Supplementary to A Regularized low tubal-rank model for high-dimensional time series data

Samrat Roy, George Michailidis

Indian Institute of Management Ahmedabad and University of Florida

Supplementary Material

Notation: One dimensional sections of a third-order tensor $\mathcal{X} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ are defined as *Column Fiber*, *Row Fiber* and *Tube Fiber* (see Figure 2.1 of Kolda and Bader (2009)) and they are denoted by $\boldsymbol{x_{:jk}}$, $\boldsymbol{x_{i:k}}$ and $\boldsymbol{x_{ij:}}$ respectively. Similarly, the two-dimensional sections of $\mathcal{X}$, namely, *Horizontal Slice*, *Lateral Slice* and *Frontal Slice* (see Figure 2.2 of Kolda and Bader (2009)) are denoted by $\boldsymbol{X_{i::}}$, $\boldsymbol{X_{:j:}}$ and $\boldsymbol{X_{::k}}$ respectively. As illustrated in Figure 2 of the main paper, the lateral slices, horizontal slices and tube fibers can be visualized as columns, rows and elements of a matrix. For any matrix $A \in \mathbb{R}^{d_1 \times d_2}$, whose $(i,j)^{th}$ element is denoted by a_{ij}, the Frobenius Norm is defined as $\|A\|_F = \sqrt{\sum_{i=1}^{d_1}\sum_{j=1}^{d_2} a_{ij}^2}$. ℓ_∞ norm of matrix A is defined by $\|A\|_\infty = \underset{i,j}{max} \mid a_{ij} \mid$. $\ell_{2,1}$ norm of A is defined as $\|A\|_{2,1} = \sum_{j=1}^{d_2}(\sum_{i=1}^{d_1} a_{ij}^2)^{\frac{1}{2}}$. Similarly, $\ell_{2,\infty}$ norm of A is given by, $\|A\|_{2,\infty} =$

$\underset{1 \le j \le d_2}{max} (\sum_{i=1}^{d_1} a_{ij}^2)^{\frac{1}{2}}$. Denoting by $\sigma_1(A), \sigma_2(A), \cdots, \sigma_d(A)$, the singular values of A, where $d = min\{d_1, d_2\}$, we define the Nuclear Norm of A by $\|A\|_* = \sum_{j=1}^{d} \sigma_j(A)$ and the Spectral Norm of A by $\|A\|_{sp} = \underset{1 \le j \le d}{max}\{\sigma_j(A)\}$.

S1 Additional notation required for Section 3

We assume that S^* is supported on a subset $E \subseteq \{1, 2, \cdots, pmT\}$, with $|E| = s_1$. We define a pair of subspaces $(\mathbb{M}(E), \mathbb{M}^{\perp}(E))$, such that, $\mathbb{M}(E) = \{M \in \mathbb{R}^{T \times pm} \mid k^{th}$ element of $M = 0, \forall k \notin E\}$ and $\mathbb{M}^{\perp}(E) = (\mathbb{M}(E))^{\perp}$. As shown in Agarwal et al. (2012) and Negahban et al. (2012), one can easily verify that for any $M_1 \in \mathbb{M}(E)$ and $M_2 \in \mathbb{M}^{\perp}(E)$, $\|M_1 + M_2\|_1 = \|M_1\|_1 + \|M_2\|_1$. This ensures that the regularizer $\|\cdot\|_1$ is *decomposable* (see Negahban et al. (2012)) with respect to the subspace pair $(\mathbb{M}(E), \mathbb{M}^{\perp}(E))$. Simplifying the notation from $(\mathbb{M}(E), \mathbb{M}^{\perp}(E))$ to $(\mathbb{M}, \mathbb{M}^{\perp})$, it is evident that, $S^* \in \mathbb{M}$, $\pi_{\mathbb{M}}(S^*) = S^*$ and $\pi_{\mathbb{M}^{\perp}}(S^*) = 0$, where $\pi_{\mathbb{M}}(\cdot)$ is the projection onto the subspace $\mathbb{M}$. Similarly we assume that B^* is supported on a subset $H \subseteq \{1, 2, \cdots, p^2m^2\}$, with $|H| = s_2$. As before, here also we define a subspace pair $(\mathbb{N}(H), \mathbb{N}^{\perp}(H))$ such that $B^* \in \mathbb{N}$, $\pi_{\mathbb{N}}(B^*) = B^*$ and $\pi_{\mathbb{N}^{\perp}}(B^*) = 0$. Finally, we define $\hat{\Delta}_L = \hat{L} - L^*$, $\hat{\Delta}_S = \hat{S} - S^*$ and $\hat{\Delta}_B = \hat{B} - B^*$ and let $\hat{\Delta}_{\mathcal{L}} \in \mathbb{R}^{p \times m \times T}$ be the tensor form of the matrix $\hat{\Delta}_L \in \mathbb{R}^{T \times pm}$. Also, $\hat{\Delta}_S^{\mathbb{M}} = \pi_{\mathbb{M}}(\hat{\Delta}_S)$, $\hat{\Delta}_S^{\mathbb{M}^{\perp}} = \pi_{\mathbb{M}^{\perp}}(\hat{\Delta}_S)$, $\hat{\Delta}_B^{\mathbb{N}} = \pi_{\mathbb{N}}(\hat{\Delta}_B)$ and

$\hat{\Delta}_B^{\mathbb{N}^\perp} = \pi_{\mathbb{N}^\perp}(\hat{\Delta}_B)$.

Lemma S1.1. *Let R denote the rank of $Circ(\boldsymbol{\mathcal{L}}^*)$. Let $C(L,S,B)$ be a weighted combination of the nuclear norm and the ℓ_1 norm regularizers as follows:*

$$C(L,S,B) = \tfrac{1}{T}\left\|Circ(\boldsymbol{\mathcal{L}})\right\|_* + \tfrac{\lambda_S}{\lambda_L}\left\|S\right\|_1 + \tfrac{\lambda_B}{\lambda_L}\left\|B\right\|_1$$

Then, for any $R = 1, 2, \cdots, \min\{pT, mT\}$, there exists a decomposition $\hat{\Delta}_L = \hat{\Delta}_L^A + \hat{\Delta}_L^B$ with Rank $(Circ(\hat{\Delta}_{\mathcal{L}}^A)) \leq 2R$, $Circ(\boldsymbol{\mathcal{L}}^*)^T Circ(\hat{\Delta}_{\mathcal{L}}^B) = 0$, $Circ(\boldsymbol{\mathcal{L}}^*)Circ(\hat{\Delta}_{\mathcal{L}}^B)^T = 0$, where $\hat{\Delta}_{\mathcal{L}}^A$ and $\hat{\Delta}_{\mathcal{L}}^B$ in $\mathbb{R}^{p\times m\times T}$ are the tensor form of the matrices $\hat{\Delta}_L^A$ and $\hat{\Delta}_L^B$ in $\mathbb{R}^{T\times pm}$ respectively. Also the following inequality holds

$$\begin{aligned} C(L^*,S^*,B^*) - C(L^* + \hat{\Delta}_L, S^* + \hat{\Delta}_S, B^* + \hat{\Delta}_B) \leq C(\hat{\Delta}_L^A, \hat{\Delta}_S^{\mathbb{M}}, \hat{\Delta}_B^{\mathbb{N}}) \\ - C(\hat{\Delta}_L^B, \hat{\Delta}_S^{\mathbb{M}^\perp}, \hat{\Delta}_B^{\mathbb{N}^\perp}) \end{aligned} \tag{S1.1}$$

Lemma S1.2. *Under the conditions $\lambda_L \geq 4\frac{1}{T}\left\|Circ(\boldsymbol{\mathcal{U}})\right\|_{sp}$, $\lambda_S \geq 8\left\|\frac{U}{\sqrt{T}}\right\|_\infty$ and $\lambda_B \geq 8\left\|\frac{Z'U}{T}\right\|_\infty$, the estimation error $(\hat{\Delta}_L, \hat{\Delta}_S, \hat{\Delta}_B)$ satisfies the following constraint:*

$$C(\hat{\Delta}_L^B, \hat{\Delta}_S^{\mathbb{M}^\perp}, \hat{\Delta}_B^{\mathbb{N}^\perp}) \leq 3C(\hat{\Delta}_L^A, \hat{\Delta}_S^{\mathbb{M}}, \hat{\Delta}_B^{\mathbb{N}}) \tag{S1.2}$$

The above lemmas characterize a set in which the error $(\hat{\Delta}_L, \hat{\Delta}_S, \hat{\Delta}_B)$ lies.

S2 Data generation process for simulation

We first begin with generating $Vec(F_t)$ from i.i.d. $N(0, \Sigma_F)$, for $t = 1, 2, \cdots, T$. Recall that, the first component of $Vec(F_t)$ is $Vec(L_t^*)$. Hence it is necessary to generate $Vec(L_t^*)$ for $t = 1, 2, \cdots, T$ from i.i.d. $N(0, \Sigma_L)$, such that the tensor $\boldsymbol{\mathcal{L}}^* \in \mathbb{R}^{p \times m \times T}$, for which the t^{th} frontal slice is created by $L_t^* \in \mathbb{R}^{p \times m}$, will have tubal rank r. The following lemma facilitates the above requirement.

Lemma S2.1. *Let $\mathcal{G} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ be a third-order tensor, whose first r lateral slices, denoted by $G_{:1:}, G_{:2:}, \cdots, G_{:r:} \in \mathbb{R}^{d_1 \times 1 \times d_3}$, are filled with entries from i.i.d. $N(0, \sigma_l^2)$ and the remaining $d_2 - r$ lateral slices are obtained by the following t-linear combination:*

$$G_{:i:} = G_{:1:} * \alpha_{1i} + G_{:2:} * \alpha_{2i} + \cdots + G_{:r:} * \alpha_{ri}, \text{ for } i = r+1, r+2, \cdots, d_2 \quad \text{(S2.1)}$$

and α_{ji}'s in $\mathbb{R}^{1 \times 1 \times d_3}$ are the tube fibers of the form $(\beta_{ji}, 0, 0, \cdots, 0)^T$, where $\beta_{ji} \in \mathbb{R}$, for $j = 1, 2, \cdots, r$ and $i = r+1, r+2, \cdots, d_2$. Then,

1. *The tensor $\mathcal{G}$ will have tubal rank r and*

2. *$Vec(G_k) \sim$ i.i.d. $N(0, \Sigma_G)$, for $k = 1, 2, \cdots, d_3$, where G_k is the k^{th} frontal slice of $\mathcal{G}$ and Σ_G will have the following form*

$$\Sigma_G = \left[\begin{array}{c|c} M_{rd_1 \times rd_1} & N_{rd_1 \times (d_2-r)d_1} \\ \hline N'_{(d_2-r)d_1 \times rd_1} & P_{(d_2-r)d_1 \times (d_2-r)d_1} \end{array}\right]$$

where,

$$M = \begin{bmatrix} \sigma_l^2 I & 0 & \cdots & 0 \\ 0 & \sigma_l^2 I & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_l^2 I \end{bmatrix}$$

$$N = \begin{bmatrix} \beta^2_{1\,\overline{r+1}}\ \sigma_l^2 I & \beta^2_{1\,\overline{r+2}}\ \sigma_l^2 I & \cdots & \beta^2_{1\,d_2}\ \sigma_l^2 I \\ \beta^2_{2\,\overline{r+1}}\ \sigma_l^2 I & \beta^2_{2\,\overline{r+2}}\ \sigma_l^2 I & \cdots & \beta^2_{2\,d_2}\ \sigma_l^2 I \\ \vdots & \vdots & \ddots & \vdots \\ \beta^2_{r\,\overline{r+1}}\ \sigma_l^2 I & \beta^2_{r\,\overline{r+2}}\ \sigma_l^2 I & \cdots & \beta^2_{r\,d_2}\ \sigma_l^2 I \end{bmatrix}$$

and

$$P = \begin{bmatrix} \sum_{j=1}^r \beta^2_{j\,\overline{r+1}}\ \sigma_l^2 I & \sum_{j=1}^r \beta_{j\,\overline{r+1}}\beta_{j\,\overline{r+2}}\ \sigma_l^2 I & \cdots & \sum_{j=1}^r \beta_{j\,\overline{r+1}}\beta_{j\,d_2}\ \sigma_l^2 I \\ \sum_{j=1}^r \beta_{j\,\overline{r+2}}\beta_{j\,\overline{r+1}}\ \sigma_l^2 I & \sum_{j=1}^r \beta^2_{j\,\overline{r+2}}\ \sigma_l^2 I & \cdots & \sum_{j=1}^r \beta_{j\,\overline{r+2}}\beta_{j\,d_2}\ \sigma_l^2 I \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{j=1}^r \beta_{j\,d_2}\beta_{j\,\overline{r+1}}\ \sigma_l^2 I & \sum_{j=1}^r \beta_{j\,d_2}\beta_{j\,\overline{r+2}}\ \sigma_l^2 I & \cdots & \sum_{j=1}^r \beta^2_{j\,d_2}\ \sigma_l^2 I \end{bmatrix}$$

where I *and* 0 *are the Identity matrix and Null matrix of order* $d_1 \times d_1$ *respectively.*

Thus, in order to generate $Vec(L_t^*)$, we first obtain r lateral slices $L_{:1:}, L_{:2:}, \cdots, L_{:r:} \in \mathbb{R}^{p \times 1 \times T}$ filled with entries from i.i.d. $N(0, \sigma_l^2)$ and then generate the remaining $m-r$ lateral slices by the t-linear combination of $L_{:1:}, L_{:2:}, \cdots, L_{:r:}$ as governed by equation (S2.1). Thus, as proved in Lemma S2.1, the resulting tensor $\boldsymbol{\mathcal{L}}^*$ will have tubal-rank r and $Vec(L_t^*) \sim$ i.i.d. $N(0, \Sigma_L)$, for $t = 1, 2, \cdots, T$, where L_t^* is the t^{th} frontal slice of $\boldsymbol{\mathcal{L}}^*$ and Σ_L will have the same form as of Σ_G.

