives and achieve an increased economic stability. This fact serves as a justification of the aforementioned low-rankness along the slices of the countries in $\hat{\mathcal{L}}$.

We now move to the estimate of the sparse component of $\mathcal{F}$, which is denoted by $\hat{\mathcal{S}}$. As discussed earlier, the lateral slice corresponding to any particular country captures the additional idiosyncrasies in the signal due to a period of financial crisis or economic boom in that country. Below we present the heatmaps of estimated lateral slices of $\hat{\mathcal{S}}$ in matrix form, with variables in one dimension and the quarters in the other dimension. Figure S8.3 and Figure S8.4 in the supplement represent 8 such slices. As shown in Figure S8.4, the slices corresponding to France and Italy indicate more or less stable economic scenarios. On the other hand, slices corresponding to Greece and Portugal display comparatively large numbers of idiosyncratic

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

elements. These can be attributed to *Greek Government-debt crisis* and *Portuguese financial crisis*. The other slices can be interpreted in a similar fashion.

Finally, we present the estimate of the sparse transition matrix, denoted by $\hat{B}$. As discussed earlier, the transition matrix characterizes the underlying temporal dependence in the data. In our case, $\hat{B}$ is of order 176×176 ($pm \times pm$) and $\hat{B}_{i_r j_r,\, i_c j_c}$ is the element that captures the effect of the i_c^{th} variable of j_c^{th} country in the past on the current value of i_r^{th} variable of the j_r^{th} country, $i_r j_r = 1, 2, \cdots, pm$ and $i_c j_c = 1, 2, \cdots, pm$. We use *Circular Network Graph* to depict the non-zero elements of $\hat{B}$, where the directed edges imply the temporal dependence of present on the past. However, for a better visual representation, we divide the whole transition matrix into m^2 block matrices, each of order $p \times p$. Out of these m^2 blocks, there will be m blocks representing the intra-country temporal dependence for m countries. On the other hand, remaining $m^2 - m$ blocks represent the cross-country temporal dependence. Figure S8.5, Figure S8.6 and Figure S8.7 illustrate a few such intra-country and cross-country temporal connectivity. The directed edges reveal meaningful economic links among the variables of past and current quarter and they are in line with the findings of the other relevant literature (Ghosh et al., 2019). For instance, in case

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

of both Italy and France, the figures show that the unemployment rate of the past quarter affects the current GDP. As discussed earlier, the countries under consideration are the members of Euro-Area and they follow coordinated economic policies. Thus the cross-country temporal connectivities are also justifiable.

As in case of simulated data, here also we use the RMSE (see Section 4) values to compare the predictive performance of our proposed model with that of the matrix-valued factor model (Wang et al., 2019). As depicted in Table 2, our model outperforms the matrix-valued factor model in terms of predictive ability.

Forecast horizon (h)	Low tubal-rank plus sparse	Matrix factor model
1	0.23	0.70
2	0.25	0.71
3	0.26	0.78

Table 2: Predictive performance using RMSE values: our model outperforms the matrix-valued factor model.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

6. Discussion

In this paper, we propose a matrix-valued time series model that first captures the underlying temporal dependence in the data and then assumes that the remainder signal is decomposed into two parts. The first one corresponds to a low tubal rank tensor, which captures the baseline signal, shared across the lateral (or, horizontal) slices. On the other hand, the second component is a third-order tensor, whose frontal slices are elementwise sparse representing the additional idiosyncrasies in the signal. This decomposition, as opposed to the tucker decomposition based factor-loading representation used in related literature, expands the scope of exploring more complex structures in the data. We develop a fast and scalable Alternating Minimization algorithm to solve our convex regularized program. In the context of theoretical development, we establish a non-asymptotic interpretable upper bound to the estimation error under Gaussian and Sub-Exponentail distributional assumptions. As described in Section 2, we have used B^* to model both M_t and E_t, and this formulation facilitates the following representation

$$Vec(X_t) = Vec(F_t) + B^* Vec(X_{t-1}) + Vec(U_t).$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

If instead, we would use two different transition matrices, say B_1^* and B_2^*, such that, $Vec(M_t) = B_1^* Vec(M_{t-1}) + Vec(F_t)$ and $Vec(E_t) = B_2^* Vec(E_{t-1}) + Vec(U_t)$, then the resulting model would not be easily tractable. This can be thought of as one of the potential future extension of this work.

7. Supplementary Material

Proofs of the theoretical results and some additional tables and figures have been presented in the online supplementary material.

8. Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C160004 to the University of Florida. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

Agarwal, A., S. Negahban, M. J. Wainwright, et al. (2012). Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics 40*(2), 1171–1197.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

Ahelegbey, D. F., M. Billio, and R. Casarin (2016a). Bayesian graphical models for structural vector autoregressive processes. *Journal of Applied Econometrics 31*(2), 357–386.

