Electronic Journal of Statistics
Vol. 16 (2022) 2683–2723
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS2004

Regularized high dimension low tubal-rank tensor regression

Samrat Roy and George Michailidis

Department of Statistics and the Informatics Institute, University of Florida
e-mail: samratroy@ufl.edu; gmichail@ufl.edu

Abstract: Tensor regression models are of emerging interest in diverse fields of social and behavioral sciences, including neuroimaging analysis, neural networks, image processing and so on. Recent theoretical advancements of tensor decomposition have facilitated significant development of various tensor regression models. The focus of most of the available literature has been on the Canonical Polyadic (CP) decomposition and its variants for the regression coefficient tensor. A CP decomposed coefficient tensor enables estimation with relatively small sample size, but it may not always capture the underlying complex structure in the data. In this work, we leverage the recently developed concept of tubal rank and develop a tensor regression model, wherein the coefficient tensor is decomposed into two components: a low tubal rank tensor and a structured sparse one. We first address the issue of identifiability of the two components comprising the coefficient tensor and subsequently develop a fast and scalable Alternating Minimization algorithm to solve the convex regularized program. Further, we provide finite sample error bounds under high dimensional scaling for the model parameters. The performance of the model is assessed on synthetic data and is also used in an application involving data from an intelligent tutoring platform.

MSC2020 subject classifications: Primary 62J05, 62J07; secondary 15A72.
Keywords and phrases: Tensor, tensor regression, CP decomposition, tubal-rank, alternating minimization.

Received May 2021.

Contents

1. Introduction

There is increased interest in tensor analysis due to both technical developments and novel applications - see [16, 7] and references therein. There has been extensive work in the literature on tensor decomposition, including the classical Tucker ([36]), Canonical Polyadic (CP) ([5]) and higher-order Singular Value Decomposition (HOSVD) ([8]) and more recently, the tensor train (TT) ([29]) and tensor SVD ones (t-SVD) ([13]).

In many applications, the interest is on building regression models based on tensor data. Examples include neuroscience and neuroimaging applications ([42, 22, 34]), applications to other image processing tasks ([31]), neural networks ([17]) and analysis of longitudinal and spatio-temporal data ([11, 39]).

A typical setting for tensor regression is as follows: one has access to predictors whose structure can be succinctly represented by a tensor and an outcome variable of interest. For example, [22] consider Magnetic Resonance Imaging technology three-dimensional scans from subjects in an Alzheimer's Disease study together with their corresponding mini-mental state exam scores. A regression model to associate the information contained in the scans and the outcome scores takes the form

$$y = \langle \mathfrak{B}, \mathfrak{X} \rangle + \varepsilon,$$

wherein $\mathfrak{B} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ is the regression coefficient tensor that captures the association between the score y and the tensor predictor (image scan) $\mathfrak{X} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$. Note that the inner product between the coefficient tensor and the predictor tensor is defined as

$$\langle \mathfrak{B}, \mathfrak{X} \rangle = Vec(\mathfrak{B})^T Vec(\mathfrak{X}) = \sum_{i_1, i_2, i_3} \mathfrak{X}_{i_1,i_2,i_3} \mathfrak{B}_{i_1,i_2,i_3} \tag{1}$$

where $\mathfrak{X}_{i_1,i_2,i_3}$ and $\mathfrak{B}_{i_1,i_2,i_3}$ are the $(i_1, i_2, i_3)^{th}$ element of $\mathfrak{X}$ and $\mathfrak{B}$, respectively. The number of parameters in $\mathfrak{B}$ increases rapidly and can easily become larger than the available sample size, thus requiring some form of regularization to estimate the regression coefficient. To that end, [42] proposed a generalized linear model with a tensor predictor in the systematic component and utilized a CP Decomposition to reduce the high dimension of the coefficient tensor $\mathfrak{B}$. [22] considered a regression model with a multivariate (vector) response and a tensor predictor and assumed that regression coefficients admit a *sparse* CP decomposition. Some other work, dealing with sparse CP decomposition, includes [2], [10] and [31]. [34] extended the regression model to accommodate a tensor response, assumed the previously mentioned sparse CP decomposition on the regression coefficient ([22]) and also developed an alternating updating algorithm to obtain a (local) minimizer of the underlying non-convex optimization problem. [21] extended the work of [42] by using a Tucker decomposition, which is a generalized version of the CP decomposition. [32] presented a general convex optimization approach for solving tensor regression problems by applying convex and *weakly decomposable* regularizers on the regression coefficient. They established the weak-decomposability of the sparsity regularizer (both element-wise

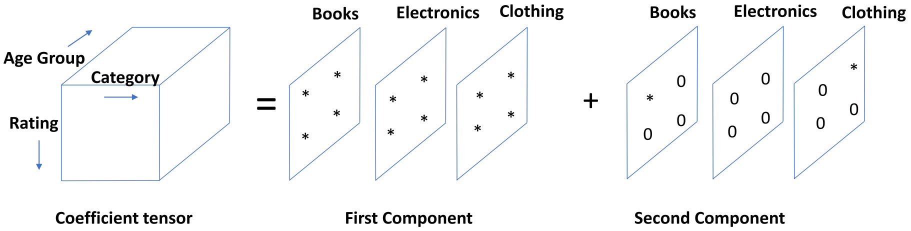


FIG 1. *Illustration of example: In the first component, all the "rating type - age group" combinations share similar baseline effects across the three categories. In the second component, some of the combinations exhibit additional effects specific to some of the categories.*

and group-wise) and low-rankness regularizer (nuclear norm, defined through its dual norm) and derived their theoretical results for these two special cases.

While a sparse CP decomposed structure on $\mathfrak{B}$ enables estimation with relatively small sample size, in many applications it may not be particularly suitable, since the data may exhibit more complex structure. For example, imaging data over different time points can be considered as a three-dimensional predictor, while the response can be a clinical outcome (see [22]). The images may have similar baseline effects across all the time points, while there might be selected segments of the images showing some additional effects, which are specific to selected time points only. A sparse CP decomposed regression model does not account for this structure, that in turn may not help researchers obtain good scientific insights from their data.

Another example relates to online commerce data, where the predictor can be organized as a three-dimensional tensor with one dimension corresponding to product categories (books, electronics and clothing accessories), the second dimension to ratings of these products (different types of ratings, including "fashionable", "durable" and so on) and the third dimension to age groups of the customers, as shown in Figure 1 (see Example 2 in [1]). The response can be the gross sales of the company. It is meaningful to decompose the effects of the coefficient tensor into two parts. In the first component, all the "rating type - age group" combinations share a similar baseline effect across the product categories. On the other hand, there might be some of the "rating type - age group" combinations, for which additional effects are present only in some specific categories. For instance, the combination of rating type label "fashionable" and age group range "16 years - 30 years" should have additional effects in the clothing accessories category. Similarly the combination of label "durable" and "31 years - 50 years" demands additional effect in the electronics category.