The second component of $Vec(F_t)$ is $Vec(S_t^*)$ and we need to generate them from i.i.d. $N(0, \Sigma_S)$, such that the matrix $S^* \in \mathbb{R}^{T \times pm}$ will have $s_1 \ll pmT$ non-zero elements. To that end, we start with drawing $Z_{ij} \overset{i.i.d.}{\sim} \text{Ber}(p_s)$, for $i = 1, 2, \cdots, p$ and $j = 1, 2, \cdots, m$, such that $E[\sum_{i=1}^{p} \sum_{j=1}^{m} Z_{ij}] = \frac{s_1}{T}$, which in other words implies that $p_s = \frac{s_1}{pmT}$. Now, for a particular pair (i, j) and given $Z_{ij} = z_{ij}$, we draw $S_t^*(i, j) \overset{i.i.d.}{\sim} z_{ij} N(0, \sigma_s^2)$ for $t = 1, 2, \cdots, T$, where $S_t^*(i, j)$ is the $(i, j)^{th}$ element of the matrix S_t^*. Thus, given $Z_{ij} = z_{ij}$, $Vec(S_t^*) \overset{i.i.d.}{\sim} N(0, \Sigma_S)$, where $\Sigma_S \in \mathbb{R}^{pm \times pm}$ is a diagonal matrix with diagonal elements as $z_{ij}^2 \sigma_s^2$ for $i = 1, 2, \cdots, p$ and $j = 1, 2, \cdots, m$. Also, $Vec(S_t^*)$ is generated independently of $Vec(L_t^*)$ for $t = 1, 2, \cdots, T$ and thus addition of these two components results in $Vec(F_t)$, where $Vec(F_t) \overset{i.i.d.}{\sim} N(0, \Sigma_F)$ for $t = 1, 2, \cdots, T$.

The transition matrix $B^* \in \mathbb{R}^{pm \times pm}$ is generated with $s_2 \ll p^2 m^2$ non-zero elements, where the non-zero elements are filled with draws from Uniform distribution. However, once the non-zero elements are drawn, the matrix is scaled down in such a way that the maximum of the absolute eigenvalues, which is the Spectral Radius, is smaller than a pre-fixed upper bound.

Finally, the errors $Vec(U_t)$ are drawn from i.i.d. $N(0, \sigma^2 I_{pm})$ distribution and the data $Vec(X_t)$ for $t = 1, 2, \cdots, T$, are generated recursively using the equation (2.3) in the main paper.

S3 Proofs

In this section, we prove the results that have been discussed in Section 3 of the main paper. Before proving the main results, we first prove a simple proposition, followed by Lemma S1.1, Lemma S1.2 and then a simple inequality, named as, *Basic Inequality.* Based on these results we then prove Lemma 3.1, which provides an upper bound to the estimation error in the deterministic case. Finally, using Lemma 3.1, we prove Theorem 3.2 and Theorem S6.1, which give the upper bound in case of Gaussian and Sub-Exponential distributional assumptions.

Proposition S3.1. $r \leq R \leq rT$

Proof. Note that, a block-circulant matrix of a third-order tensor in $\mathbb{R}^{p \times m \times T}$, can be expressed as $[B_1|B_2|\cdots|B_m]$, where the j^{th} block B_j is a matrix of dimension $pT \times T$, $j = 1, 2, \cdots m$. In each B_j, the first column is $Vec(j^{th}$ Lateral Slice of the tensor) and the remaining $(T-1)$ columns are just a circulant rearrangement of the first column. Since the tubal rank of the tensor is r, there will be r blocks among these m blocks such that:

- any column of any of the remaining $m-r$ blocks can be written as a linear combination of the columns of the aforementioned r blocks and
- any column of j_1^{th} block is linearly independent of any column of j_2^{th} block, where $j_1 \neq j_2, j_1, j_2 = 1, 2, \cdots, r$.

So the rank of the block-circulant matrix will depend on the intra-block linear dependence of these r blocks. If all the columns within each of the r blocks are linearly independent, then there will be $r \times T$ linearly independent columns in the full block-circulant matrix. At the other extreme, if there is only one linearly independent column in each of the r blocks, then there will be r linearly independent columns in the block-circulant matrix. Hence the proof. □

S3.1 Proof of Lemma S1.1

Proof. Note that $Circ(\boldsymbol{\mathcal{L}}^*)$ and $Circ(\hat{\Delta}_{\mathcal{L}})$ are the two matrices of the same dimension. Using Lemma 3.4 of Recht et al. (2010), it is possible to decompose $Circ(\hat{\Delta}_{\mathcal{L}})$ as $Circ(\hat{\Delta}_{\mathcal{L}}^{A}) + Circ(\hat{\Delta}_{\mathcal{L}}^{B})$, such that, Rank$(Circ(\hat{\Delta}_{\mathcal{L}}^{A})) \leq 2$ Rank $(Circ(\boldsymbol{\mathcal{L}}^{*})) = 2R$ and $Circ(\boldsymbol{\mathcal{L}}^{*})^{T} Circ(\hat{\Delta}_{\mathcal{L}}^{B}) = 0$, $Circ(\boldsymbol{\mathcal{L}}^{*}) Circ(\hat{\Delta}_{\mathcal{L}}^{B})^{T} = 0$. The reader may visit Recht et al. (2010) to know the details on how to derive such decomposition. It is worth mentioning that, Agarwal et al. (2012) uses the same tool while proving their Lemma 1. However, as Lemma 2.3 of Recht et al. (2010) proves, the last two equalities are essentially the sufficient condition of the additivity of nuclear norm. In other words, these imply

$$\left\| Circ(\boldsymbol{\mathcal{L}}^{*}) + Circ(\hat{\Delta}_{\mathcal{L}}^{B}) \right\|_{*} = \left\| Circ(\boldsymbol{\mathcal{L}}^{*}) \right\|_{*} + \left\| Circ(\hat{\Delta}_{\mathcal{L}}^{B}) \right\|_{*} \tag{S3.1}$$

We use the above finding later in this proof. It now remains to show that inequality (S1.1) holds for such decomposition. Note that,

$$C(L^* + \hat{\Delta}_L, S^* + \hat{\Delta}_S, B^* + \hat{\Delta}_B)$$

$$= \frac{1}{T} \left\| Circ(\boldsymbol{\mathcal{L}}^*) + Circ(\hat{\Delta}_{\mathcal{L}}) \right\|_* + \frac{\lambda_S}{\lambda_L} \left\| S^* + \hat{\Delta}_S \right\|_1 + \frac{\lambda_B}{\lambda_L} \left\| B^* + \hat{\Delta}_B \right\|_1$$

,by the definition of $C(L, S, B)$ and the fact that $Circ(\cdot)$ is additive

$$= \frac{1}{T} \left\| Circ(\boldsymbol{\mathcal{L}}^*) + Circ(\hat{\Delta}^A_{\mathcal{L}}) + Circ(\hat{\Delta}^B_{\mathcal{L}}) \right\|_* + \frac{\lambda_S}{\lambda_L} \left\| S^* + \hat{\Delta}^{\mathbb{M}}_S + \hat{\Delta}^{\mathbb{M}^\perp}_S \right\|_1 + \frac{\lambda_B}{\lambda_L} \left\| B^* + \hat{\Delta}^{\mathbb{N}}_B + \hat{\Delta}^{\mathbb{N}^\perp}_B \right\|_1$$

,by the aforementioned decomposition and the property of projection

$$\geq \frac{1}{T} \left\| Circ(\boldsymbol{\mathcal{L}}^*) + Circ(\hat{\Delta}^B_{\mathcal{L}}) \right\|_* - \frac{1}{T} \left\| Circ(\hat{\Delta}^A_{\mathcal{L}}) \right\|_* + \frac{\lambda_S}{\lambda_L} \left\| S^* + \hat{\Delta}^{\mathbb{M}^\perp}_S \right\|_1 - \frac{\lambda_S}{\lambda_L} \left\| \hat{\Delta}^{\mathbb{M}}_S \right\|_1 + \frac{\lambda_B}{\lambda_L} \left\| B^* + \hat{\Delta}^{\mathbb{N}^\perp}_B \right\|_1 - \frac{\lambda_B}{\lambda_L} \left\| \hat{\Delta}^{\mathbb{N}}_B \right\|_1$$

, by the Triangle Inequality

$$\geq \frac{1}{T} \left\| Circ(\boldsymbol{\mathcal{L}}^*) \right\|_* + \frac{1}{T} \left\| Circ(\hat{\Delta}^B_{\mathcal{L}}) \right\|_* - \frac{1}{T} \left\| Circ(\hat{\Delta}^A_{\mathcal{L}}) \right\|_* + \frac{\lambda_S}{\lambda_L} \left\| S^* \right\|_1 + \frac{\lambda_S}{\lambda_L} \left\| \hat{\Delta}^{\mathbb{M}^\perp}_S \right\|_1 - \frac{\lambda_S}{\lambda_L} \left\| \hat{\Delta}^{\mathbb{M}}_S \right\|_1 + \frac{\lambda_B}{\lambda_L} \left\| B^* \right\|_1 + \frac{\lambda_B}{\lambda_L} \left\| \hat{\Delta}^{\mathbb{N}^\perp}_B \right\|_1 - \frac{\lambda_B}{\lambda_L} \left\| \hat{\Delta}^{\mathbb{N}}_B \right\|_1$$

, by equation (S3.1) and the Decomposability of $\|\cdot\|_1$

The proof follows from the above inequality and the definition of $C(L^*, S^*, B^*)$.

□

S3.2 Proof of Lemma S1.2

Proof. We start with defining the following function:

$$f(\Delta_L, \Delta_S, \Delta_B) = L(L^* + \Delta_L, S^* + \Delta_S, B^* + \Delta_B) - L(L^*, S^*, B^*) + \lambda_L\{C(L^* + \Delta_L, S^* + \Delta_S, B^* + \Delta_B) - C(L^*, S^*, B^*).\}$$

$L(L, S, B)$ is used to denote the loss function given by, $\frac{1}{2T}\left\|Y - L - S - ZB^T\right\|_F^2$. Since $f(0, 0, 0) = 0$ and $(\hat{\Delta}_L, \hat{\Delta}_S, \hat{\Delta}_B)$ is the optimal error, one must have, $f(\hat{\Delta}_L, \hat{\Delta}_S, \hat{\Delta}_B) \leq f(0, 0, 0) = 0$. Recall that, we already have a lower bound of $C(L^* + \hat{\Delta}_L, S^* + \hat{\Delta}_S, B^* + \hat{\Delta}_B) - C(L^*, S^*, B^*)$ from equation (S1.1). Now our job is to find a lower bound to $L(L^* + \hat{\Delta}_L, S^* + \hat{\Delta}_S, B^* + \hat{\Delta}_B) - L(L^*, S^*, B^*)$. These two bounds, along with the fact that $f(\hat{\Delta}_L, \hat{\Delta}_S, \hat{\Delta}_B) \leq 0$, will prove the result.

Since $\lambda_L \frac{1}{T}\left\|Circ(\boldsymbol{\mathcal{L}}^*)\right\|_* + \lambda_S \left\|S^*\right\|_1 + \lambda_B \left\|B^*\right\|_1 = \lambda_L C(L^*, S^*, B^*)$, one can think C as an alternative regularizer and λ_L as the associated parameter for our problem. Now as Negahban et al. (2012) derives while proving their Lemma 1, using the convexity of the loss function and dual-norm inequality, we get the following:

$$L(L^* + \hat{\Delta}_L, S^* + \hat{\Delta}_S, B^* + \hat{\Delta}_B) - L(L^*, S^*, B^*) \geq - C^*(\nabla L(L^*, S^*, B^*))C(\hat{\Delta}_L, \hat{\Delta}_S, \hat{\Delta}_B) \tag{S3.2}$$

Where, C^* is the dual norm associated with the regularizer C. It is easy to check that $\nabla L(L^*, S^*, B^*) = [-\frac{U}{T}, -\frac{U}{T}, -\frac{Z'U}{T}]$. Now, from the given conditions on the regularizer parameters and using the similar argument as in the proof of Lemma 1 in Agarwal et al. (2012) $C^*(\nabla L(L^*, S^*, B^*))$ can be shown to be bounded above by $\frac{\lambda_L}{2}$. Also, it is easy to check that $C(\hat{\Delta}_L, \hat{\Delta}_S, \hat{\Delta}_B) \leq C(\hat{\Delta}_L^A, \hat{\Delta}_S^{\mathbb{M}}, \hat{\Delta}_B^{\mathbb{N}}) + C(\hat{\Delta}_L^B, \hat{\Delta}_S^{\mathbb{M}^\perp}, \hat{\Delta}_B^{\mathbb{N}^\perp})$. Thus (S3.2) reduces to

$$\begin{aligned} &L(L^* + \hat{\Delta}_L, S^* + \hat{\Delta}_S, B^* + \hat{\Delta}_B) - L(L^*, S^*, B^*) \geq \\ &-\frac{\lambda_L}{2}(C(\hat{\Delta}_L^A, \hat{\Delta}_S^{\mathbb{M}}, \hat{\Delta}_B^{\mathbb{N}}) + C(\hat{\Delta}_L^B, \hat{\Delta}_S^{\mathbb{M}^\perp}, \hat{\Delta}_B^{\mathbb{N}^\perp})) \end{aligned} \tag{S3.3}$$

Finally the rest of the proof follows simply from (S1.1),(S3.3) and from the fact that $f(\hat{\Delta}_L, \hat{\Delta}_S, \hat{\Delta}_B) \leq 0$. □

S3.3 Basic Inequality

$$\begin{aligned} \frac{1}{2T}\left\|\hat{\Delta}_L + \hat{\Delta}_S + Z\hat{\Delta}_B^T\right\|_F^2 \leq \frac{1}{T}\langle U, \hat{\Delta}_L + \hat{\Delta}_S + Z\hat{\Delta}_B^T\rangle + \lambda_L C(L^*, S^*, B^*) \\ - \lambda_L C(L^* + \hat{\Delta}_L, S^* + \hat{\Delta}_S, B^* + \hat{\Delta}_B) \end{aligned} \tag{S3.4}$$

Proof. By the optimality of $(\hat{L}, \hat{S}, \hat{B})$ and the feasibility of (L^*, S^*, B^*) we have the following inequality:

$$\begin{aligned} &\frac{1}{2T}\left\|Y - \hat{L} - \hat{S} - Z\hat{B}^T\right\|_F^2 + \lambda_L \frac{1}{T}\left\|Circ(\hat{\boldsymbol{\mathcal{L}}})\right\|_* + \lambda_S\left\|\hat{S}\right\|_1 + \lambda_B\left\|\hat{B}\right\|_1 \\ &\leq \frac{1}{2T}\left\|Y - L^* - S^* - ZB^{*T}\right\|_F^2 + \lambda_L\frac{1}{T}\left\|Circ(\boldsymbol{\mathcal{L}}^*)\right\|_* + \lambda_S\left\|S^*\right\|_1 + \lambda_B\left\|B^*\right\|_1 \end{aligned}$$

Now, from $Y = L^* + S^* + ZB^{*T} + U$, we will have,

$$\begin{aligned}
&\left\|Y - \hat{L} - \hat{S} - Z\hat{B}^T\right\|_F^2 \\
&= \left\|Y - (L^* + S^* + ZB^{*T}) - (\hat{L} + \hat{S} + Z\hat{B}^T) + (L^* + S^* + ZB^{*T})\right\|_F^2 \\
&= \left\|U - (\hat{\Delta}_L + \hat{\Delta}_S + Z\hat{\Delta}_B^T)\right\|_F^2 \\
&= \|U\|_F^2 + \left\|\hat{\Delta}_L + \hat{\Delta}_S + Z\hat{\Delta}_B^T\right\|_F^2 - 2\langle U, \hat{\Delta}_L + \hat{\Delta}_S + Z\hat{\Delta}_B^T\rangle
\end{aligned}$$

Using the above decomposition along with the earlier inequality in the proof, we arrive at the following inequality:

$$\begin{aligned}
&\frac{1}{2T}\|U\|_F^2 + \frac{1}{2T}\left\|\hat{\Delta}_L + \hat{\Delta}_S + Z\hat{\Delta}_B^T\right\|_F^2 - \frac{1}{T}\langle U, \hat{\Delta}_L + \hat{\Delta}_S + Z\hat{\Delta}_B^T\rangle + \\
&\lambda_L \frac{1}{T}\left\|Circ(\hat{\boldsymbol{\mathcal{L}}})\right\|_* + \lambda_S\left\|\hat{S}\right\|_1 + \lambda_B\left\|\hat{B}\right\|_1 \\
&\leq \frac{1}{2T}\|U\|_F^2 + \lambda_L\frac{1}{T}\|Circ(\boldsymbol{\mathcal{L}}^*)\|_* + \lambda_S\|S^*\|_1 + \lambda_B\|B^*\|_1
\end{aligned}$$

Hence the proof.