Ahelegbey, D. F., M. Billio, and R. Casarin (2016b). Sparse graphical vector autoregression: a bayesian approach. *Annals of Economics and Statistics/Annales d'Économie et de Statistique* (123/124), 333–361.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica 71*(1), 135–171.

Bai, J. and P. Wang (2015). Identification and bayesian estimation of dynamic factor models. *Journal of Business & Economic Statistics 33*(2), 221–240.

Bańbura, M., D. Giannone, and L. Reichlin (2010). Large bayesian vector auto regressions. *Journal of applied Econometrics 25*(1), 71–92.

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

Bernanke, B. S., J. Boivin, and P. Eliasz (2005). Measuring the effects of

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

monetary policy: a factor-augmented vector autoregressive (favar) approach. *The Quarterly journal of economics 120*(1), 387–422.

Billio, M., R. Casarin, and L. Rossini (2019). Bayesian nonparametric sparse var models. *Journal of Econometrics 212*(1), 97–115.

Blanchard, O. and R. Perotti (2002). An empirical characterization of the dynamic effects of changes in government spending and taxes on output. *the Quarterly Journal of economics 117*(4), 1329–1368.

Candès, E. J., X. Li, Y. Ma, and J. Wright (2011). Robust principal component analysis? *Journal of the ACM (JACM) 58*(3), 1–37.

Chakraborty, N., K. Khare, and G. Michailidis (2023). A bayesian framework for sparse estimation in high-dimensional mixed frequency vector autoregressive models. *Statistica Sinica 33*, 1629–1652.

Chang, J., B. Guo, and Q. Yao (2018). Principal component analysis for second-order stationary vector time series. *The Annals of Statistics 46*(5), 2094–2124.

Chen, E. Y. and R. Chen (2022). Modeling dynamic transport network with matrix factor models: an application to international trade flow. *Journal of Data Science 21*(3), 490–507.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

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

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

De Mol, C., D. Giannone, and L. Reichlin (2008). Forecasting using a large number of predictors: Is bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics 146*(2), 318–328.

George, E. I., D. Sun, and S. Ni (2008). Bayesian stochastic search for var model restrictions. *Journal of Econometrics 142*(1), 553–580.

Ghosh, S., K. Khare, and G. M. and (2019). High-dimensional posterior consistency in bayesian vector autoregressive models. *Journal of the American Statistical Association 114*(526), 735–748. PMID: 31474783.

Kilmer, M. E., K. Braman, N. Hao, and R. C. Hoover (2013). Third-order tensors as operators on matrices: A theoretical and computational framework with applications in imaging. *SIAM Journal on Matrix Analysis and Applications 34*(1), 148–172.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

Kilmer, M. E. and C. D. Martin (2011). Factorization strategies for third-order tensors. *Linear Algebra and its Applications 435*(3), 641–658.

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

Kolda, T. G. and B. W. Bader (2009). Tensor decompositions and applications. *SIAM review 51*(3), 455–500.

Lam, C., Q. Yao, et al. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics 40*(2), 694–726.

Liu, X. and T. Zhang (2022). Estimating change-point latent factor models for high-dimensional time series. *Journal of Statistical Planning and Inference 217*, 69–91.

Liu, X.-Y., S. Aeron, V. Aggarwal, and X. Wang (2016). Low-tubal-rank tensor completion using alternating minimization. In *Modeling and Simulation for Defense Systems and Applications XI*, Volume 9848, pp. 984809. International Society for Optics and Photonics.

Loh, P.-L. and M. J. Wainwright (2012). High-dimensional regression with

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics 40*(3), 1637 – 1664.

Lu, C., J. Feng, Y. Chen, W. Liu, Z. Lin, and S. Yan (2016). Tensor robust principal component analysis: Exact recovery of corrupted low-rank tensors via convex optimization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5249–5257.

Lu, C., J. Feng, Z. Lin, and S. Yan (2018, 7). Exact low tubal rank tensor recovery from gaussian measurements. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pp. 2504–2510. International Joint Conferences on Artificial Intelligence Organization.

McCracken, M. and S. Ng (2020). Fred-qd: A quarterly database for macroeconomic research. Technical report, National Bureau of Economic Research.

Michailidis, G. and F. d'Alché Buc (2013). Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues. *Mathematical biosciences 246*(2), 326–334.

Seth, A. K., A. B. Barrett, and L. Barnett (2015). Granger causality anal-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0009

ysis in neuroscience and neuroimaging. *Journal of Neuroscience 35*(8), 3293–3297.

Stock, J. H. and M. W. Watson (2005). An empirical comparison of methods for forecasting using many predictors. *Manuscript, Princeton University 46*.

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

Wang, D., Y. Zheng, and G. Li (2024). High-dimensional low-rank tensor autoregressive time series modeling. *Journal of Econometrics 238*(1), 105544.

Zhang, Z. and S. Aeron (2016). Exact tensor completion using t-svd. *IEEE Transactions on Signal Processing 65*(6), 1511–1526.