In this paper, we propose a tensor regression model with a scalar response, a *third-order tensor* predictor and consider a low-dimensional structure on the coefficient tensor suitable for the previously mentioned examples. The key technical assumption is that one component of the regression coefficient $\mathfrak{B}$ exhibits *low tubal rank*, a concept introduced for tensor decompositions in [14] and [13] and briefly explained next. As depicted in Figure 2, in the case of a third-order

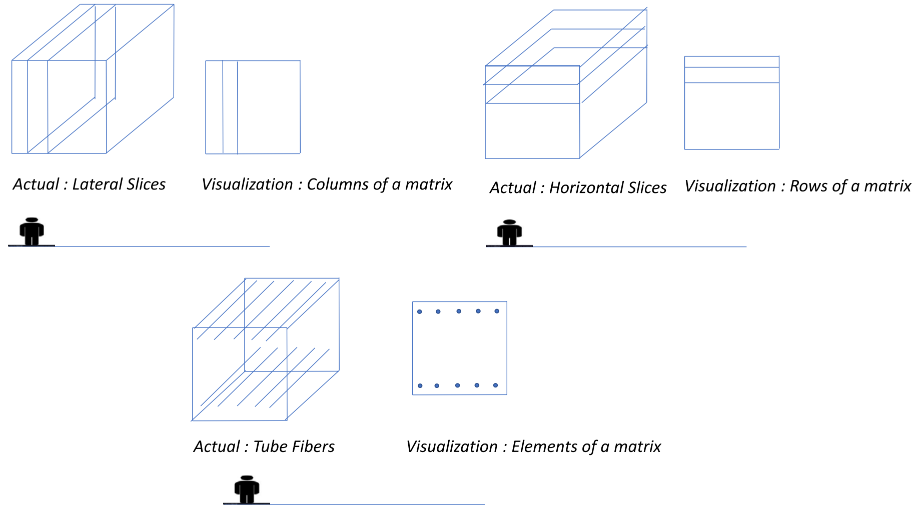

Actual : Lateral Slices
Visualization : Columns of a matrix
Actual : Horizontal Slices
Visualization : Rows of a matrix
Actual : Tube Fibers
Visualization : Elements of a matrix

Fig 2. *Matrix-type view of a third-order tensor: Lateral Slices, Horizontal Slices and Tube Fibers can be visualized as columns, rows and elements respectively*

tensor, the horizontal slices, lateral slices and the tube fibers (see *Notation* for rigorous definitions and Figures 2.1 and 2.2 in [16] for a pictorial illustration) play the roles of rows, columns and elements of a matrix, respectively. The question addressed in the aforementioned papers is "How does one extend the well known concepts of matrix algebra, e.g., linear combination, linear dependence, rank and so on to this case?". Their key technical development lies in the novel concept of *t-product* [Definition C.1] between a lateral slice (or, horizontal slice) and a tube, which is the tensor counterpart of the product between a column (or, row) and a scalar. Along the same lines, the *t-linear combination* [Definition C.7] of lateral slices and the *Range* of the tensor are defined, which can be conceptualized as the tensor counterparts of the linear combination of the columns and the column space, respectively. Finally, [14] shows that the number of elements needed to generate any element in the range, is the same as the *tubal rank* [Definition C.9]. Thus, just like the rank of a matrix, tubal-rank of a third-order tensor determines the number of lateral (and horizontal) slices that are *t-linearly independent* (see the discussion after Definition C.7, Figure 3 and Section 2 for more details). In addition to the low-dimensional structure as expressed through a low-tubal rank, to capture idiosyncratic effects, we assume that the coefficient tensor $\boldsymbol{\mathcal{B}}$ can be decomposed as follows: $\boldsymbol{\mathcal{B}} = \boldsymbol{\mathcal{L}} + \boldsymbol{\mathcal{S}}$, wherein the first component is a low tubal-rank tensor, where the baseline effects are shared across the slices and the second component is sparse, reflecting additional idiosyncratic effects (see Figure 1).

The key novel contribution of this paper is to develop the algorithm and technical tools to obtain estimates of the two components of the regression coefficient $\boldsymbol{\mathcal{B}}$ and establish a non-asymptotic upper bound to their estimation error. This type of decomposition has been employed earlier in another line of research that deals with tensor recovery (see [23]). However, to the best of our knowledge, the proposed methodology and the subsequent theoretical analysis are new in the context of tensor regression. [20] considered a tensor regression model with sparse tubal-regularized penalization. However, the latter paper is motivated by a different perspective vis-a-vis the current work. As mentioned in the previous paragraph, the current work characterizes the baseline effects as the low tubal-rank tensor and the idiosyncratic effects as the sparse tensor. Thus, this can be conceptualized as the third-order generalization of the matrix low-rank plus sparse approach in [1]. On the other hand, [20] does not consider any such decomposition of their coefficient tensor into baseline ($\boldsymbol{\mathcal{L}}$ in our case) and idiosyncratic parts ($\boldsymbol{\mathcal{S}}$ in our case). Rather, in order to reduce the dimension of the regression coefficient, they simply assume *both* low tubal-rank and sparse structure on the coefficient tensor $\boldsymbol{\mathcal{W}}$. Consequently, the objective functions considered in the two papers are different. Finally, in terms of theoretical developments, a novel incoherence condition is needed for identifiability of the two components in our case and we also provide a detailed derivation and interpretation of the non-asymptotic upper bound of the estimation error. The bounds, as discussed in Section 3, are in line with the matrix case results [1]. On the other hand, [20] did not make any such attempt in their paper.

The remainder of the paper is organized as follows: In section 2, we develop and provide intuition on how to interpret the low tubal rank regression model. Section 2.1 discusses the convex relaxation and presents a block alternating minimization algorithm to estimate the unknown model parameters from data. Theoretical results related to estimation error bounds are discussed in Section 3. In Sections 4 and 5, we evaluate the performance of our model on synthetic and real data, respectively. Finally, we conclude with a discussion in Section 6. Background material on the t-product and the corresponding t-Singular Value Decomposition and related concepts are provided in Appendix C.

Notation: The *order* of a tensor is the number of its dimensions. For a tensor of order N, we use $d_1, d_2, \cdots, d_N$ to denote the *size* of the tensor along each of the N dimensions. Throughout the paper, tensors of order three or higher are denoted by boldface Euler script letters, e.g., $\boldsymbol{\mathcal{X}}$. We write $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}$ to represent an N-order tensor of size $d_1 \times d_2 \times ... \times d_N$. For any third-order tensor $\boldsymbol{\mathcal{X}}$, $\mathcal{X}_{ijk}$ denotes the $(i,j,k)^{th}$ element of $\boldsymbol{\mathcal{X}}$. One dimensional sections of a third-order tensor $\boldsymbol{\mathcal{X}}$, namely, *Column Fiber*, *Row Fiber* and *Tube Fiber* (see Figure 2.1 of [16]) are denoted by $\boldsymbol{x_{:jk}}$, $\boldsymbol{x_{i:k}}$ and $\boldsymbol{x_{ij:}}$ respectively. Similarly, the two-dimensional sections, namely, *Horizontal Slice*, *Lateral Slice* and *Frontal Slice* (see Figure 2.2 of [16]) are denoted by $\boldsymbol{X_{i::}}$, $\boldsymbol{X_{:j:}}$ and $\boldsymbol{X_{::k}}$ respectively. As illustrated in Figure 2, 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 $\|A\|_\infty = \max\limits_{i,j} \mid a_{ij} \mid$. $\ell_{2,1}$ 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, $\|A\|_{2,\infty} = \max_{1\leq j\leq d_2} (\sum_{i=1}^{d_1} a_{ij}^2)^{\frac{1}{2}}$. 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 $\|A\|_{sp} = \max_{1\leq j\leq d} \{\sigma_j(A)\}$.