□

S3.4 Proof of Lemma 3.1

Proof. To avoid complex notations, in this proof, we initially ignore the factor $1/T$ in the definition of $C(L, S, B)$ and adjust that later towards the end of the proof. The reader may note that Lemma S1.1, Lemma

S1.2 and Basic Inequality hold good, with the earlier assumption $\lambda_L \geq 4\frac{1}{T}\left\|Circ(\boldsymbol{\mathcal{U}})\right\|_{sp}$ is now replaced by $\lambda_L \geq 4\left\|Circ(\boldsymbol{\mathcal{U}})\right\|_{sp}$. We start with the left hand side of the Basic Inequality (S3.4), which can be written as follows:

$$\frac{1}{2T}\left\|\hat{\Delta}_L\right\|_F^2+\frac{1}{2T}\left\|\hat{\Delta}_S\right\|_F^2+\frac{1}{2T}\left\|Z\hat{\Delta}_B^T\right\|_F^2+\frac{1}{T}\langle\hat{\Delta}_L,\hat{\Delta}_S\rangle+\frac{1}{T}\langle\hat{\Delta}_F,Z\hat{\Delta}_B^T\rangle \quad \text{(S3.5)}$$

where $\hat{\Delta}_F = \hat{\Delta}_L + \hat{\Delta}_S$. Now one can see that,

$$\begin{aligned}
&\frac{1}{T}|\langle\hat{\Delta}_L,\hat{\Delta}_S\rangle| \\
\leq &\left\|Circ(\hat{\Delta}_{\mathcal{L}})\right\|_\infty \left\|\frac{\hat{\Delta}_S}{\sqrt{T}}\right\|_1, \text{using Dual-Norm inequality} \\
\leq &\{\left\|Circ(\hat{\boldsymbol{\mathcal{L}}})\right\|_\infty + \left\|Circ(\boldsymbol{\mathcal{L}}^*)\right\|_\infty\}\left\|\frac{\hat{\Delta}_S}{\sqrt{T}}\right\|_1 \\
\leq &\frac{2\alpha_1}{\sqrt{pmT}}\left\|\frac{\hat{\Delta}_S}{\sqrt{T}}\right\|_1, \text{using Assumption 2}
\end{aligned}$$

On the other hand, from the last component of equation (S3.5) one can get,

$$\begin{aligned}
&\frac{1}{T}|\langle\hat{\Delta}_F,Z\hat{\Delta}_B^T\rangle| \\
= &\frac{1}{T}|\langle\hat{\Delta}_F^T Z,\hat{\Delta}_B\rangle| \\
\leq &\left\|\hat{\Delta}_F^T Z\right\|_\infty \left\|\hat{\Delta}_B\right\|_1, \text{using Dual-Norm inequality} \\
\leq &\left\|\hat{\Delta}_F\right\|_\infty \left\|Z\right\|_{sp} \left\|\hat{\Delta}_B\right\|_1 \\
\leq &\{\left\|\hat{F}\right\|_\infty + \left\|F\right\|_\infty\}\left\|Z\right\|_{sp}\left\|\hat{\Delta}_B\right\|_1 \\
\leq &\frac{2\alpha_2}{\sqrt{pmT}}\left\|\hat{\Delta}_B\right\|_1, \text{using Assumption 2}
\end{aligned}$$

Hence using these along with the RSC Assumption, it is easy to derive the following from equation (S3.5):

$$
\begin{aligned}
\frac{1}{2T}\left\|\hat{\Delta}_L + \hat{\Delta}_S + Z\hat{\Delta}_B^T\right\|_F^2 \geq\ & \frac{1}{2}\left\|\frac{\hat{\Delta}_L}{\sqrt{T}}\right\|_F^2 + \frac{1}{2}\left\|\frac{\hat{\Delta}_S}{\sqrt{T}}\right\|_F^2 + \frac{\gamma}{2}\left\|\hat{\Delta}_B\right\|_F^2 \\
& - \frac{2\alpha_1}{\sqrt{pmT}}\left\|\frac{\hat{\Delta}_S}{\sqrt{T}}\right\|_1 - \frac{2\alpha_2}{\sqrt{pmT}}\left\|\hat{\Delta}_B\right\|_1 \\
\geq\ & \frac{\gamma'}{2}(\left\|\frac{\hat{\Delta}_L}{\sqrt{T}}\right\|_F^2 + \left\|\frac{\hat{\Delta}_S}{\sqrt{T}}\right\|_F^2 + \left\|\hat{\Delta}_B\right\|_F^2) - \frac{\lambda_S}{2}\left\|\frac{\hat{\Delta}_S}{\sqrt{T}}\right\|_1 \\
& - \frac{\lambda_B}{2}\left\|\hat{\Delta}_B\right\|_1 \\
& \text{by Assumption 1 and taking } \gamma' = \min\{\gamma, 1\} \\
\geq\ & \frac{\gamma'}{2}(\left\|\frac{\hat{\Delta}_L}{\sqrt{T}}\right\|_F^2 + \left\|\frac{\hat{\Delta}_S}{\sqrt{T}}\right\|_F^2 + \left\|\hat{\Delta}_B\right\|_F^2) \\
& - \frac{\lambda_L}{2}c(\frac{\hat{\Delta}_L}{\sqrt{T}}, \frac{\hat{\Delta}_S}{\sqrt{T}}, \hat{\Delta}_B)
\end{aligned}
$$

Thus we have a lower bound to the left hand side of the Basic Inequality (S3.4).

Now we move on to the right hand side of Basic Inequality. Note that, the term $\frac{1}{T}\langle U, \hat{\Delta}_L + \hat{\Delta}_S + Z\hat{\Delta}_B^T\rangle$ can be written as follows:

$$
\begin{aligned}
&\langle \frac{Circ(\mathfrak{U})}{T}, \frac{Circ(\hat{\Delta}_{\mathcal{L}})}{T}\rangle + \langle \frac{U}{\sqrt{T}}, \frac{\hat{\Delta}_S}{\sqrt{T}}\rangle + \langle \frac{Z'U}{T}, \hat{\Delta}_B\rangle \\
\leq &\left\| \frac{Circ(\mathfrak{U})}{T} \right\|_{sp} \left\| Circ(\frac{\hat{\Delta}_{\mathcal{L}}}{\sqrt{T}}) \right\|_* + \left\| \frac{U}{\sqrt{T}} \right\|_\infty \left\| \frac{\hat{\Delta}_S}{\sqrt{T}} \right\|_1 + \left\| \frac{Z'U}{T} \right\|_\infty \left\| \hat{\Delta}_B \right\|_1 \\
\leq &\frac{\lambda_L}{4}\{c(\frac{\hat{\Delta}_L^A}{\sqrt{T}}, \frac{\hat{\Delta}_S^{\mathbb{M}}}{\sqrt{T}}, \hat{\Delta}_B^{\mathbb{N}}) + c(\frac{\hat{\Delta}_L^B}{\sqrt{T}}, \frac{\hat{\Delta}_S^{\mathbb{M}^\perp}}{\sqrt{T}}, \hat{\Delta}_B^{\mathbb{N}^\perp})\}
\end{aligned}
$$

,by Assumption 3 and the definition of $C(L, S, B)$.

Using the above inequality, along with Lemma S1.1, it is easy to show that the right hand side of the Basic Inequality can be upper bounded by $\frac{3\lambda_L}{2}c(\frac{\hat{\Delta}_L^A}{\sqrt{T}}, \frac{\hat{\Delta}_S^{\mathbb{M}}}{\sqrt{T}}, \hat{\Delta}_B^{\mathbb{N}})$.

With the above upper bound of the right hand side of Basic Inequality and the lower bound of the left hand side of the Basic Inequality, it is easy to show that,

$$
\frac{\gamma'}{2}(\left\| \frac{\hat{\Delta}_L}{\sqrt{T}} \right\|_F^2 + \left\| \frac{\hat{\Delta}_S}{\sqrt{T}} \right\|_F^2 + \left\| \hat{\Delta}_B \right\|_F^2) \leq 4\lambda_L c(\frac{\hat{\Delta}_L^A}{\sqrt{T}}, \frac{\hat{\Delta}_S^{\mathbb{M}}}{\sqrt{T}}, \hat{\Delta}_B^{\mathbb{N}}) \qquad \text{(S3.6)}
$$

Now recall from Lemma S1.1 that, rank of $Circ(\hat{\Delta}_{\mathcal{L}}^A)$ is at most $2R$. This fact, along with the concept of *Compatibility Constant* defined in Agarwal et al. (2012), reveals that

$$\begin{aligned}&\lambda_L c(\frac{\hat{\Delta}_L^A}{\sqrt{T}}, \frac{\hat{\Delta}_S^{\mathbb{M}}}{\sqrt{T}}, \hat{\Delta}_B^{\mathbb{N}})\\ \leq&\sqrt{2R}\lambda_L \left\| Circ(\frac{\hat{\Delta}_{\mathcal{L}}}{\sqrt{T}})\right\|_F + \lambda_S\ \sqrt{s_1}\left\|\frac{\hat{\Delta}_S}{\sqrt{T}}\right\|_F + \lambda_B\ \sqrt{s_2}\left\|\hat{\Delta}_B\right\|_F\end{aligned}$$

Now, we finally adjust the factor $\frac{1}{T}$, that we ignored in the beginning of the proof and add the same prior to $\left\|Circ(\frac{\hat{\Delta}_{\mathcal{L}}}{\sqrt{T}})\right\|_F$ in the above expression. Then using the facts that $\left\|Circ(\frac{\hat{\Delta}_{\mathcal{L}}}{\sqrt{T}})\right\|_F = \sqrt{T}\left\|\frac{\hat{\Delta}_L}{\sqrt{T}}\right\|_F$ and $R \leq rT$, the above inequality reduces to

$$\begin{aligned}&\lambda_L c(\frac{\hat{\Delta}_L^A}{\sqrt{T}}, \frac{\hat{\Delta}_S^{\mathbb{M}}}{\sqrt{T}}, \hat{\Delta}_B^{\mathbb{N}})\\ \leq&\sqrt{2r}\lambda_L \left\|\frac{\hat{\Delta}_L}{\sqrt{T}}\right\|_F + \lambda_S\ \sqrt{s_1}\left\|\frac{\hat{\Delta}_S}{\sqrt{T}}\right\|_F + \lambda_B\ \sqrt{s_2}\left\|\hat{\Delta}_B\right\|_F\end{aligned}$$

The reader may note that the above inequality is exactly the same as the one obtained in Agarwal et al. (2012), towards the very end of the proof of their Theorem 1. Hence, as done in Agarwal et al. (2012), we substitute the above inequality into inequality (S3.6) and then following the exact same steps as in Agarwal et al. (2012), we complete the proof. □

S3.5 Proof of Theorem 3.2

Proof. We start with finding a suitable choice for λ_B. As mentioned before the statement of Theorem 3.2, we recall that the processes $\{p_{1t}\}$ and $\{p_{2t}\}$ are centered, stationary, Gaussian processes with $Cov(p_{1t}, p_{2t}) = 0 \forall t$ and the maximum eigenvalue of the cross spectral density corresponding to the joint process $[{p_{1t}}', {p_{2t}}']'$ is bounded a.e. on $[-\pi, \pi]$. Hence the joint process $[{p_{1t}}', {p_{2t}}']'$ satisfy the assumption of Proposition 2.4(b) in Basu et al. (2015). Thus by applying their Proposition to our data matrices Z and U, we can say that there exists a constant $c > 0$ such that for any $u, v \in \mathbb{R}^{pm}$ such that $\|u\| \leq 1$ and $\|v\| \leq 1$, we will have

$$Pr\{|u'(\frac{Z'U}{T})v| \ > 2\pi \ \mathbb{Q}(B^*, \sigma^2, \Sigma_F)\eta\} \leq 6 \ exp[-cT \ min\{\eta^2, \eta\}] \tag{S3.7}$$

where $\mathbb{Q}(B^*, \sigma^2, \Sigma_F) = \mathcal{M}(f_{p_1}) + \mathcal{M}(f_{p_2}) + \mathcal{M}(f_{p_1,p_2})$. Now, as in the proof of Proposition 4.3 in Basu et al. (2015), we take $u = e_i$, $v = e_j$ and then applying union bound over p^2m^2 elements of $Z'U$ and finally taking $\eta = \sqrt{\frac{2\log(pm)}{T}}$, we get,

$$Pr\{\frac{\|Z'U\|_\infty}{T} > 2\pi \ \mathbb{Q}(B^*, \sigma^2, \Sigma_F)\sqrt{\frac{2\log(pm)}{T}}\} \leq 6 \ exp[-c_1 \ (2\log(pm))] \tag{S3.8}$$

Hence, we choose $\lambda_B = c_1^* \ \mathbb{Q}(B^*, \sigma^2, \Sigma_F)\sqrt{\frac{2\log(pm)}{T}} + \frac{4\alpha_2}{\sqrt{pmT}}$ for some suitably chosen constant c_1^*.