2. Low tubal-rank tensor regression model and its estimation

Let $y \in \mathbb{R}$ be a scalar response and $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1\times d_2\times d_3}$ be a third-order tensor predictor. We propose the following tensor regression model,

$$y = \langle \boldsymbol{\mathcal{B}}^*, \boldsymbol{\mathcal{X}} \rangle + \varepsilon \tag{2}$$

The coefficient tensor $\boldsymbol{\mathcal{B}}^* \in \mathbb{R}^{d_1\times d_2\times d_3}$ captures the association between the scalar response and the tensor predictor. The inner product is defined as in (1). Further, it is assumed that $\varepsilon \sim N(0, \sigma^2)$. To deal with the large number of regression coefficients, we posit that $\boldsymbol{\mathcal{B}}^* = \boldsymbol{\mathcal{L}}^* + \boldsymbol{\mathcal{S}}^*$, where $\boldsymbol{\mathcal{L}}^*$ and $\boldsymbol{\mathcal{S}}^*$ are characterized by two complementary types of low dimensional structure, discussed next.

The $\boldsymbol{\mathcal{L}}^*$ component corresponds to a low tubal-rank tensor. As shown in [13] and discussed in Section 1, the tubal-rank of a third-order tensor is analogous (in the appropriate algebra, see Appendix C) 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 the tubal-rank characterizes a similar type of dependence, namely *t-linear* dependence (see Definition C.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 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 [13]. Although the formal definitions are deferred to Appendix C, Figures 3 and 4 illustrate the key ideas. As Figure 3 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 3 also provides an alternative representation of t-dependence in terms of a block-circulant matrix [see Notation B.2]. Figure 4 provides some numerical examples to show the relation between tubal-rank and t-dependence among the lateral slices. Based on this brief discussion, the purpose of the first component $\boldsymbol{\mathcal{L}}^*$ becomes to capture similar baseline effects.

The second component, $\boldsymbol{\mathcal{S}}^*$ consists of the additional effects, which might be present only in some of the specific lateral slices and they can be at element, row or column level within each slice. For the posited model, it is assumed without loss of generality that the slice specific effects are present column-wise in $\boldsymbol{\mathcal{S}}^*$. However, one may also assume any of the other two possibilities with few minor adjustments, as discussed in the sequel. It is interesting to note that, one can

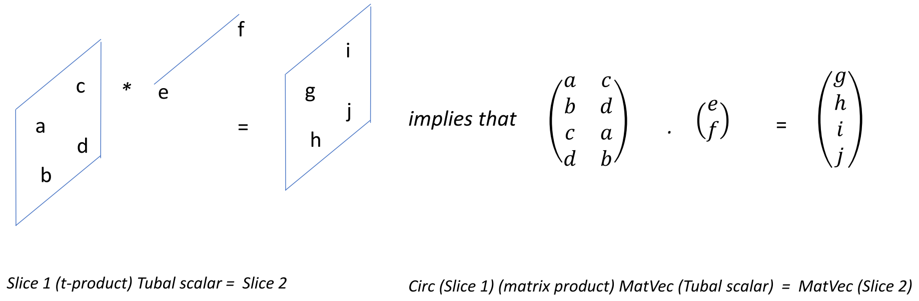
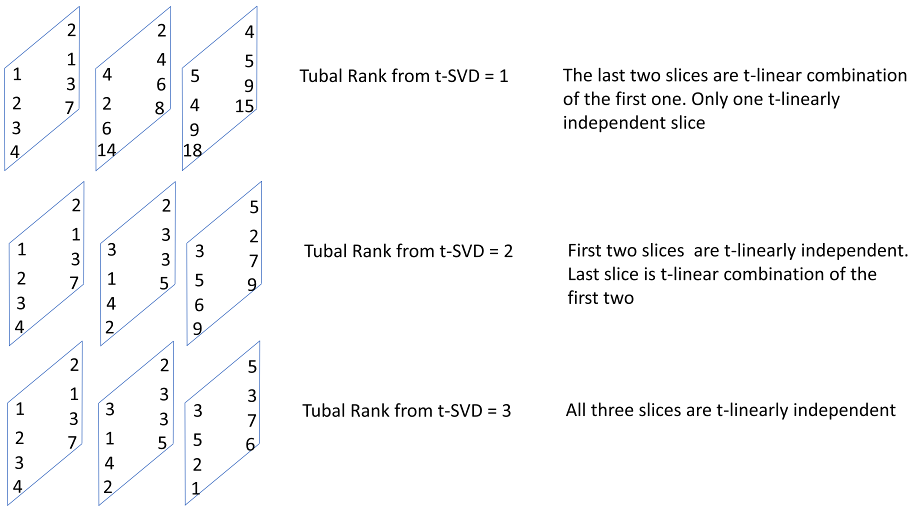

a
b
c
d
*
e
f
=
g
h
i
j
implies that
$\begin{pmatrix} a & c \\ b & d \\ c & a \\ d & b \end{pmatrix} \cdot \begin{pmatrix} e \\ f \end{pmatrix} = \begin{pmatrix} g \\ h \\ i \\ j \end{pmatrix}$
Slice 1 (t-product) Tubal scalar = Slice 2
Circ (Slice 1) (matrix product) MatVec (Tubal scalar) = MatVec (Slice 2)

FIG 3. *t-linear dependence between two lateral slices*



Tubal Rank from t-SVD = 1

The last two slices are t-linear combination of the first one. Only one t-linearly independent slice

Tubal Rank from t-SVD = 2

First two slices are t-linearly independent. Last slice is t-linear combination of the first two

Tubal Rank from t-SVD = 3

All three slices are t-linearly independent

FIG 4. *Tubal Rank and t-linear dependence*

have an alternative representation of $\mathcal{S}^*$, which is easier to work with, namely that its frontal slices are column-wise sparse.

Based on the above discussion, we rewrite the regression model in more explicit form as follows:

$$y = \langle \mathcal{L}^* + \mathcal{S}^*, \mathcal{X} \rangle + \varepsilon, \tag{3}$$

where $\mathcal{L}^*$ is a low-tubal rank tensor, $\mathcal{S}^*$ is a tensor whose frontal slices are column-wise sparse and $\varepsilon \sim N(0, \sigma^2)$. The goal becomes to estimate both $\mathcal{L}^*$ and $\mathcal{S}^*$ based on available data.

Note that similarly to the matrix case [1], an additional constraint is needed in order to make the model identifiable. Specifically, $\mathcal{L}^*$ needs to be "incoherent" with $\mathcal{S}^*$, an issue addressed in Section 2.1.