Next, we move on to finding a suitable choice for λ_L. Note that $Circ(\mathfrak{U})$ is a Gaussian Ensemble with zero mean and covariance matrix, say, Σ_{Circ}. Hence from Lemma H.1 of Negahban and Wainwright (2011)we get,

$$Pr\{\|Circ(\mathfrak{U})\|_{sp} \geq 12\ \rho(\Sigma_{Circ})(\sqrt{pT} + \sqrt{mT}) + a\} \leq exp(-\frac{a^2}{2\ \rho^2(\Sigma_{Circ})}) \tag{S3.9}$$

where $\rho^2(\Sigma_{Circ}) = \underset{\|u\|_2=1,\|v\|_2=1}{sup} Var(u'\tilde{X}v)$, for any random matrix $\tilde{X}$ sampled from Σ_{Circ}-Gaussian Ensemble. In our case, with some simple algebraic steps, it is easy to show that $\rho^2(\Sigma_{Circ}) \leq \sigma^2 c$, for some constant c and hence by choosing $a = 2\sigma\sqrt{c}(\sqrt{pT} + \sqrt{mT})$ we get,

$$Pr\{\frac{1}{T}\|Circ(\mathfrak{U})\|_{sp} \geq \frac{c_2\sigma(\sqrt{p} + \sqrt{m})}{\sqrt{T}}\} \leq exp(-2T(p+m)) \tag{S3.10}$$

Hence we choose $\lambda_L = \frac{c_2^*\sigma}{\sqrt{T}}(\sqrt{p} + \sqrt{m})$, for some suitably chosen constant c_2^*.

Finally we need to make a suitable choice of λ_S. Using the Gaussian Tail bound and then union bound over pmT elements we get,

$$Pr\{\|U\|_\infty \geq a\} \leq 2\ exp(-\frac{a^2}{2\sigma^2}) + \log(pmT) \tag{S3.11}$$

Now choosing $a = 4\sigma\sqrt{\log(pmT)}$, we get

$$Pr\{\left\|\frac{U}{\sqrt{T}}\right\|_\infty \geq \frac{4\sigma\sqrt{\log(pmT)}}{\sqrt{T}}\} \leq 2\ exp(-7\log(pmT)) \tag{S3.12}$$

Hence we choose $\lambda_S = \frac{c_3^* \sigma \sqrt{\log(pmT)}}{\sqrt{T}} + \frac{4\alpha_1}{\sqrt{pmT}}$, for some suitably chosen constant c_3^*.

Hence we have shown that the regularizer parameters satisfy conditions in Assumption 3 with high probability. Also, using Proposition 2.3 of Basu et al. (2015) and employing similar steps as in the proof of Proposition 4.2 of Basu et al. (2015), it is easy to check that the Restricted Strong Convexity holds with high probability. Hence using Lemma 3.1 and following the similar steps as in Agarwal et al. (2012), we see that with probability greater than $1 - 6\ exp\{-c_1^*(2\log(pm))\}$ we will have the following upper bound of $e^2(\hat{L}, \hat{S}, \hat{B})$

$$c_1\ \sigma^2 \frac{r(p+m)}{T} + c_2\ [\sigma^2 \frac{s_1 \log(pmT)}{T} + \frac{\alpha_1^2 s_1}{pmT}] + c_3\ [\mathbb{Q}^2(B^*, \sigma^2, \Sigma_F) \frac{s_2\ 2\log(pm)}{T} + \frac{\alpha_2^2 s_2}{pmT}] \tag{S3.13}$$

Hence the proof is complete. □

Before presenting the proof of Theorem S6.1, we first state and prove the following lemmas, which will be useful for the proof. To that end, we first define the linear process of the following form:

$$X_t = \sum_{l=0}^{\infty} B_l w_{t-l} \tag{S3.14}$$

In the case where the process is Gaussian, the w_t's correspond to Gaussian white noise process. However, we assume that w_t is a white noise process whose coordinates gave the following α-sub-exponential tail decay, that is, there exist two constants a, b such that the following holds:

$$Pr\{|w_{tj}| \geq \xi\} \leq a \ exp(-\frac{\xi^\alpha}{b}), \forall \xi > 0 \tag{S3.15}$$

The following lemma generalizes a Hanson-Wright type concentration inequality to the samples from a linear process X_t as defined in equation (S3.14).

Lemma S3.1. *Consider some generic p-dimensional linear processes given in the form of $X_t = \sum_{l=0}^{\infty} \Phi_l u_{t-l}$, where u_t's are i.i.d. and their coordinates follow α-sub-exponential tail decay, as characterized by equation (S3.15). Denote its realization by $X \in \mathbb{R}^{n\times p}$ with n consecutive observations stacked in its rows. Then, for a deterministic $np \times np$ matrix A, there exists some constant C such that the following holds:*

$$Pr\{|Vec(X')'\ AVec(X') - Exp[Vec(X')'\ AVec(X')]| > 2\pi\eta\mathcal{M}(f_X)\} \leq \tau(\eta, \alpha, a), \tag{S3.16}$$

where, $\mathcal{M}(f_X)$ *is as defined in the main paper and*

$$\tau(\eta, \alpha, A) = 2exp[-Cmin\{\frac{\eta^2}{rank(A)\left\|A\right\|_{op}^2}, (\frac{\eta}{\left\|A\right\|_{op}})^{\frac{\alpha}{2}}\}] \tag{S3.17}$$

Proof. Let $Vec(X') \stackrel{d}{=} \Omega^{\frac{1}{2}}Z$, where Ω is the covariance matrix of the np-dimensional random vector $Vec(X')$ and Z satisfies $Exp(Z) = 0$ and $Exp(ZZ')$ is I_{np}. Now applying Proposition 1.1 in Götze et al. (2019) gives

$$Pr\{|Vec(X')'\ AVec(X') - Exp[Vec(X')'\ AVec(X')]| > 2\pi\eta\mathcal{M}(f_X)\}$$
$$=Pr\{|Z'\Omega^{\frac{1}{2}}A\Omega^{\frac{1}{2}}Z - Exp[Z'\Omega^{\frac{1}{2}}A\Omega^{\frac{1}{2}}Z]| > 2\pi\eta\mathcal{M}(f_X)\}$$
$$\leq 2exp(-c_0 \cdot \nu(\Omega^{\frac{1}{2}}A\Omega^{\frac{1}{2}}, \alpha, 2\pi\eta\mathcal{M}(f_X)))$$

where

$$\nu(A, \alpha, t) = min\{\frac{t^2}{M^4\left\|A\right\|_F^2}, (\frac{t}{M^2\left\|A\right\|_{op}})^{\frac{\alpha}{2}}\} \tag{S3.18}$$

Here both c_0 and M are constants that depend on a and b. Next, we consider the bounds for various norms on $\Omega^{\frac{1}{2}}A\Omega^{\frac{1}{2}}$ as follows:

- $\left\|\Omega^{\frac{1}{2}}A\Omega^{\frac{1}{2}}\right\|_{op} \leq \left\|\Omega\right\|_{op}\left\|A\right\|_{op} \leq 2\pi\mathcal{M}(f_X)\left\|A\right\|_{op}$, where the last inequality follows from Proposition 2.3 in Basu et al. (2015), which applies to general linear process.

- $\left\|\Omega^{\frac{1}{2}}A\Omega^{\frac{1}{2}}\right\|_F \leq \sqrt{rank(\Omega^{\frac{1}{2}}A\Omega^{\frac{1}{2}})}\left\|\Omega^{\frac{1}{2}}A\Omega^{\frac{1}{2}}\right\|_{op} \leq 2\pi\sqrt{rank(A)}\left\|A\right\|_{op}\mathcal{M}(f_X)$

The proof follows by putting these bounds in $\nu(\Omega^{\frac{1}{2}}A\Omega^{\frac{1}{2}}, \alpha, 2\pi\eta\mathcal{M}(f_X))$. □

Our next lemma is a generalization of Proposition 2.4 in Basu et al. (2015), to the case where the underlying process is characterized by the equations (S3.14) and (S3.15).

Lemma S3.2. *Consider a generic p-dimensional linear process in the form of* $X_t = \sum_{l=0}^{\infty} \Phi_l u_{t-l}$*, where the coordinates of* u_t *have* α*-sub-exponential tail decay as characterizd by equation (S3.15). Let* $\Sigma_X(0) = Cov(X_t, X_t)$*. Denote the realization of* X_t *by* $X \in \mathbb{R}^{n\times p}$ *and the sample covariance by* $S = \frac{1}{n}X'X$*. Then*

(i) For unit vectors v_1 *and* v_2 *satisfying* $\|v_1\| \le 1$, $\|v_2\| \le 1$*, the following bound holds:*

$$Pr\{|v_1'(S - \Sigma_X(0))v_1| > 2\pi\eta\mathcal{M}(f_X)\} \le \tau'(\eta, \alpha, n)$$

$$\text{and } Pr\{|v_1'(S - \Sigma_X(0))v_2| > 6\pi\eta\mathcal{M}(f_X)\} \le 2\tau'(\eta, \alpha, n)$$

(ii) Consider a q-dimensional linear process $Z_t = \sum_{l=0}^{\infty} \Psi_l w_{t-l}$*, where the coordinates of* w_t *have* α*-sub-exponential tail decay, as characterized by equation (S3.15). Also* $Cov(X_t, Z_t) = 0 \forall t$ *and the data matrix* $Z \in \mathbb{R}^{n\times q}$ *is similarly defined. Then the following bound holds:*

$$Pr\{|v_1'(X'Z)v_2| \ > 2\pi\eta(\mathcal{M}(f_X) + \mathcal{M}(f_Z) + \mathcal{M}(f_{X,Z}))\} \le 3\tau'(\eta, \alpha, n),$$

where $\mathcal{M}(f_{X,Z})$ *is defined the same way as in the main paper .*

Here τ' *is defined as* $\tau'(\eta, \alpha, n) = c_1 exp[-c_2 min\{n\eta^2, (n\eta)^{\frac{\alpha}{2}}\}]$*, for some constants* c_1 *and* c_2*.*

Proof. First we note that with $A = I_n$ and the definition of $\tau(\eta, \alpha, A)$ the following holds for some constant $C > 0$

$$\tau(n\eta, \alpha, A) = 2exp[-Cmin\{n\eta^2, (n\eta)^{\frac{\alpha}{2}}\}] \tag{S3.19}$$

Let $y_t = v_1' X_t$ and $Y = Xv_1 \in \mathbb{R}^n$ be n consecutive observations of the scalar process $\{y_t\}$. Then, we will have $v_1' S v_1 \overset{d}{=} \frac{1}{n} Y'Y$ and $v_1' \Sigma_X(0) v_1 = Exp[\frac{Y'Y}{n}]$. Applying Lemma S3.1 to the process $\{y_t\}$ with $A = I_n$ (since moment properties are preserved under linear transformation), we obtain the following:

$$\begin{aligned} Pr\{|v_1'(S - \Sigma_X(0))v_1| > 2\pi\eta\mathcal{M}(f_Y)\} &= Pr\{|Y'Y - Exp(Y'Y)| > 2\pi n\eta\mathcal{M}(f_Y)\} \\ &\leq \tau'(\eta, \alpha, n) \end{aligned} \tag{S3.20}$$

Further by Lemma C.6 of Sun et al. (2018), it follows that $\mathcal{M}(f_Y) \leq \|v_1\|^2 \mathcal{M}(f_X) = \mathcal{M}(f_X)$. Hence the following bound holds:

$$Pr\{|v_1'(S - \Sigma_X(0))v_1| > 2\pi\eta\mathcal{M}(f_X)\} \leq \tau'(\eta, \alpha, n) \tag{S3.21}$$

This proves the first part in (i). The rest of the proof follows along the similar lines to the derivation of Proposition 2.4 in Basu et al. (2015) and an outline is as follows:

For $|v_1'(S-\Sigma_X(0))v_2|$, one considers the following decomposition:

$$2|v_1'(S-\Sigma_X(0))v_2| \leq |v_1'(S-\Sigma_X(0))v_1| + |v_2'(S-\Sigma_X(0))v_2| + |(v_1+v_2)' \\ (S-\Sigma_X(0))(v_1+v_2)| \tag{S3.22}$$

with $\|(v_1+v_2)\| \leq 2$. Now repeating the steps as in (i) for each of the three components above yields the desired result.

For $|v_1'(X'Z)v_2|$, let $\tilde{y}_t = v_2'Z_t$ and thus $v_1'(X'Z)v_2 = \frac{1}{n}\sum_{t=1}^{n} y_t\tilde{y}_t$ and it satisfies the following decomposition:

$$\begin{aligned}\frac{2}{n}\sum_{t=1}^{n} y_t\tilde{y}_t &= [\frac{1}{n}\sum_{t=1}^{n}(y_t+\tilde{y}_t)^2 - Var(y_t+\tilde{y}_t)] - [\frac{1}{n}\sum_{t=1}^{n} {y_t}^2 - Var(y_t)] \\ &- [\frac{1}{n}\sum_{t=1}^{n} {\tilde{y}_t}^2 - Var(\tilde{y}_t)] \\ &= \frac{1}{n}[G'G - Exp(G'G)] - \frac{1}{n}[Y'Y - Exp(Y'Y)] \\ &- \frac{1}{n}[\tilde{Y}'\tilde{Y} - Exp(\tilde{Y}'\tilde{Y})],\end{aligned}$$

where $g_t = y_t + \tilde{y}_t$ is the summation process and G and $\tilde{Y}$ are defined analogously to the definition of Y. Now the proof of (ii) follows by repeating the same steps as in the proof of the second part of (i) and noting the fact that $\mathcal{M}(f_G) \leq \mathcal{M}(f_Z) + \mathcal{M}(f_X) + \mathcal{M}(f_{X,Z})$. □

Our next lemma can be considered as a generalization of the deviation bound derived in Basu et al. (2015).

Lemma S3.3. *There exist positive constants C, c_1 and c_2 such that the following deviation bound holds:*

$$\|X'E\|_\infty \leq C\mathbb{Q}\frac{(\log p + \log q)^{\frac{1}{\alpha}}}{\sqrt{n}} \tag{S3.23}$$

with probability at least $1 - c_1 exp\{-c_2(\log(pq))^{\frac{2}{\alpha}}\}$, for any random realizations $X \in \mathbb{R}^{n\times p}$ and $E \in \mathbb{R}^{n\times q}$, drawn from the p-dimensional linear processes $\{X_t\}$ and q-dimensional linear processes $\{\varepsilon_t\}$ respectively, where the coordinates of X_t and ε_t have α-sub-exponential tail decay and $\mathbb{Q} = \mathcal{M}(f_X) + \mathcal{M}(f_\varepsilon) + \mathcal{M}(f_{X,\varepsilon})$

Proof. The proof follows by applying part (ii) of Lemma S3.2 with $v_1 = e_i$ and $v_2 = e_j$, then taking union bound over all pq elements and finally choosing $\eta = c_0\mathbb{Q}\frac{(\log p+\log q)^{\frac{1}{\alpha}}}{\sqrt{n}}$ for some suitably chosen constant c_0. □

The following lemma verifies the Restricted Strong Convexity condition and thus can be considered as a generalization of Proposition 4.2 of Basu et al. (2015).

Lemma S3.4. *Consider a random realization $X \in \mathbb{R}^{n\times p}$ drawn from the p-dimensional linear process $X_t = \sum_{l=0}^{\infty} \Phi_l u_{t-l}$, where each coordinates of u_t has α-sub-exponential tail decay. Then RSC holds for X with parameter $\alpha_{RSC} = \pi m(f_X)$ and tolerance $\tau = c_0\alpha_{RSC}\frac{\log p}{n^{\frac{\alpha}{2}}}$ with probability at least $1 - c_1 exp(-c_2 n^{\frac{\alpha}{2}})$, where the definition of RSC and $m(f_X)$ are the same as* *defined in Basu et al. (2015).*

Proof. Let $S = \frac{1}{n}X'X$. First suppose that we have the following:

$$\frac{1}{2}v'Sv = \frac{1}{2}v'(\frac{X'X}{n})v \geq \frac{\alpha_{RSC}}{2}\left\|v\right\|_2^2 - \tau\left\|v\right\|_1^2, \forall v \in \mathbb{R}^p \tag{S3.24}$$

Then, for all $\Delta \in \mathbb{R}^{p\times p}$ and letting Δ_j denote the j^{th} column, the RSC condition automatically holds since

$$\begin{aligned}\frac{1}{2T}\left\|X\Delta\right\|_F^2 &= \frac{1}{2}\sum_{j=1}^{p}\Delta_j'(\frac{X'X}{n})\Delta_j \\ &\geq \frac{\alpha_{RSC}}{2}\sum_{j=1}^{p}\left\|\Delta_j\right\|_2^2 - \tau\sum_{j=1}^{p}\left\|\Delta_j\right\|_1^2 \\ &\geq \frac{\alpha_{RSC}}{2}\left\|\Delta\right\|_F^2 - \tau\left\|\Delta\right\|_1^2\end{aligned}$$

Therefore it suffices to verify that (S3.24) holds. Now applying the discretization argument as in Lemma F.2 and Lemma F.3 in Basu et al. (2015), define $\mathbb{K}(2s) = \{v \in \mathbb{R}^p, \|v\| \leq 1, \|v\|_0 \leq 2s\}$ and taking the union bound in this $2s$-sparse cone gives the following inequality:

$$\begin{aligned}&Pr\{\sup_{v\in\mathbb{K}(2s)}|v'(S-\Sigma_X(0))v| > 2\pi\mathcal{M}(f_X)\eta\} \\ \leq& 2\cdot min\{p^s, (21e\cdot\frac{p}{s})^s\}\tau'(\eta,\alpha,n) \\ =& 2c_1 exp[-c_2\ min\{n\eta^2, (n\eta)^{\frac{\alpha}{2}}\} + s\ min\{\log p, \log(21e\frac{p}{s})\}]\end{aligned}$$

Let $\eta = \frac{m(f_X)}{54\mathcal{M}(f_X)}$. Then applying the results from Lemma 12 in Loh et al.

(2012) with $\Gamma = S - \Sigma_X(0)$ and $\delta = \pi \frac{m(f_X)}{27}$ the following holds:

$$\frac{1}{2} v' S v \geq \frac{\alpha_{RSC}}{2} \|v\|^2 - \frac{\alpha_{RSC}}{2s} \|v\|_1^2, \text{where } \alpha_{RSC} = \pi m(f_X) \tag{S3.25}$$

with probability at least $1 - 2min\{p^s, (21e \cdot \frac{p}{s})^s\} \tau'(\eta, \alpha, n)$. By letting $s = c_0 \frac{n^{\frac{\alpha}{2}}}{\log p}$ for some small constant c_0, τ can be expressed as $\tau = c_0 \alpha_{RSC} \frac{\log p}{n^{\frac{\alpha}{2}}}$ and thus the bound holds with the probability as given in the statement. □

The following lemma can be considered as a generalization of Lemma H.1 in Negahban and Wainwright (2011) from the Gaussian case to $\alpha-$sub-exponential decay family.

Lemma S3.5. *Let $M \in \mathbb{R}^{m_1 \times m_2}$ be a random matrix where the entries follow $\alpha-$sub-exponential tail decay, as defined in the main paper. For every $\lambda \in (0, b)$ and for every $u \in \mathbb{R}^{m_1}$, $v \in \mathbb{R}^{m_2}$ with $\|u\|_2 = 1$ and $\|v\|_2 = 1$, we assume that, $\log Exp[e^{\lambda \{u' M v\}}] \leq \sup_{(u,v)} \psi_{u,v}(\lambda)$. Define $\psi^{*^{-1}}(y) = \inf_{\lambda \in (0,b)} \frac{y + \sup_{(u,v)} \psi_{u,v}(\lambda)}{\lambda}$. Then we have,*

$$Exp[\|M\|_{op}] \leq 2\psi^{*^{-1}}(m_1 + m_2) \tag{S3.26}$$

and

$$Pr\{\|M\|_{op} \geq Exp[\|M\|_{op}] + t\} \leq exp(-\frac{t^\alpha}{\rho}) \tag{S3.27}$$

where $\rho = \sup_{(u,v)} c_2(u, v)$ and $c_2(u, v)$ is the "c_2-parameter" corresponding to the variable $u' M v$, as in the definition of the sub-exponential tail decay.

Proof. The proof follows mostly along the same lines as of the proof of Lemma H.1 in Negahban and Wainwright (2011). The key technical difference is that, instead of applying standard bounds on Gaussian maxima, here we employ the bound provided by Theorem 2.5 of Boucheron et al. (2013) and arrive at the following inequality:

$$Exp[\|M\|_{op}] \leq 2{\psi^*}^{-1}(\log(M_1M_2)) = 2{\psi^*}^{-1}(\log(M_1) + \log(M_2)) \quad \text{(S3.28)}$$

where, M_1 and M_2 are such that $\{u^1, u^2, \cdots, u^{M_1}\}$ and $\{v^1, v^2, \cdots, v^{M_2}\}$ are $\frac{1}{4}$ coverings of S^{m_1-1} and S^{m_2-1} respectively. The definitions of S^{m_1-1} and S^{m_2-1} are same as in Negahban and Wainwright (2011). Finally, as in the proof of Lemma H.1 in Negahban and Wainwright (2011), there exist $\frac{1}{4}$ coverings of S^{m_1-1} and S^{m_2-1} with $\log(M_1) \leq m_1 \log 8$ and $\log(M_2) \leq m_2 \log 8$. Also $\psi^*(\cdot)$ is non-decreasing by Lemma 2.4 of Boucheron et al. (2013) and thus ${\psi^*}^{-1}(\cdot)$ is also non-decreasing by its definition. Hence we get $Exp[\|M\|_{op}] \leq 2{\psi^*}^{-1}(m_1 + m_2)$ and the proof is complete. □

As mentioned earlier, the above lemma can be considered as a generalization of the Lemma H.1 in Negahban and Wainwright (2011). In their case, they consider Gaussian random variables and thus $\psi_{u,v}(\lambda)$ takes the form $\frac{\lambda^2}{2}Var(u'Mv)$. Hence following their notation, $\underset{(u,v)}{sup}\ \psi_{u,v}(\lambda)$ boils down to $\frac{\lambda^2\rho^2(\Sigma)}{2}$ and finally from the definition of ${\psi^*}^{-1}$, it is easy to derive that

$Exp[\|M\|_{sp}] \leq c\rho(\Sigma)(\sqrt{m_1} + \sqrt{m_2})$.

S3.6 Proof of Theorem S6.1

Proof. We first start with finding a suitable choice for λ_S. By the definition of $\alpha-$sub-exponential tail decay, we get

$$Pr\{|U_{ij}| > a\} \leq c_1 \cdot exp(-\frac{a^\alpha}{c_2}) \tag{S3.29}$$

for some suitably chosen constant c_1 and c_2. Now by taking union bound over all the pmT entries and choosing $a = \text{const.}\{c_2 \log(pmT)\}^{\frac{1}{\alpha}}$ we get the following:

$$Pr\{\frac{\|U\|_\infty}{\sqrt{T}} > \frac{\text{const.}\{c_2 \log(pmT)\}^{\frac{1}{\alpha}}}{\sqrt{T}}\} \leq c_1 \cdot exp(-\text{const.} \log(pmT)) \tag{S3.30}$$

Hence we choose $\lambda_S = \frac{c_1^*\{c_2 \log(pmT)\}^{\frac{1}{\alpha}}}{\sqrt{T}} + \frac{4\alpha_1}{\sqrt{pmT}}$.

Next we find a suitable choice for λ_L. Applying Lemma S3.5 on matrix $Circ(\boldsymbol{\mathfrak{U}})$ we get,

$$Exp[\|Circ(\boldsymbol{\mathfrak{U}})\|_{sp}] \leq 2\psi^{*^{-1}}(mT + pT) \tag{S3.31}$$

and

$$Pr\{\|Circ(\boldsymbol{\mathfrak{U}})\|_{sp} \geq 2\psi^{*^{-1}}(mT + pT) + a\} \leq exp(-\frac{a^\alpha}{\rho}) \tag{S3.32}$$

Finally choosing $a = \rho^{\frac{1}{\alpha}} {\psi^*}^{-1}(mT + pT)$, we arrive at the following

$$Pr\{\frac{\|Circ(\mathfrak{U})\|_{sp}}{T} \geq \frac{c^* {\psi^*}^{-1}(mT + pT)}{T}\} \leq exp(-\{{\psi^*}^{-1}(mT + pT)\}^{\alpha}) \tag{S3.33}$$

for some suitably chosen constant c^*. Hence we choose $\lambda_L = \frac{c^* {\psi^*}^{-1}(mT+pT)}{T}$.

Now we obtain a suitable choice for λ_B. Note that, $\{Vec(X_t)\}$ can be considered as a linear process of the form $Vec(X_t) = \sum_{l=0}^{\infty} B_l \cdot w_{t-l}$, where each coordinate of w_t has the form as characterized by equation (S3.15) and B_l's are suitably adjusted matrices. Hence by applying Lemma S3.3 on the processes $\{Vec(X_t)\}$ and $\{Vec(U_t)\}$ we get,

$$Pr\{\frac{\|Z'U\|_{\infty}}{T} \geq \mathbb{Q}\frac{\{2\log(pm)\}^{\frac{1}{\alpha}}}{\sqrt{T}}\} \leq c_1 \cdot exp(-c_2\{2\log(pm)\}^{\frac{2}{\alpha}}) \tag{S3.34}$$

where $\mathbb{Q} = \mathcal{M}(f_X) + \mathcal{M}(f_U) + \mathcal{M}(f_{X,U})$. So we choose $\lambda_B = c_2^* \, \mathbb{Q}\frac{\{2\log(pm)\}^{\frac{1}{\alpha}}}{\sqrt{T}} + \frac{4\alpha_2}{\sqrt{pmT}}$. Finally, using Lemma S3.4, the assumption of Restricted Strong Convexity holds with high probability. Hence the proof follows using Lemma 3.1. □

S3.7 Proof of Lemma S2.1

Proof. The first part of the proof follows from the construction of the tensor $\mathfrak{G}$ and from the fact that tubal rank is same as the number of t-linearly independent lateral slices, as shown in Kilmer and Martin (2011) and Kilmer et al. (2013). The second part of the proof also follows from the construction of $\mathcal{G}$. □

S4 Matrix-type view of third-order tensor

We first start with understanding the matrix-type view of a third-order tensor. Three basic elements of interest are the lateral slices, horizontal slices and the tube fibers, defined rigorously under the section *Notations*. Recalling the definitions, lateral slices of $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ are d_2 laterally oriented matrices of dimension $d_1 \times d_3$. As mentioned in Kilmer et al. (2013), by staring at these laterally oriented matrices straight from the front, one will actually see them as column vectors of length d_1. Hence, the reader can envisage a three-dimensional tensor as a display of such lateral slices, placed side by side, playing the role of columns in a matrix. Similarly, the horizontal slices can be visualized as the row vectors of length d_2 and one can imagine that these slices play the roles of the rows of a matrix. An immediate question the reader may have is what exactly plays the role of an element? One can find the answer to this question in a similar fashion. By viewing the tube fibers of the tensor from the front, one would visualize them as the elements of a matrix. Figure 2 of the main paper aims to provide the reader a pictorial representation of this discussion. Note that, lateral and horizontal slices, although being matrices, can be considered as third-order tensors in $R^{d_1\times 1\times d_3}$ and $R^{1\times d_2\times d_3}$ respectively. Similarly, a tube fiber, although a vector, can be considered as a third-order tensor in $R^{1\times 1\times d_3}$. Kilmer et al. (2013) refer such elements in $R^{1\times 1\times d_3}$ as *Tubal Scalar*.

Given that the matrix-type view of a third-order tensor is now well understood, in order to proceed further, we first present the notion of *Block Circulant Matrices* and some related facts from Kilmer et al. (2013) and Kilmer and Martin (2011).

Notation S4.1. *For any vector* $\boldsymbol{a} = [a_0, a_1, a_2, a_3]^T$*, the Circulant Matrix associated with* $\boldsymbol{a}$*, denoted by Circ(*$\boldsymbol{a}$*), is defined as follows*

$$Circ(\boldsymbol{a}) = \begin{bmatrix} a_0 & a_3 & a_2 & a_1 \\ a_1 & a_0 & a_3 & a_2 \\ a_2 & a_1 & a_0 & a_3 \\ a_3 & a_2 & a_1 & a_0 \end{bmatrix}$$

Fact S4.1. *As discussed in Golub and Loan (1996), Circulant matrices can be diagonalized with the normalized Discrete Fourier Transform (DFT) matrix. In terms of commonly used notations, for any vector* $\boldsymbol{a}$ *of length* n*, let* F_n *denote the* $n \times n$ *DFT matrix and* F_n^* *denote its conjugate transpose.*

Then, F_n Circ($\boldsymbol{a}$) F_n^* is a diagonal matrix.

Fact S4.2. *diag(F_n Circ($\boldsymbol{a}$) F_n^*) = fft($\boldsymbol{a}$), Here fft($\boldsymbol{a}$) is the result of applying the Fast Fourier Transform to $\boldsymbol{a}$.*

The way circulant matrix is defined, in the same spirit, one can construct the *Block Circulant Matrix* using the frontal slices of a third-order tensor. In order to avoid complications, here we slightly modify our previous notation of frontal slices. The earlier notation $\boldsymbol{X_{::k}}$ for the k^{th} frontal slice is now simply replaced by $\boldsymbol{X_k}$.