2.1. Estimation of the tensor regression coefficient

Analogously to the matrix case, the tubal-rank is non-convex. Hence, for estimation purposes we aim to leverage a convex relaxation. To that end, from equation (31) and the related discussion in Appendix B, it can be seen that $\frac{1}{d_3}\left\|Circ(\boldsymbol{\mathcal{L}^*})\right\|_*$ corresponds to such a convex relaxation, where $Circ(\boldsymbol{\mathcal{L}}^*)$ is the *block-circulant matrix* associated with the tensor $\boldsymbol{\mathcal{L}}^*$ (see notation B.2 in Appendix) and $\|\cdot\|_*$ denotes the matrix nuclear norm. Indeed, it is an alternative form of the Tensor Nuclear Norm defined in [26] (see Definition 7 in [26] and Equation (12) in [25]) and imposing a restriction on this norm translates into an analogous restriction on the tubal-rank. Further, since $\boldsymbol{\mathcal{S}}^*$ consists of column-wise sparse frontal slices, one can simply treat each frontal slice as a matrix and employ the usual column-wise $\ell_{2,1}$ norm for each of them. Hence, we consider a regularizer $\sum_{k=1}^{d_3}\|\boldsymbol{S^*_{::k}}\|_{2,1}$ to constrain the sparse component.

The objective function for estimating the tensor regression coefficient $\boldsymbol{\mathcal{B}}$ based on the posited low-tubal rank and structured sparse decomposition is given by:

$$\min_{\boldsymbol{\mathcal{L}},\boldsymbol{\mathcal{S}}}\{\frac{1}{2n}\sum_{i=1}^{n}(y_i-\langle\boldsymbol{\mathcal{L}}+\boldsymbol{\mathcal{S}},\boldsymbol{\mathcal{X}}_i\rangle)^2+\lambda_L\frac{1}{d_3}\left\|Circ(\boldsymbol{\mathcal{L}})\right\|_*+\lambda_S\sum_{k=1}^{d_3}\|\boldsymbol{S_{::k}}\|_{2,1}\} \tag{4}$$

wherein λ_L and λ_S are non-negative regularization parameters corresponding to low tubal-rank and sparse components, respectively. The factor $\frac{1}{d_3}$ can be interpreted as follows: for any third-order tensor in $\mathbb{R}^{d_1\times d_2\times d_3}$, the associated block-circulant matrix consists of d_2 blocks. Each block corresponds to a particular lateral slice and contains all of the d_3 possible block-circulant arrangements of that lateral slice. Hence, besides causing inter-slice t-dependence, the penalty on $\left\|Circ(\boldsymbol{\mathcal{L}}^*)\right\|_*$ also induces intra-slice dependence among the d_3 block-circulant arrangements of the slice. The factor $\frac{1}{d_3}$ thus adjusts for this additional penalization.

Before presenting an algorithm for estimating $(\boldsymbol{\mathcal{L}}^*,\boldsymbol{\mathcal{S}}^*)$, we address the issue of identifiability of these parameters. In the matrix case, an *incoherence condition* is required and usually operationalized through conditions on the singular vectors of the low rank component obtained from the SVD (see, e.g., [6], [4] and [37]). We adapt the approach used in [1] to the low tubal rank tensor $\boldsymbol{\mathcal{L}}^*$ and the structured sparse tensor $\boldsymbol{\mathcal{S}}^*$ and require

$$\|Circ(\boldsymbol{\mathcal{L}}^*)\|_{2,\infty}\le\frac{\alpha}{\sqrt{d_2}},$$

for some fixed parameter $\alpha>0$. Note that based on Proposition A.1, a low tubal rank for $\boldsymbol{\mathcal{L}}^*$ translates to low 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 imposing this “spikeness” restriction on the columns of $Circ(\boldsymbol{\mathcal{L}}^*)$, one can ensure a sufficient number of non-zero columns in $Circ(\boldsymbol{\mathcal{L}}^*)$ and thus in each of the frontal slices of $\boldsymbol{\mathcal{L}}^*$. However, each of the last $d_2(d_3-1)$ columns of $Circ(\boldsymbol{\mathcal{L}}^*)$ can be written by rearranging elements of any one of its first d_2 columns. Hence by restricting the "spikiness" of only the first d_2 columns, one can essentially control the "spikiness" of all the columns of $Circ(\boldsymbol{\mathcal{L}}^*)$, which leads to the posited incoherence conditions.

Note that the objective function (4), denoted by $f(\boldsymbol{\mathcal{L}}, \boldsymbol{\mathcal{S}})$, is jointly convex and hence the following alternating block minimization procedure summarized in Algorithm 1, will obtain the desired minimizer. The details of Steps 1 and 2, that update $\boldsymbol{\mathcal{L}}$ and $\boldsymbol{\mathcal{S}}$ alternatively, are discussed next.

Algorithm 1 Alternating Block Minimization Procedure for minimizing $f(\boldsymbol{\mathcal{L}}, \boldsymbol{\mathcal{S}})$

```
Input: data $\{(y_i, \boldsymbol{\mathcal{X}}_i), i = 1, 2, \cdots, n\}, \lambda_L, \lambda_S$
Initialize: $\boldsymbol{\mathcal{L}}^{(0)}, \boldsymbol{\mathcal{S}}^{(0)} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$
repeat
    Step 1: Update $\boldsymbol{\mathcal{L}}^{(t+1)} = \arg\min_{\boldsymbol{\mathcal{L}}} f(\boldsymbol{\mathcal{L}}, \boldsymbol{\mathcal{S}}^{(t)})$, given $\boldsymbol{\mathcal{S}}^{(t)}$
    Step 2: Update $\boldsymbol{\mathcal{S}}^{(t+1)} = \arg\min_{\boldsymbol{\mathcal{S}}} f(\boldsymbol{\mathcal{L}}^{(t+1)}, \boldsymbol{\mathcal{S}})$, given $\boldsymbol{\mathcal{L}}^{(t+1)}$
until $f(\boldsymbol{\mathcal{L}}^{(t+1)}, \boldsymbol{\mathcal{S}}^{(t+1)})$ converges
```

Step 1: Step 1 updates the value of $\boldsymbol{\mathcal{L}}$ given $\boldsymbol{\mathcal{S}}$. For a given value of $\boldsymbol{\mathcal{S}}$, letting $u_i = y_i - \langle \boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{X}}_i \rangle$, the problem then reduces to minimizing $g(\boldsymbol{\mathcal{L}}) = \frac{1}{2n} \sum_{i=1}^{n} (u_i - \langle \boldsymbol{\mathcal{L}}, \boldsymbol{\mathcal{X}}_i \rangle)^2 + \lambda_L \frac{1}{d_3} \|Circ(\boldsymbol{\mathcal{L}})\|_*$, with respect to $\boldsymbol{\mathcal{L}}$. Further, denoting the matrix $\frac{Circ(\boldsymbol{\mathcal{L}})}{d_3}$ by W and $Circ(\boldsymbol{\mathcal{X}}_i)$ by V_i, some simple algebraic steps show that, minimizing $g(\boldsymbol{\mathcal{L}})$ with respect to $\boldsymbol{\mathcal{L}}$, is equivalent to minimizing $g(W) = \frac{1}{2n} \sum_{i=1}^{n} (u_i - Tr(W^T V_i))^2 + \lambda_L \|W\|_*$ with respect to W. This minimization problem shows up in various applications of machine learning, such as matrix classification, multi-task learning and matrix completion (see [3, 35]). [12] consider a general class of optimization problems that includes the above formulation: specifically, for a matrix variable M, the objective function of interest is given by $\min_{M} \{F(M) + \lambda \|M\|_*\}$, wherein $F(\cdot)$ is a smooth convex function. [12] proposed an Extended Gradient Algorithm and Accelerated Gradient Algorithm to obtain the minimizer M and also addressed convergence issues. A direct application of the aforementioned algorithms provides the optimal solution W and thus eventually, the optimal $\boldsymbol{\mathcal{L}}$.

Step 2: In Step 2, for a given value of $\boldsymbol{\mathcal{L}}$, letting $z_i = y_i - \langle \boldsymbol{\mathcal{L}}, \boldsymbol{\mathcal{X}}_i \rangle$, the problem boils down to minimizing $h(\boldsymbol{\mathcal{S}}) = \frac{1}{2n} \sum_{i=1}^{n} (z_i - \langle \boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{X}}_i \rangle)^2 + \lambda_S \sum_{k=1}^{d_3} \|\boldsymbol{S}_{\boldsymbol{::k}}\|_{2,1}$, with respect to $\boldsymbol{\mathcal{S}}$. We now construct the matrix $\boldsymbol{\mathcal{S}}_{Mat}$ of dimension $d_1 \times d_2 d_3$ (and similarly $\boldsymbol{\mathcal{X}}_{iMat}$), by placing the frontal slices of $\boldsymbol{\mathcal{S}}$ (and of $\boldsymbol{\mathcal{X}}_i$) side by side. One can easily check that $\langle \boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{X}}_i \rangle = Tr(\boldsymbol{\mathcal{S}}_{Mat}^T \boldsymbol{\mathcal{X}}_{iMat})$ and $\sum_{k=1}^{d_3} \|\boldsymbol{S}_{\boldsymbol{::k}}\|_{2,1} = \|\boldsymbol{\mathcal{S}}_{Mat}\|_{2,1}$ and thus $h(\boldsymbol{\mathcal{S}}) = \frac{1}{2n} \sum_{i=1}^{n} (z_i - Tr(\boldsymbol{\mathcal{S}}_{Mat}^T \boldsymbol{\mathcal{X}}_{iMat}))^2 + \lambda_S \|\boldsymbol{\mathcal{S}}_{Mat}\|_{2,1}$. Finally, vectorizing $\boldsymbol{\mathcal{S}}_{Mat}$ and $\boldsymbol{\mathcal{X}}_{iMat}$, the problem takes the form of a group-lasso penalized learning problems, as discussed in [38]. Assuming that the loss function admits the Quadratic Majorization condition, [38] develops Groupwise-Majorization-Descent (GMD) algorithm in order to solve such problems. We directly employ that method to obtain the optimal $\boldsymbol{\mathcal{S}}$.

The above algorithm requires a slight modification to accommodate sparsity at the element level for the component $\mathcal{S}^*$. Specifically, in Step 2 the objective function $h(\mathcal{S})$ penalizes the ℓ_1 norm of $vec(\mathcal{S}_{Mat})$, and thus the optimal $\mathcal{S}$ is obtained by solving a *lasso* problem.

3. Theoretical results

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

$$e^2(\hat{\mathcal{L}}, \hat{\mathcal{S}}) = \left\|\hat{\mathcal{L}} - \mathcal{L}^*\right\|_F^2 + \left\|\hat{\mathcal{S}} - \mathcal{S}^*\right\|_F^2 \tag{5}$$

Next, we introduce some additional notation needed in the sequel.

Additional notation: For $\mathcal{L}^*$ with tubal rank $r \ll \min\{d_1, d_2\}$, the rank of the associated block-circulant matrix is denoted by R and is bounded above by $r \times d_3$ (see Proposition A.1). $\mathcal{S}^*_{Mat}$ is a matrix of dimension $d_1 \times d_2 d_3$, that is constructed by placing the frontal slices of $\mathcal{S}^*$ side by side. We assume that $\mathcal{S}^*_{Mat}$ has $s \ll d_2 d_3$ non-zero columns. More specifically, suppose that $\mathcal{S}^*_{Mat}$ is supported on a subset $E \subseteq \{1, 2, \cdots, d_2 d_3\}$, with $|E| = s$. We define a pair of subspaces $(\mathbb{M}(E), \mathbb{M}^\perp(E))$, such that, $\mathbb{M}(E) = \{M \in \mathbb{R}^{d_1 \times d_2 d_3} \mid k^{th}$ column of $M = 0, \forall k \notin E\}$ and $\mathbb{M}^\perp(E) = (\mathbb{M}(E))^\perp$. As shown in [1, 28], one can easily verify that for any $M_1 \in \mathbb{M}(E)$ and $M_2 \in \mathbb{M}^\perp(E)$, $\|M_1 + M_2\|_{2,1} = \|M_1\|_{2,1} + \|M_2\|_{2,1}$. This ensures that the regularizer $\|\cdot\|_{2,1}$ is *decomposable* (see [28]) 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, $\mathcal{S}^*_{Mat} \in \mathbb{M}$, $\pi_{\mathbb{M}}(\mathcal{S}^*_{Mat}) = \mathcal{S}^*_{Mat}$ and $\pi_{\mathbb{M}^\perp}(\mathcal{S}^*_{Mat}) = 0$, where $\pi_{\mathbb{M}}(\cdot)$ is the projection onto the subspace $\mathbb{M}$. We define, $\hat{\Delta}_{\mathcal{L}} = \hat{\mathcal{L}} - \mathcal{L}^*$, $\hat{\Delta}_{\mathcal{S}} = \hat{\mathcal{S}} - \mathcal{S}^*$. $\hat{\Delta}_{\mathcal{S}Mat}$ is the matrix constructed by placing the frontal slices of $\hat{\Delta}_{\mathcal{S}}$ side by side. $\hat{\Delta}^{\mathbb{M}}_{\mathcal{S}Mat} = \pi_{\mathbb{M}}(\hat{\Delta}_{\mathcal{S}Mat})$ and $\hat{\Delta}^{\mathbb{M}^\perp}_{\mathcal{S}Mat} = \pi_{\mathbb{M}^\perp}(\hat{\Delta}_{\mathcal{S}Mat})$. $\hat{\Delta}^{\mathbb{M}}_{\mathcal{S}}$ and $\hat{\Delta}^{\mathbb{M}^\perp}_{\mathcal{S}}$ are the tensor counterparts of $\hat{\Delta}^{\mathbb{M}}_{\mathcal{S}Mat}$ and $\hat{\Delta}^{\mathbb{M}^\perp}_{\mathcal{S}Mat}$ respectively.

The roadmap of the technical developments is as follows: Lemmas 3.1 and 3.2 characterize the set to which the errors $(\hat{\Delta}_{\mathcal{L}}, \hat{\Delta}_{\mathcal{S}})$ belong. In addition, we assume that Restricted Strong Convexity of the loss function on this set holds (Assumption 1). For deterministic realizations of the predictors and the error terms and under certain regularity conditions, Lemma 3.3 establishes the bound on the estimation error $e^2(\hat{\mathcal{L}}, \hat{\mathcal{S}})$. Theorem 3.4 extends the result to random realizations of the predictors and the errors, while Corollary 3.4.1 presents the error bound when $\mathcal{S}^*$ has element-wise sparse frontal slices.

The proofs of the results are delegated to Appendix A.