Notation S4.2. *For any third-order tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$, let $\boldsymbol{A_1}$, $\boldsymbol{A_2}$, $\ldots$, $\boldsymbol{A_{d_3}}$ be the frontal slices. Then the Block Circulant matrix associated with $\boldsymbol{\mathcal{A}}$, denoted by Circ($\boldsymbol{\mathcal{A}}$), is the following matrix of order $d_1 d_3 \times d_2 d_3$*

$$Circ(\boldsymbol{\mathcal{A}}) = \begin{bmatrix} \boldsymbol{A_1} & \boldsymbol{A_{d_3}} & \boldsymbol{A_{d_3-1}} & \cdots & \boldsymbol{A_2} \\ \boldsymbol{A_2} & \boldsymbol{A_1} & \boldsymbol{A_{d_3}} & \cdots & \boldsymbol{A_3} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \boldsymbol{A_{d_3}} & \boldsymbol{A_{d_3-1}} & \ddots & \boldsymbol{A_2} & \boldsymbol{A_1} \end{bmatrix}$$

Fact S4.3. *Similar to Fact S4.1, a block circulant matrix can be block diagonalized. As before, suppose we have a DFT matrix F_{d_3} of order $d_3 \times d_3$*

and its conjugate transpose $F^*_{d_3}$. *Then the block-diagonalization is achieved as follows:*

$$(F_{d_3} \otimes I_{d_1}) \cdot Circ(\boldsymbol{\mathcal{A}}) \cdot (F_{d_3} \otimes I_{d_2}) = \begin{bmatrix} \boldsymbol{D_1} & & & \\ & \boldsymbol{D_2} & & \\ & & \ddots & \\ & & & \boldsymbol{D_{d_3}} \end{bmatrix}$$

Here $\otimes$ *represents the Kronecker product.*

Fact S4.4. *There is an alternative way to arrive at the above block diagonals. If one applies Fast Fourier Transformation along each tube of* $\boldsymbol{A}$ *and obtains a tensor* $\boldsymbol{\mathfrak{D}}$, *then the above block diagonals are actually the frontal slices of this newly obtained tensor* $\boldsymbol{\mathfrak{D}}$.

Notation S4.3. *MatVec operator arranges the frontal slices one below other and creates a matrix of order* $d_1 d_3 \times d_2$ *as follows*

$$MatVec(\boldsymbol{\mathcal{A}}) = \begin{bmatrix} \boldsymbol{A_1} \\ \boldsymbol{A_2} \\ \vdots \\ \boldsymbol{A_{d_3}} \end{bmatrix}$$

Notation S4.4. *fold operator converts MatVec($\boldsymbol{\mathcal{A}}$) back into the tensor $\boldsymbol{\mathcal{A}}$. Hence fold(MatVec($\boldsymbol{\mathcal{A}}$)) = $\boldsymbol{\mathcal{A}}$.*

Equipped with these notations, we are now in a position to describe the notion of *t-product* between two tensors. The idea of t-product was introduced in Kilmer et al. (2008) and some of its important theoretical properties were developed in Kilmer and Martin (2011)

Definition S4.1. *Given $\boldsymbol{\mathcal{A}} \in R^{d_1 \times d_2 \times d_3}$ and $\boldsymbol{\mathcal{B}} \in R^{d_2 \times l \times d_3}$ the t-product $\boldsymbol{\mathcal{A}} * \boldsymbol{\mathcal{B}}$ is defined to be a tensor $\boldsymbol{\mathcal{C}} \in R^{d_1 \times l \times d_3}$, where,*

$$\boldsymbol{\mathcal{C}} = fold\ (Circ\ (\boldsymbol{\mathcal{A}}) \cdot MatVec\ (\boldsymbol{\mathcal{B}}))$$

where $Circ(\cdot)$ and $MatVec(\cdot)$ are defined by Notation S4.2 and Notation S4.3 in Appendix S4.

Example S4.1. *Suppose $d_3 = 2$. Then the above definition expands as*

$$\boldsymbol{\mathcal{C}} = fold\left(\begin{bmatrix} \boldsymbol{A_1} & \boldsymbol{A_2} \\ \boldsymbol{A_2} & \boldsymbol{A_1} \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{B_1} \\ \boldsymbol{B_2} \end{bmatrix}\right)$$

Fact S4.5. *When a third-order tensor in $R^{d_1 \times d_2 \times d_3}$ is viewed as a $d_1 \times d_2$ matrix of tubes, then t-product between two tensors can be considered as* *matrix-matrix multiplication, with the exception that the operation between the scalars is now replaced by circular convolution between the tubes. Here, for any two vectors* $\boldsymbol{p}$ *and* $\boldsymbol{q}$*, circular convolution between them is defined as* $Circ(\boldsymbol{p}) \cdot q$

Fact S4.6. *t-product can be computed efficiently in three steps. First, apply FFT on* $\boldsymbol{\mathcal{A}}$ *and* $\boldsymbol{\mathfrak{B}}$ *along each tube and denote the resulting tensors as* $\tilde{\boldsymbol{\mathcal{A}}}$ *and* $\tilde{\boldsymbol{\mathfrak{B}}}$ *respectively. Then multiply each frontal slice of* $\tilde{\boldsymbol{\mathcal{A}}}$ *by the corresponding frontal slice of* $\tilde{\boldsymbol{\mathfrak{B}}}$*. Finally, apply inverse FFT along the tubes of the result.*

Next we discuss the notion of Identity tensor, inverse and transpose of a tensor and orthogonal tensor.

Definition S4.2. *The* $n \times n \times l$ *Identity Tensor, denoted by* $\boldsymbol{\mathfrak{I}}_{\boldsymbol{nnl}}$*, is defined to be a tensor, whose first frontal slice is a* $n \times n$ *identity matrix and all the other frontal slices are zeros.*

One can easily verify that $\boldsymbol{\mathcal{A}} * \boldsymbol{\mathfrak{I}} = \boldsymbol{\mathcal{A}} = \boldsymbol{\mathfrak{I}} * \boldsymbol{\mathcal{A}}$

Definition S4.3. $\boldsymbol{\mathcal{A}} \in R^{n \times n \times l}$ *is said to have an inverse* $\boldsymbol{\mathfrak{B}}$*, if* $\boldsymbol{\mathcal{A}} * \boldsymbol{\mathfrak{B}} = \boldsymbol{\mathfrak{I}}$ $= \boldsymbol{\mathfrak{B}} * \boldsymbol{\mathcal{A}}$

Definition S4.4. *Transpose of* $\boldsymbol{\mathcal{A}} \in R^{d_1 \times d_2 \times d_3}$*, denoted by* $\boldsymbol{\mathcal{A}}^T$*, is the* $d_2 \times d_1 \times d_3$ *tensor obtained by transposing each of the frontal slices and then reversing the order of the transposed frontal slices* 2 *through* d_3*.*

Example S4.2. *Suppose $d_3 = 4$. Then from the above definition,*

$$\mathcal{A}^T = fold\left(\begin{bmatrix} \boldsymbol{A_1^T} \\ \boldsymbol{A_4^T} \\ \boldsymbol{A_3^T} \\ \boldsymbol{A_2^T} \end{bmatrix}\right)$$

Definition S4.5. $\mathfrak{Q} \in R^{n\times n\times l}$ *is said to be orthogonal tensor if* $\mathfrak{Q}^T * \mathfrak{Q} = \mathfrak{Q} * \mathfrak{Q}^T = \mathfrak{I}$

Definition S4.6. *The collection of lateral slices* $\boldsymbol{Q_{:1:}}$, $\boldsymbol{Q_{:2:}}$, $\cdots$,$\boldsymbol{Q_{:n:}}$ *of* $\mathfrak{Q}$ *is said to form an orthogonal set if*

$$\boldsymbol{Q}^T_{\boldsymbol{:i:}} * \boldsymbol{Q_{:j:}} = \begin{cases} \alpha_i \boldsymbol{e1}, & \textit{if } i = j \\ 0, & \textit{if } i \neq j \end{cases}$$

where α_i is a nonzero scalar. The set is orthonormal if $\alpha_i = 1$.

Fact S4.7. $\mathfrak{Q} \in R^{n\times n\times l}$ *is orthogonal tensor iff the collection of the lateral slices* $\{\boldsymbol{Q_{:1:}}, \boldsymbol{Q_{:2:}}, \cdots, \boldsymbol{Q_{:n:}}\}$ *forms an orthonormal set.*

Suppose an orthogonal set of elements in $R^{m\times 1\times l}$ contains m elements. Comparing this framework to usual matrix algebra, it would be of great use , if one could reconstruct any element in $R^{m\times 1\times l}$ from those m elements. As discussed in Kilmer et al. (2013), one could achieve this by extending the concept of usual linear combination to t-linear combination, where, lateral slices act as columns and tubal scalars play the role of scalars.

Definition S4.7. *Given d_2 lateral slices, $\boldsymbol{X}_{:1:}$, $\boldsymbol{X}_{:2:}$, $\cdots$, $\boldsymbol{X}_{:d_2:}$ of $\mathcal{X} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ and d_2 tubal scalars $\boldsymbol{c_1}$, $\boldsymbol{c_2}$, ..., $\boldsymbol{c_{d_2}}$, the t-linear combination of the lateral slices is defined as $\boldsymbol{X}_{:1:} * \boldsymbol{c_1} + \boldsymbol{X}_{:2:} * \boldsymbol{c_2} + \ldots + \boldsymbol{X}_{:d_2:} * \boldsymbol{c_{d_2}}$, where, the tubal-scalars are the elements in $\mathbb{R}^{1 \times 1 \times d_3}$ and $*$ denote the t-product defined above.*

Employing the definition of t-linear combination, one can now define the range of the tensor $\boldsymbol{\mathcal{A}} \in R^{d_1 \times d_2 \times d_3}$, denoted by $\boldsymbol{R}(\boldsymbol{\mathcal{A}})$, as the set of all possible t-linear combinations of its lateral slices. Similarly, extending the notion of the usual linear dependence of two columns, one can say that the lateral slice $\boldsymbol{A}_{:j_2:}$ is t-linearly dependent on the lateral slice $\boldsymbol{A}_{:j_1:}$, if there exists a tubal scalar $\boldsymbol{c}$, such that, $\boldsymbol{A}_{:j_2:} = \boldsymbol{A}_{:j_1:} * \boldsymbol{c}$. Figure 3 in the main paper furnishes further clarity of this idea by demonstrating a simple example. keeping this framework in mind, it would be very useful if one would know the minimum number of elements in $R^{d_1 \times 1 \times d_3}$, that is required to construct any arbitrary element in $\boldsymbol{R}(\boldsymbol{\mathcal{A}})$. As described in Kilmer et al. (2013), this number is characterized by *Tubal Rank*, which is the last topic of our discussion under this section. Before moving on to that discussion, we need to describe one more notation.

Notation S4.5. *An f-diagonal tensor, denoted by* $\boldsymbol{\mathfrak{F}}$*, is a third-order tensor, whose each frontal slice is a diagonal matrix. In terms of notation,* $\mathfrak{F}_{ijk} = 0$*, for* $i \neq j, \forall k$*.*

Definition S4.8. *For any* $\boldsymbol{\mathcal{A}} \in R^{d_1 \times d_2 \times d_3}$*, t-SVD of* $\boldsymbol{\mathcal{A}}$ *is given as follows:*

$$\boldsymbol{\mathcal{A}} = \boldsymbol{\mathfrak{U}} * \boldsymbol{\mathfrak{S}} * \boldsymbol{\mathfrak{V}}^{\boldsymbol{T}}$$

Here $\boldsymbol{\mathfrak{U}}$ *and* $\boldsymbol{\mathfrak{V}}$ *are orthogonal tensors in* $R^{d_1 \times d_1 \times d_3}$ *and* $R^{d_2 \times d_2 \times d_3}$ *respectively.* $\boldsymbol{\mathfrak{S}}$ *is a f-diagonal tensor in* $R^{d_1 \times d_2 \times d_3}$*.*

Definition S4.9. *For any third-order tensor, Tubal-rank, denoted by r, is defined to be the number of non zero tubes in the f-diagonal tensor* $\boldsymbol{\mathfrak{S}}$ *in its t-SVD factorization. Hence,* $r = \#\ \{i\text{: } \boldsymbol{s_{ii:}} \neq 0\}$*, where* $\boldsymbol{s_{ii:}}$ *denote the* i^{th} *diagonal tube of* $\boldsymbol{\mathfrak{S}}$*.*

Like matrix singular value decomposition, in this case too, $\boldsymbol{R(\mathcal{A})}$ can be written unambiguously by the lateral slices of $\boldsymbol{\mathfrak{U}}$. Also, the number of elements in $R^{d_1 \times 1 \times d_3}$, required to construct any element in $\boldsymbol{R(\mathcal{A})}$, is same as the tubal rank of $\boldsymbol{\mathcal{A}}$. The reader may visit Kilmer et al. (2013) for the proofs and further details. Hence, as rank of a matrix decides the number of linearly independent columns of a matrix, tubal rank plays similar role in case of a third order tensor. Indeed, lower the value of tubal rank, higher the number of t-linearly dependent lateral slices. Figure S8.1 displays the tubal ranks and the lateral slices of three different tensors. In the first case, only the first slice from the left is t-linearly independent. Both the remaining slices are t-linear combination of the first one. Hence the tubal rank in this case is one. Similar justification follows for the other two cases too.

It is possible to compute t-SVD by performing matrix SVD d_3 times in the Fourier domain. The reader may see Kilmer and Martin (2011) for more details. However, Lu et al. (2018) recently proposed a more efficient algorithm for computing t-SVD. This algorithm requires one to perform matrix SVD only $\lceil \frac{d_3+1}{2} \rceil$ times, instead of d_3 times. Lu et al. (2018) defines the elements of the first frontal slice of the f-diagonal tensor $\boldsymbol{\mathcal{S}}$, that is $\boldsymbol{S_{::1}}$, as the *Singular values* of the tensor $\boldsymbol{\mathcal{A}}$ and argues that, the number of non-zero singular values is equivalent to the tubal-rank defined in S4.9. In terms of the notations used here, r = # $\{i\colon \boldsymbol{s_{ii:}} \neq 0\}$ = # $\{i\colon \mathcal{S}_{ii1} \neq 0\}$. So, by penalizing high value of $\sum_{i=1}^{r} \mathcal{S}_{ii1}$, one can actually restrict the value of the tubal-rank to an upper bound. In Lu et al. (2018), the quantity $\sum_{i=1}^{r} \mathcal{S}_{ii1}$ is defined as *Tensor Nuclear Norm* of the tensor $\boldsymbol{\mathcal{A}}$. Just as a side note, this definition of Tensor Nuclear Norm is slightly different from the one in Zhang et al. (2014). However using Definition 7 of Lu et al. (2018) and equation 12 of Lu et al. (2016), one can derive the following relationship between Tensor Nuclear Norm and Block Circulant matrix.

$$\sum_{i=1}^{r} \mathcal{S}_{ii1} = \frac{1}{d_3} \left\| Circ(\boldsymbol{\mathcal{A}}) \right\|_* \tag{S4.1}$$

It is evident that, in order to restrict the tubal rank of a tensor, one can impose penalty on the right hand side of the above equation. We utilize this fact while we discuss the convex relaxation of our proposed model in Section 2.1.