Lemma 3.1. *Let R denote the rank of $Circ(\mathcal{L}^*)$. Let $C(\mathcal{L}, \mathcal{S})$ be a weighted combination of the nuclear norm and the $\ell_{2,1}$ norm regularizers as follows:*

$$C(\mathcal{L}, \mathcal{S}) = \frac{1}{d_3} \|Circ(\mathcal{L})\|_* + \frac{\lambda_S}{\lambda_L} \|\mathcal{S}_{Mat}\|_{2,1}$$

Then, for any $R = 1, 2, \cdots, \min\{d_1 d_3, d_2 d_3\}$*, there exists a decomposition* $\hat{\Delta}_{\mathcal{L}} = \hat{\Delta}^A_{\mathcal{L}} + \hat{\Delta}^B_{\mathcal{L}}$ *with Rank* $(Circ(\hat{\Delta}^A_{\mathcal{L}})) \leq 2R$*,* $Circ(\mathcal{L}^*)^T Circ(\hat{\Delta}^B_{\mathcal{L}}) = 0$*,* $Circ(\mathcal{L}^*) Circ(\hat{\Delta}^B_{\mathcal{L}})^T = 0$ *and*

$$C(\mathcal{L}^*, \mathcal{S}^*) - C(\mathcal{L}^* + \hat{\Delta}_{\mathcal{L}}, \mathcal{S}^* + \hat{\Delta}_{\mathcal{S}}) \leq C(\hat{\Delta}^A_{\mathcal{L}}, \hat{\Delta}^{\mathbb{M}}_{\mathcal{S}}) - C(\hat{\Delta}^B_{\mathcal{L}}, \hat{\Delta}^{\mathbb{M}^\perp}_{\mathcal{S}}) \qquad (6)$$

Lemma 3.2. *Suppose the predictors* $\mathfrak{X}_i$*'s and the errors* ϵ_i*'s are deterministic. Define a third-order tensor* $\mathfrak{D} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ *as follows:*

$$\mathfrak{D} = \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \mathfrak{X}_i$$

Also let $\mathfrak{D}_{Mat}$ *be a matrix obtained by placing the frontal slices of* $\mathfrak{D}$ *side by side. Then under the conditions* $\lambda_L \geq 4\frac{1}{d_3} \|Circ(\mathfrak{D})\|_{sp}$ *and* $\lambda_s \geq 4 \|\mathfrak{D}_{Mat}\|_{2,\infty}$*, the estimation error* $(\hat{\Delta}_{\mathcal{L}}, \hat{\Delta}_{\mathcal{S}})$ *satisfies the following constraint:*

$$C(\hat{\Delta}^B_{\mathcal{L}}, \hat{\Delta}^{\mathbb{M}^\perp}_{\mathcal{S}}) \leq 3C(\hat{\Delta}^A_{\mathcal{L}}, \hat{\Delta}^{\mathbb{M}}_{\mathcal{S}}) \qquad (7)$$

As mentioned earlier, the above lemmas characterize a set in which the error $(\hat{\Delta}_{\mathcal{L}}, \hat{\Delta}_{\mathcal{S}})$ lies. Given this set, we are now in a position to summarize all the assumptions that we make. We first prepare a list of the assumptions and then provide further details on each of those assumptions.

Assumption 1. *The loss function* $L(\mathcal{L}^*, \mathcal{S}^*)$ *satisfies Restricted Strong Convexity with curvature* $\gamma > 0$ *(and tolerance* $\tau_L = 0$*) over the set, characterized by Lemma 3.1 and Lemma 3.2. In other words, there exists a positive constant* $\gamma > 0$ *such that*

$$\frac{1}{2n} \sum_{i=1}^{n} \{\langle \Delta_{\mathcal{L}} + \Delta_{\mathcal{S}}, \mathfrak{X}_i \rangle\}^2 \geq \frac{\gamma}{2} \|\Delta_{\mathcal{L}} + \Delta_{\mathcal{S}}\|_F^2, \text{for all } (\Delta_{\mathcal{L}}, \Delta_{\mathcal{S}}) \text{ satisfying equation } (7) \qquad (8)$$

Assumption 2. $\|Circ(\mathcal{L}^*)\|_{2,\infty} \leq \frac{\alpha}{\sqrt{d_2}}$*, for some fixed parameter* α

Assumption 3. *When the predictors* $\mathfrak{X}_i$*'s and the errors* ϵ_i*'s are deterministic, the regularizer parameters* (λ_L, λ_S) *satisfy*

$$\lambda_L \geq 4\frac{1}{d_3} \|Circ(\mathfrak{D})\|_{sp} \text{ and } \lambda_s \geq 4 \|\mathfrak{D}_{Mat}\|_{2,\infty} + \frac{4\gamma\alpha}{\sqrt{d_2}} \qquad (9)$$

where $\mathfrak{D}$ *and* $\mathfrak{D}_{Mat}$ *are as defined in Lemma 3.2.*

- Assumption 1 ensures that the loss function exhibits strong convexity over some restricted set of interest, as defined in equation (7). This is a fairly standard assumption in the high-dimensional literature [1].

- Assumption 2 is aimed to ensure that the low-tubal rank component $\boldsymbol{\mathcal{L}}^*$ is incoherent with the sparse component $\boldsymbol{\mathcal{S}}^*$, as discussed in Section 2.1. It is worth recalling that this assumption is a straightforward application of the 'spikiness' restriction on the columns of the low-rank matrix, as introduced in [1]. 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 other Tensor Incoherence conditions, including those in [24, 40], which involve the components of the t-SVD.
- Assumption 3 imposes a certain lower bound to the two regularizer parameters, a common requirement in the high-dimensional literature.

The following lemma establishes an upper bound to $e^2(\hat{\boldsymbol{\mathcal{L}}}, \hat{\boldsymbol{\mathcal{S}}})$ in the case of deterministic predictors and errors.

Lemma 3.3. *Suppose the predictors $\boldsymbol{\mathcal{X}}_i$'s and the errors ϵ_i's are deterministic. Then, under Assumptions (1), (2) and (3), the estimation error $e^2(\hat{\boldsymbol{\mathcal{L}}}, \hat{\boldsymbol{\mathcal{S}}})$ satisfies the following:*

$$e^2(\hat{\boldsymbol{\mathcal{L}}}, \hat{\boldsymbol{\mathcal{S}}}) \preceq \lambda_L^2 \; r + \lambda_S^2 \; s \tag{10}$$

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

Note that the result is broadly in line with Theorem 1 in [1]; specifically, when the loss function satisfies the Restricted Strong Convexity and the parameters of interest are exactly (not approximately) Low Rank and Sparse, a similar form error bound is obtained. In the current setting, the tubal-rank (instead of the matrix rank) enters the bound, as well as the columnwise sparsity s reflecting the nature of $\boldsymbol{\mathcal{S}}^*_{Mat}$.

Next, the above result is extended to the case of stochastic errors and predictors. To that end, we assume that ϵ_i's are i.i.d. $N(0, \sigma^2)$ and use the notation $X^{(i)} = \text{Vec}(\boldsymbol{\mathcal{X}}_i)$ in order to denote the vectorized form of the i^{th} predictor $\boldsymbol{\mathcal{X}}_i$. We write

$$X = ((X^{(1)})^T, (X^{(2)})^T, \cdots, (X^{(n)})^T)^T \in \mathbb{R}^{nd_1d_2d_3} \tag{11}$$

to denote the combined predictors from all the n samples in vectorized form. As in [32], we assume that $X \sim N(0, \Sigma)$, where $\Sigma = \text{Cov}(X) \in \mathbb{R}^{nD \times nD}$ and $D = d_1d_2d_3$. Note that this assumption does not require the data tensors $\boldsymbol{\mathcal{X}}_i$'s to be independent. We assume that Σ has bounded eigenvalues. Let $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the smallest and largest eigenvalues of a matrix respectively. We assume in the sequel that

$$c_l^2 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq c_u^2 \tag{12}$$

for some constants $0 < c_l \leq c_u < \infty$. As mentioned in [32], it is evident that in particular if all the covariates $\{X^{(i)} : i = 1, 2, \cdots, n\}$ are independent and identically distributed, then Σ will have a block-diagonal structure and in that case, the condition in equation (12) reduces to the similar conditions on $\text{Cov}(X^{(i)})$.

With this Gaussian assumption on the predictors and errors, we establish the following result.

Theorem 3.4. *Suppose ϵ_i's are i.i.d. $N(0, \sigma^2)$ and the predictors follow a Gaussian distribution, characterized by equation (11) and equation (12). Suppose that Assumption 2 holds. Then it can be shown that the conditions in Assumption 1 and Assumption 3 are satisfied with high probability and we will have*

$$e^2(\hat{\mathcal{L}}, \hat{\mathcal{S}}) \leq c_1 \ \sigma^2 c_u^2 \frac{r(d_1+d_2)}{n} + c_2 \ [\sigma^2 c_u^2 \frac{sd_1}{n} + \sigma^2 c_u^2 \frac{s\log(d_2 d_3)}{n} + \frac{\alpha^2 s}{d_2}] \quad (13)$$

with probability greater than $1 - \exp(-9\log(d_2 d_3))$, where c_u^2 is defined in (12).

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 $\sigma^2 \frac{w(m_1+m_2)}{n}$. 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 rd_1 (or, rd_2) corresponds to the r t-independent lateral slices (or, horizontal slices) and estimation of the d_1 (or, d_2) tubes in that slice. The multiplicative factor remains the same, except the term c_u^2 that appears additionally in this case in order to accommodate the variability in the predictors.

The second part of the error bound is related to the sparse component and can be interpreted as follows: the first term $\sigma^2 c_u^2 \frac{sd_1}{n}$ arises as a result of estimating sd_1 non-zero parameters in $\mathcal{S}^*_{Mat}$ and the second term $\sigma^2 c_u^2 \frac{s\log(d_2d_3)}{n}$ is devoted to the selection of s positions to place the non-zero columns in $\mathcal{S}^*_{Mat}$. This selection problem induces the term $\log(\binom{d_2 d_3}{s}) \approx s\log(d_2 d_3)$. Finally, the last term $\frac{\alpha^2 s}{d_2}$ appears due to the non-identifiability of the model.

When the sparsity in the regression coefficient tensor arises element-wise, instead of columnwise, the estimation error bound can be obtained in an analogous manner, with the columnwise $\ell_{(2,1)}$ norm replaced with the elementwise ℓ_1 norm, that imposes the elementwise sparsity in $\mathcal{S}^*_{Mat}$. Further, instead of restricting the spikinesss of the columns, the incoherence condition now controls the elementwise spikiness of $Circ(\mathcal{L}^*)$ and thus Assumption 2 takes the form $\|Circ(\mathcal{L}^*)\|_\infty \leq \frac{\alpha}{\sqrt{d_1 d_2 d_3}}$. Also in Assumption 3, the second inequality becomes $\lambda_s \geq 4\,\|\mathcal{D}_{Mat}\|_\infty + \frac{4\gamma\alpha}{\sqrt{d_1 d_2 d_3}}$. With these modified versions of the assumptions and denoting the number of non-zero elements in $\mathcal{S}^*_{Mat}$ by s, we present next the following Corollary to the Theorem 3.4, that provides the estimation error bound in case of elementwise sparsity in $\mathcal{S}^*$.

Corollary 3.4.1. *Suppose the errors ϵ_i's are i.i.d. $N(0, \sigma^2)$, the predictors follow a Gaussian distribution, characterized by equation (11) and equation (12) and the modified version of the Assumption 2 holds. Then, it can be shown that the conditions in Assumption 1 and in modified version of Assumption 3 are* *satisfied with high probability and we obtain*

$$e^2(\hat{\boldsymbol{\mathcal{L}}}, \hat{\boldsymbol{\mathcal{S}}}) \leq c_1 \ \sigma^2 c_u^2 \frac{r(d_1+d_2)}{n} + c_2 \ [\sigma^2 c_u^2 \frac{s \log(d_1 d_2 d_3)}{n} + \frac{\alpha^2 s}{d_1 d_2 d_3}] \qquad (14)$$

with probability greater than $1-\exp(-9\log(d_1d_2d_3))$*, where* c_u^2 *is defined in (12).*

4. Performance evaluation

We illustrate the performance of our estimation procedure described in Section 2.1, based on synthetic data under different settings. We start by describing how the true $\boldsymbol{\mathcal{L}}^*$ and $\boldsymbol{\mathcal{S}}^*$ are generated.

For the $\boldsymbol{\mathcal{L}}^*$, we start by generating a third-order tensor in $\mathbb{R}^{d_1\times d_2\times d_3}$ with Uniform $(0,1)$ entries and then obtain its t-SVD (see Definition C.8) using the `rTensor` `R` package ([19]). Let $\boldsymbol{\mathcal{U}}$ and $\boldsymbol{\mathcal{V}}$ denote the two orthogonal tensors (Definition C.5) and let $\boldsymbol{\mathcal{K}}$ be the f-diagonal tensor (Notation C.1) of the t-SVD. For any $r = 1, 2, \cdots, d = \min\{d_1, d_2\}$, we randomly select $d-r$ diagonal tubes of $\boldsymbol{\mathcal{K}}$ and make them zero, whereas the remaining tubes remain non-zero. Denoting the resulting f-diagonal tensor by $\boldsymbol{\mathcal{K}_1}$, $\boldsymbol{\mathcal{L}}^*$, with tubal rank r, is then generated as $\boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{K}_1} * \boldsymbol{\mathcal{V}^T}$.

To generate $\boldsymbol{\mathcal{S}}^*$, we start with a third-order tensor with Uniform$(0,1)$ entries as before. Then, for the k^{th} frontal slice, with $k = 1, 2, \cdots, d_3$, we randomly choose $s_k(\ll d_2)$ columns and set all the remaining $d_2 - s_k$ columns to zero. With this construction and denoting $\sum_{k=1}^{d_3} s_k$ by s, $\boldsymbol{\mathcal{S}}^*_{Mat}$ will have $s(\ll d_2 d_3)$ non-zero columns, as assumed in Section 3. However, for simplicity, in the simulations we assume that $s_1 = s_2 = \cdots = s_{d_3} = s^*$.