S5 Details of Alternating Minimization Algorithm

Step 1: Step 1 updates the value of L given S and B. For a given value of S and B, letting $U = Y - S - ZB^T$, the problem then reduces to minimizing $g(L) = \frac{1}{2} \left\| U - L \right\|_F^2 + \lambda_L \frac{1}{T} \left\| Circ(\boldsymbol{\mathcal{L}}) \right\|_*$ with respect to L, where $\boldsymbol{\mathcal{L}} \in \mathbb{R}^{p \times m \times T}$ is the tensor form of $L \in \mathbb{R}^{T \times pm}$. Now we create a tensor $\boldsymbol{\mathcal{U}} \in \mathbb{R}^{p \times m \times T}$, whose t^{th} frontal slice is essentially the matrix version of the t^{th} row of U. Given these notations and noting the fact that $\left\| Circ(\boldsymbol{\mathcal{U}}) - Circ(\boldsymbol{\mathcal{L}}) \right\|_F^2 = T \times \left\| U - L \right\|_F^2$ and denoting $Circ(\boldsymbol{\mathcal{L}})$ by W, one can rewrite the aforementioned problem as minimizing $g(W) = \frac{1}{2} \left\| W - Circ(\boldsymbol{\mathcal{U}}) \right\|_F^2 + \lambda_L \left\| W \right\|_*$ with respect to W. As discussed in Cai et al. (2010), the solution to this minimization problem involves two simple steps: first obtaining the *Singular*

Value Decomposition of $Circ(\mathcal{U})$ and then applying soft-thresholding on the singular values. This technique of *Singular Value Thresholding* shows up in various machine learning problems, including matrix classification, multi-task learning and matrix completion. Ji and Ye (2009) formalize this problem in their Theorem 3.1, which we apply directly to obtain the optimal solution W and thus eventually, the optimal L.

Step 2: In Step 2, we update the value of S given L and B. As in Step 1, for given values of L and B, we let $V = Y - L - ZB^T$ and the problem then boils down to minimizing $g(S) = \frac{1}{2}\|S - V\|_F^2 + \lambda_S \|S\|_1$ with respect to S. Thus the problem reduces to *Soft Thresholding* (see Friedman et al. (2001)), which is solved by shrinking the elements of V towards zero by the factor λ_S.

Step 3: Finally, in Step 3, we update the value of B given L and S. For a given value of L and S, letting $G = Y - L - S$, the problem then reduces to minimizing $g(B) = \frac{1}{2}\left\|G - ZB^T\right\|_F^2 + \lambda_B \left\|B^T\right\|_1$ with respect to B. This is a typical setting of penalized multivariate regression and we directly employ the algorithm used in Lin et al. (2016) as it has better computational efficiency and desired convergence properties.

S6 Theoretical results under Sub-exponential distribution

Here we extend the above result to the case when the signals and the errors will have *α-sub-exponential tail decay*. As defined in Götze et al. (2019), a random variable E is said to have α-sub-exponential tail decay if the following holds: $Pr\{|E| > t\} \leq c_1 \cdot exp(-\frac{t^\alpha}{c_2})$, for some constants c_1 and c_2, where the parameter $\alpha \in (0, 1] \cup \{2\}$. The above definition covers a variety of distribution depending on the chosen value of α. Examples include, *sub-Gaussian* distribution, *sub-Exponential* distribution such as Poisson or Weibull random variables and so on (see Götze et al. (2019)).

In our case, we have two white noise processes $Vec(F_t)$ and $Vec(U_t)$ and given the above discussion, we assume that each coordinate of $Vec(F_t)$ and $Vec(U_t)$ has α-sub-exponential tail decay, as defined above. In addition to that, we also assume $Cov(Vec(U_t), Vec(F_{t'})) = 0 \forall (t, t')$ as in Section 2 of the main paper and in the earlier result under Gaussian assumption. Finally, for every $\lambda \in (0, b)$ and for every $u \in \mathbb{R}^{pT}$, $v \in \mathbb{R}^{mT}$ with $\|u\|_2 = 1$ and $\|v\|_2 = 1$, we assume that, $\log Exp[e^{\lambda\{u' Circ(\mathfrak{U}) v\}}] \leq \sup_{(u,v)} \psi_{u,v}(\lambda)$. This assumption has been used to prove Theorem 2.5 of Boucheron et al. (2013), which we will use while proving our next result. As in Lemma 2.4 of

Boucheron et al. (2013), we define $\psi^{*^{-1}}(y) = \underset{\lambda \in (0,b)}{inf} \frac{y + \underset{(u,v)}{sup}\, \psi_{u,v}(\lambda)}{\lambda}$. Given these distributional assumptions and definitions, we now state our next result.

Theorem S6.1. *Suppose the coordinates of the signals and the errors follow α-sub-exponential tail decay with the assumptions discussed above. Then with probability greater than $1 - c_1^* exp\{-c_2^*(2\log(pm))^{\frac{2}{\alpha}}\}$ we will have,*

$$e^2(\hat{L}, \hat{S}, \hat{B}) \leq \hat{c_1}\ r\ \{\frac{\psi^{*^{-1}}(mT + pT)}{T}\}^2 + \hat{c_2}\ [\frac{s_1\{c_2\ \log(pmT)\}^{\frac{2}{\alpha}}}{T} + \frac{\alpha_1^2 s_1}{pmT}] + \hat{c_3}\ [\mathbb{Q}^2 \frac{s_2\ \{2\log(pm)\}^{\frac{2}{\alpha}}}{T} + \frac{\alpha_2^2 s_2}{pmT}] \tag{S6.1}$$

where, $c_1^*, c_2^*, \hat{c_1}, \hat{c_2}, \hat{c_3}$ are the suitable constants, c_2 is the "c_2 parameter" in the definition of α-sub exponential tail decay for the coordinates of the errors and $\mathbb{Q}$ has similar definition as of $\mathbb{Q}(B^*, \sigma^2, \Sigma_F)$ in Theorem 3.2 of the main paper.

It is worth mentioning that, in the Gaussian case, $\alpha = 2$ and $\underset{(u,v)}{sup}\ \psi_{u,v}(\lambda) = \frac{\lambda^2\ \rho^2(\Sigma_{Circ})}{2}$, where $\rho^2(\Sigma_{Circ}) = \underset{(u,v)}{sup}\ Var(u^T Circ(\mathbf{\mathcal{U}})v)$, as defined in Lemma H.1 of Negahban and Wainwright (2011). Hence recalling the definition of $\psi^{*^{-1}}(y)$, it is easy to show that $\psi^{*^{-1}}(mT + pT)$ attains the infimum with respect to λ when $\lambda = \frac{\sqrt{2(mT+pT)}}{\rho(\Sigma_{Circ})}$. Putting this optimal λ, some easy algebraic steps show that the first term $r\{\frac{\psi^{*^{-1}}(mT+pT)}{T}\}^2$ in equation (S6.1) boils down to $\sigma^2 \frac{r(p+m)}{T}$, which is the first term of equation (3.3) of the main paper. For $\alpha = 2$, the remaining two terms of Equation S6.1 are also in line with those of equation (3.3) of the main paper. Thus the bound in Theorem S6.1 is justified.

S7 AIC

As mentioned in the simulation studies of the main paper, while working with real data, the true rank and sparsity levels are unknown. In such situations, we choose the values of λ_L, λ_S and λ_B in such a way that the AIC, as defined below, is minimized.

AIC: We define AIC as $T\log(\frac{RSS}{T}) + 2$ Rank $(Circ(\hat{\boldsymbol{\mathcal{L}}})) + 2k_1 + 2k_2$, where, RSS is defined as $\left\|Y - \hat{L} - \hat{S} - Z\hat{B}^T\right\|_F^2$ (see equation (2.5) in the main paper) and k_1 and k_2 are the number of non-zero elements in $\hat{\boldsymbol{\mathcal{S}}}$ and $\hat{B}$ respectively. This formulation is quite common in literature, which essentially rewards goodness of fit and at the same time penalizes overfitting.

S8 Addtional Tables and Figures

		Sub-case 1		Sub-case 2		Sub-case 3	
Sample Size	Criteria	NZ = 2	NZ = 4	ED = 0.04	ED = 0.06	r = 2	r = 3
T=60	*Tubal-Rank*	4	4	4	4	1	4
	RE	0.52	0.64	0.52	0.57	0.35	0.52
	SN_S	0.83	0.81	0.83	0.82	0.85	0.83
	SP _S	0.80	0.80	0.80	0.80	0.88	0.80
	SN_A	0.81	0.80	0.81	0.80	0.91	0.81
	SP _A	0.83	0.82	0.83	0.81	0.85	0.83
T=80	*Tubal-Rank*	4	4	4	4	2	4
	RE	0.30	0.37	0.30	0.44	0.27	0.30
	SN_S	0.90	0.87	0.90	0.86	0.90	0.90
	SP _S	0.85	0.80	0.85	0.81	0.91	0.85
	SN_A	0.91	0.90	0.91	0.90	0.95	0.91
	SP _A	0.90	0.88	0.90	0.89	0.90	0.90
T=120	*Tubal-Rank*	3	3	3	3	2	3
	RE	0.16	0.18	0.16	0.18	0.10	0.16
	SN_S	0.91	0.90	0.91	0.91	0.92	0.91
	SP _S	0.93	0.92	0.93	0.90	0.95	0.93
	SN_A	0.95	0.94	0.95	0.93	0.98	0.95
	SP _A	0.91	0.90	0.91	0.90	0.93	0.91

Table S8.1: Scenario 1: $p = 10, m = 10$

		Sub-case 1		Sub-case 2		Sub-case 3	
Sample Size	Criteria	NZ = 2	NZ = 4	ED = 0.04	ED = 0.06	r = 2	r = 3
T=80	*Tubal-Rank*	4	4	4	4	1	4
	RE	0.55	0.68	0.54	0.58	0.40	0.55
	SN_S	0.81	0.80	0.82	0.82	0.84	0.84
	SP _S	0.80	0.81	0.81	0.80	0.87	0.80
	SN_A	0.81	0.81	0.80	0.81	0.92	0.82
	SP _A	0.83	0.83	0.83	0.82	0.84	0.83
T=120	*Tubal-Rank*	4	4	4	4	2	4
	RE	0.35	0.40	0.33	0.46	0.31	0.36
	SN_S	0.89	0.86	0.90	0.86	0.90	0.90
	SP _S	0.85	0.80	0.86	0.81	0.92	0.87
	SN_A	0.92	0.91	0.90	0.90	0.96	0.91
	SP _A	0.91	0.89	0.92	0.89	0.90	0.91
T=140	*Tubal-Rank*	3	3	3	3	2	3
	RE	0.19	0.23	0.20	0.22	0.15	0.18
	SN_S	0.92	0.91	0.90	0.90	0.92	0.92
	SP _S	0.93	0.93	0.92	0.90	0.94	0.94
	SN_A	0.94	0.94	0.95	0.93	0.97	0.95
	SP _A	0.91	0.91	0.90	0.90	0.92	0.92

Table S8.2: Scenario 2: $p = 20, m = 20$. In Sub-case 1, $r = 3$ and edge-density of $B^* = 0.04$, but we vary the number of non-zero elements in $\mathcal{S}^*$ from 2 to 4 per slice. In Sub-case 2, we fix $r = 3$ and the number of non-zero elements in each slice of $\mathcal{S}^* = 2$, but vary the edge-density of B^* from 0.04 to 0.06. Finally, in Sub-case 3, 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 $\mathcal{L}^*$ from 2 to 3

Variable	Abbreviation	Source	Transformation
Interest Rate of Long-Term Government Bond Yields	GOV. BOND	EUROSTAT	Δ
Consumer Price Index: All Items	CPI	IMF	$\Delta^2 \ln$
Producer Price Index: All Commodities	PPI	IMF	$\Delta^2 \ln$
Total Share Prices for All Shares	Tot_Share	FRED	$\Delta^2 \ln$
Final Consumption Expenditure	Cons_Exp	IMF	$\Delta \ln$
Capacity Utilization	Cap_Util	FRED	Δ
All Employees	Empl	FRED	$\Delta \ln$
Civilian Unemployment Rate	Un_Rate	FRED	Δ
Compensation of Employees	Comp	IMF	$\Delta \ln$
National Income	Nat_Income	IMF	$\Delta \ln$
Effective Exchange Rate (based on Unit-Labor-Cost)	EER	IMF	Δ
Industrial Production Index	IPI	IMF	Δ
Total Reserves	Tot_Res	IMF	$\Delta^2 \ln$
External Balance of Goods and Services	BGS	IMF	$\Delta \ln$
Broad Money Liabilities	M_2	IMF	$\Delta^2 \ln$
Gross Domestic Product deflator	GDP	IMF	$\Delta^2 \ln$

Table S8.3: Details of the Macroeconomic Variables

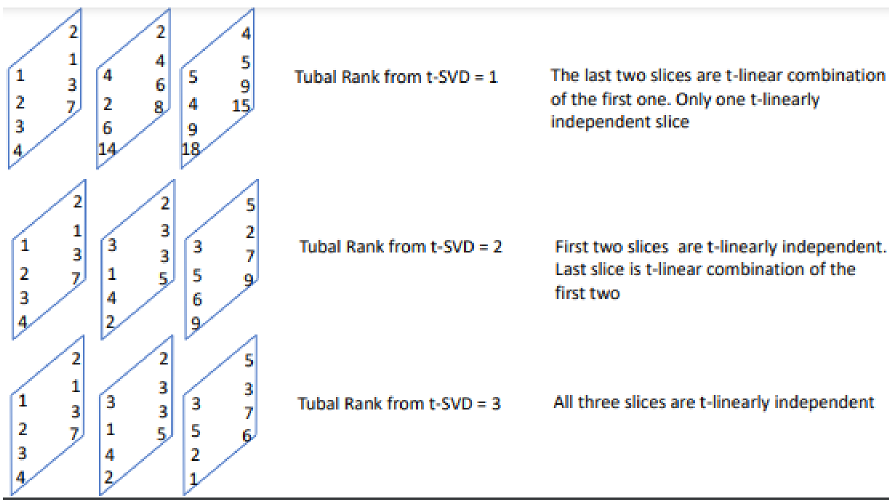

Figure S8.1: Tubal Rank and t-linear dependence

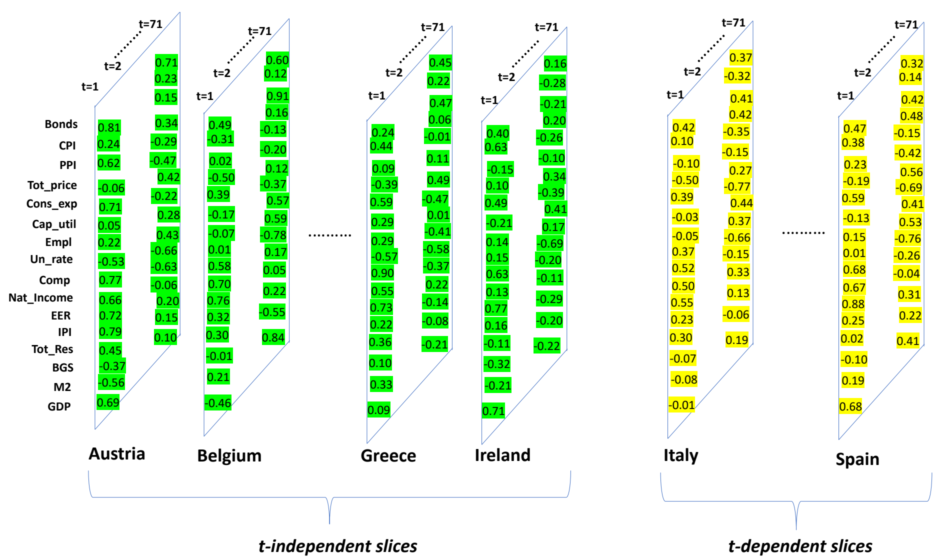

Figure S8.2: Partial Screenshot of Estimated Low Tubal-Rank Component $\hat{\mathcal{L}}$. Estimated Tubal-Rank is 7. Slices marked in green are t-linearly independent slices (total 7) and the remaining ones that are marked in yellow are t-linearly dependent on the previous set

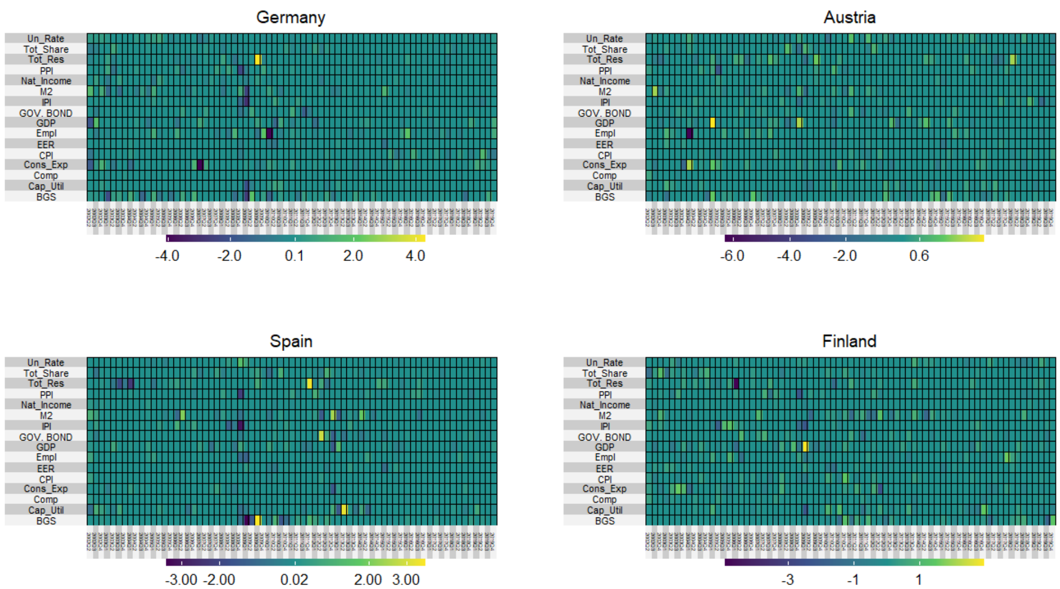

Figure S8.3: Heatmaps of the slices (Germany, Spain, Austria and Finland) of estimated sparse component $\hat{\mathcal{S}}$

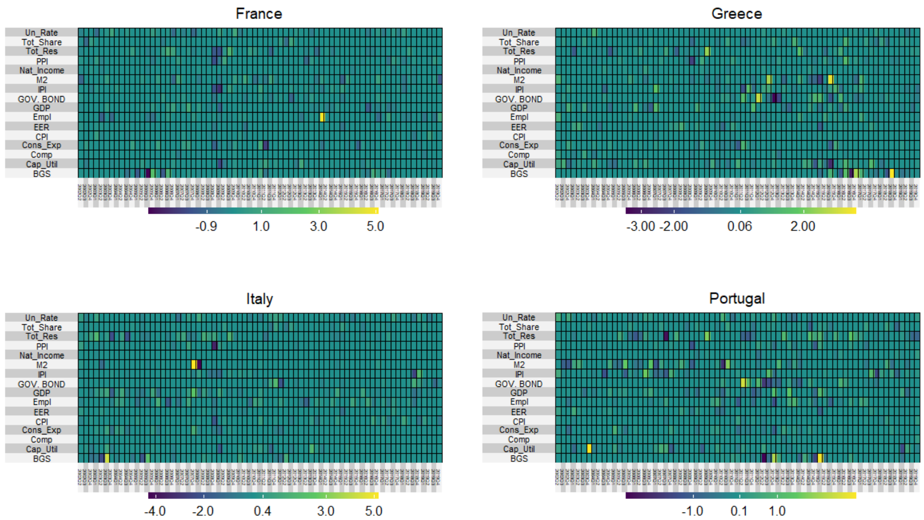

Figure S8.4: Heatmaps of the slices (France, Italy, Greece and Portugal) of estimated sparse component $\hat{\mathcal{S}}$

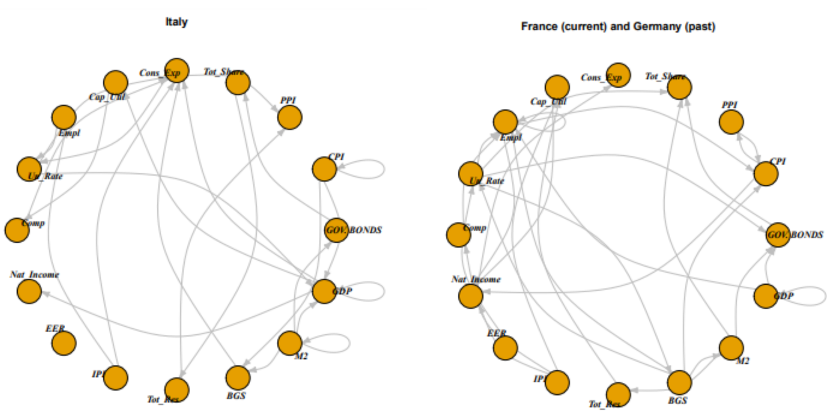

Figure S8.5: Intra-country temporal connectivity for Italy and cross-country temporal connectivity for France (current) and Germany (past)

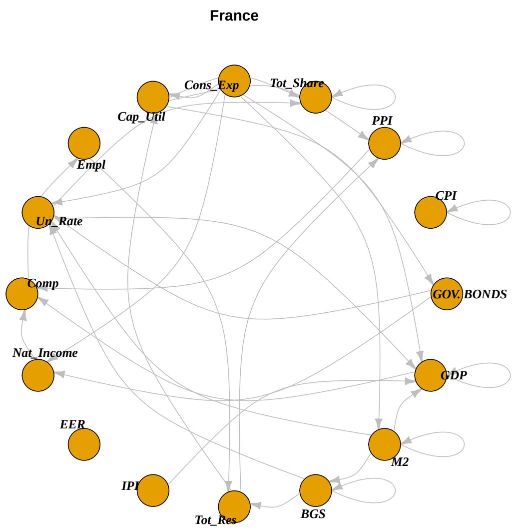

Figure S8.6: Intra-country temporal connectivity for France

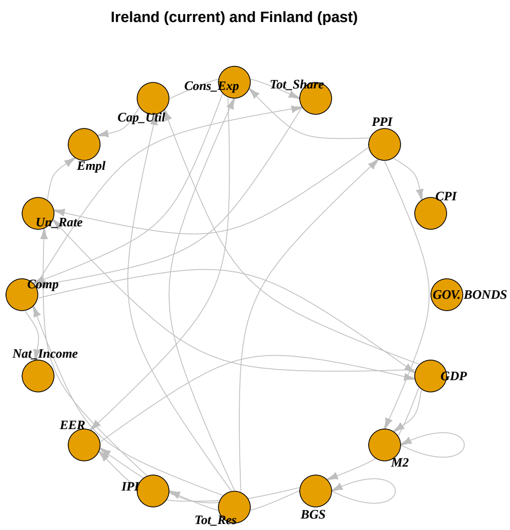

Figure S8.7: Cross-country temporal connectivity for Ireland (current) and Finland (past)

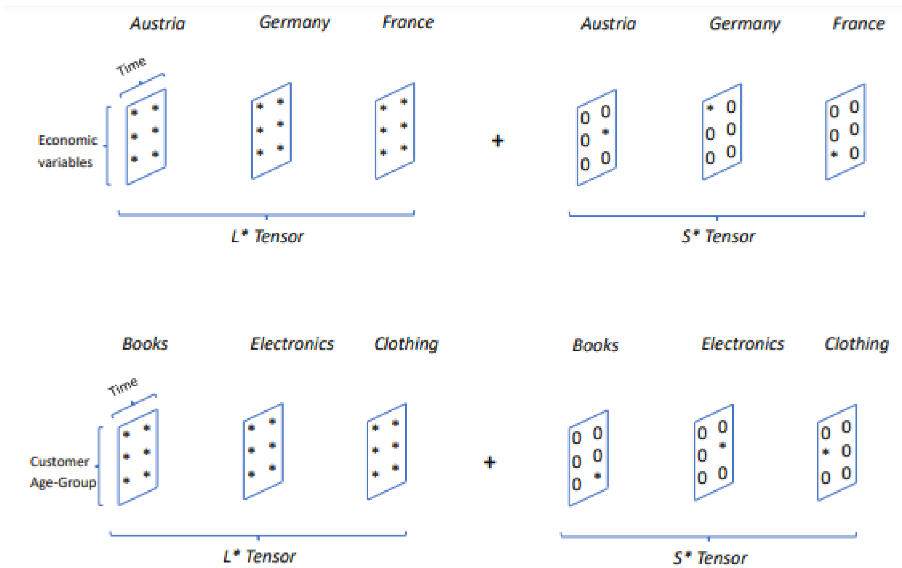

Figure S8.8: Low tubal-rank and sparse components: Lateral Slices correspond to different countries in the first example and different product-categories in the second example

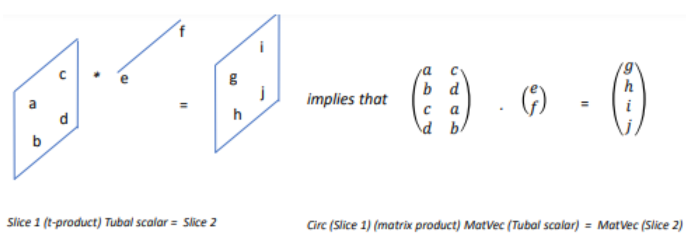

Figure S8.9: t-linear dependence between two lateral slices

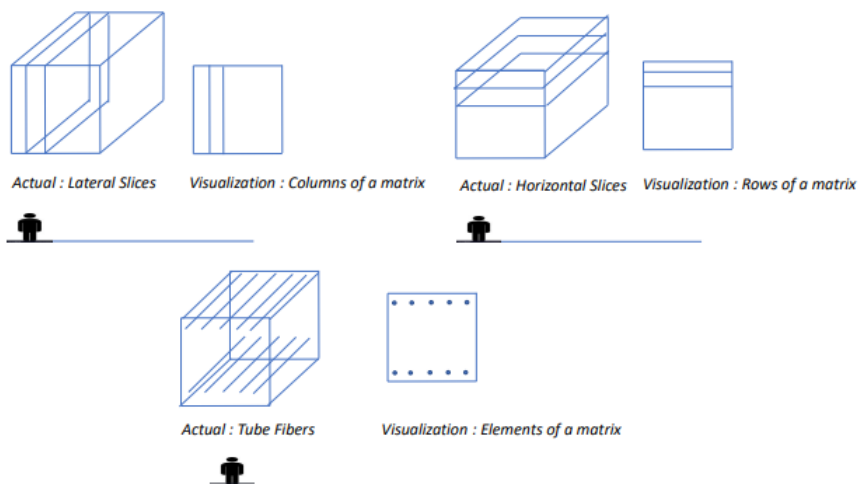

Figure S8.10: Matrix-type view of a third-order tensor: Lateral Slices, Horizontal Slices and Tube Fibers can be visualized as columns, rows and elements respectively

Bibliography

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.

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

Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Cai, J.-F., E. J. Candès, and Z. Shen (2010). A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization 20*(4), 1956–1982.

Friedman, J., T. Hastie, and R. Tibshirani (2001). *The elements of statistical learning*, Volume 1. Springer series in statistics New York.

Golub, G. H. and C. F. V. Loan (1996). *Matrix Computations (Johns Hopkins Studies in Mathematical Sciences)(3rd Edition)*. Johns Hopkins University Press.

Götze, F., H. Sambale, and A. Sinulis (2019). Concentration inequalities for polynomials in \α-sub-exponential random variables. *arXiv preprint arXiv:1903.05964*.

Ji, S. and J. Ye (2009). An accelerated gradient method for trace norm minimization. In *Proceedings of the 26th annual international conference on machine learning*, pp. 457–464.

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.

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

Kilmer, M. E., C. D. Martin, and L. Perrone (2008). A third-order generalization of the matrix svd as a product of third-order tensors. *Tufts University, Department of Computer Science, Tech. Rep. TR-2008-4*.

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

Lin, J., S. Basu, M. Banerjee, and G. Michailidis (2016). Penalized maximum likelihood estimation of multi-layered gaussian graphical models. *The Journal of Machine Learning Research 17*(1), 5097–5147.

Loh, P.-L., M. J. Wainwright, et al. (2012). High-dimensional regression with 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). Exact low tubal rank tensor recovery from gaussian measurements. *arXiv preprint arXiv:1806.02511*.

Negahban, S. and M. J. Wainwright (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 1069–1097.

Negahban, S. N., P. Ravikumar, M. J. Wainwright, B. Yu, et al. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science 27*(4), 538–557.

Recht, B., M. Fazel, and P. A. Parrilo (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review 52*(3), 471–501.

Zhang, Z., G. Ely, S. Aeron, N. Hao, and M. Kilmer (2014). Novel methods for multilinear data completion and de-noising based on tensor-svd. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3842–3849.