The predictors $\boldsymbol{\mathcal{X}}_i$'s are sampled independently, where in each of the predictors, the entries are i.i.d. $N(0,1)$. Finally we simulate independent and identically distributed entries of Gaussian noise for the error term and generate the responses based on Model 2. Given simulated data, we employ the algorithm, discussed in Section 2.1, to obtain $\hat{\boldsymbol{\mathcal{L}}}$ and $\hat{\boldsymbol{\mathcal{S}}}$. The regularization parameters λ_L and λ_S are selected by a two-dimensional grid search method. We run the algorithm and obtain the estimates for different grids of the pair (λ_L, λ_S) and select that pair for which the rank of $Circ(\hat{\boldsymbol{\mathcal{L}}})$ and the positions of the non-zero columns in $\hat{\boldsymbol{\mathcal{S}}}_{Mat}$ are as close as possible to the rank of $Circ(\boldsymbol{\mathcal{L}}^*)$ and the positions of the non-zero columns in $\boldsymbol{\mathcal{S}}^*_{Mat}$ 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 rank and sparsity level are unknown to us.

Performance Evaluation: We use Relative Error, Rank of $Circ(\hat{\boldsymbol{\mathcal{L}}})$, Sensitivity and Specificity as the criteria of evaluation. Small values of relative error, along with the closeness of rank of $Circ(\hat{\boldsymbol{\mathcal{L}}})$ and rank of $Circ(\boldsymbol{\mathcal{L}}^*)$, characterize the quality of the estimation. In addition to that, sensitivity and specificity together assess the ability of support recovery. Below we provide the rigorous definitions of these criteria.

1. Relative Error (RE): Considering the definition of Estimation Error provided in equation (5), the Relative Error is defined as $\frac{\|\hat{\boldsymbol{\mathcal{L}}}-\boldsymbol{\mathcal{L}}^*\|_F^2+\|\hat{\boldsymbol{\mathcal{S}}}-\boldsymbol{\mathcal{S}}^*\|_F^2}{\|\boldsymbol{\mathcal{L}}^*\|_F^2+\|\boldsymbol{\mathcal{S}}^*\|_F^2}$
2. Specificity (SP): Specificity is defined as 1− False Positive Rate (FPR), where, FPR is defined as follows:
$$\frac{\text{number of non-zero elements in } \hat{\boldsymbol{\mathcal{S}}}\text{, which are actually zero in } \boldsymbol{\mathcal{S}}^*}{\text{number of elements that are zero in } \boldsymbol{\mathcal{S}}^*}$$
3. Sensitivity (SN): Sensitivity, also known as True Positive Rate (TPR), which is defined as follows:
$$\frac{\text{number of non-zero elements in } \hat{\boldsymbol{\mathcal{S}}}\text{, which are actually non-zero in } \boldsymbol{\mathcal{S}}^*}{\text{number of non-zero elements in } \boldsymbol{\mathcal{S}}^*}$$

Using the above-mentioned criteria we evaluate the performance of our method under four different scenarios. Each scenario corresponds to specific values of the triplet (d_1, d_2, d_3). Furthermore, within each scenario, we obtain the estimates under four different sub-cases, where each sub-case corresponds to a particular combinations of the true tubal-rank and sparsity level. 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: $d_1 = 10, d_2 = 10, d_3 = 8$, Scenario 2: $d_1 = 20, d_2 = 20, d_3 = 8$, Scenario 3: $d_1 = 10, d_2 = 10, d_3 = 18$, Scenario 4: $d_1 = 20, d_2 = 20, d_3 = 18$
- Sub-case 1: $(r, s^*) = (2, 1)$, Sub-case 2: $(r, s^*) = (2, 2)$, Sub-case 3: $(r, s^*) = (3, 1)$, Sub-case 4: $(r, s^*) = (3, 2)$. In all of these cases, $\boldsymbol{\mathcal{L}}^*$ has been generated in such a way that the right-hand side of the Proposition A.1 follows with equality. More specifically, rank of the true block-circulant matrix is simply $r \times d_3$. As an example, in Table 1, since d_3 is 8, we will have R as 16 and 24 when r is 2 and 3 respectively.

As depicted in Table 1, in all four sub-cases, the relative error decrease as the sample size increases. Moreover, as the estimation is equipped with more and

TABLE 1

Performance Evaluation under Scenario 1 : $d_1 = 10, d_2 = 10, d_3 = 8$*; Relative Error, Rank of the estimated Block-Circulant matrix, Specificity and Sensitivity are reported for four different combinations of true rank and true sparsity level.*

	r = 2, s* = 1, R = 16				r = 2, s* = 2, R = 16			
Sample Size	RE	R	SP	SN	RE	R	SP	SN
400	0.48	21	0.90	1	0.62	19	0.94	1
800	0.33	17	0.99	1	0.39	17	0.94	1
1100	0.29	17	0.99	1	0.36	17	0.95	1

	r = 3, s* = 1, R = 24				r = 3, s* = 2, R = 24			
Sample Size	RE	R	SP	SN	RE	R	SP	SN
400	0.60	25	0.97	1	0.65	26	0.91	0.94
800	0.37	24	1	1	0.42	25	0.92	1
1100	0.31	24	1	1	0.38	22	1	1

Table 2

Performance Evaluation under Scenario 2 : $d_1 = 20, d_2 = 20, d_3 = 8$; Relative Error, Rank of the estimated Block-Circulant matrix, Specificity and Sensitivity are reported for four different combinations of true rank and true sparsity level.

	r = 2, s* = 1, R = 16				r = 2, s* = 2, R = 16			
Sample Size	RE	R	SP	SN	RE	R	SP	SN
400	0.50	17	0.93	1	0.65	19	0.90	1
800	0.35	18	0.95	1	0.40	18	0.91	1
1100	0.32	17	0.96	1	0.38	18	0.94	1

	r = 3, s* = 1, R = 24				r = 3, s* = 2, R = 24			
Sample Size	RE	R	SP	SN	RE	R	SP	SN
400	0.75	28	0.90	1	0.90	25	0.89	0.89
800	0.46	24	0.93	1	0.63	24	0.94	1
1100	0.37	25	1	1	0.53	24	0.97	1

Table 3

Performance Evaluation under Scenario 3 : $d_1 = 10, d_2 = 10, d_3 = 18$; Relative Error, Rank of the estimated Block-Circulant matrix, Specificity and Sensitivity are reported for four different combinations of true rank and true sparsity level.

	r = 2, s* = 1, R = 36				r = 2, s* = 2, R = 36			
Sample Size	RE	R	SP	SN	RE	R	SP	SN
500	0.77	37	0.86	1	0.81	37	0.86	0.90
900	0.62	35	0.92	1	0.66	36	0.92	0.97
1200	0.58	34	0.94	1	0.62	36	0.94	1

	r = 3, s* = 1, R = 54				r = 3, s* = 2, R = 54			
Sample Size	RE	R	SP	SN	RE	R	SP	SN
500	0.84	51	0.95	0.90	0.90	53	0.91	0.88
900	0.68	50	0.96	1	0.75	52	0.97	0.94
1200	0.65	50	0.96	1	0.72	50	0.97	0.94

more samples, it becomes easier to achieve the target rank of the true block-circulant matrix. Finally, the values of the specificity and sensitivity approaches to 1, with increase in the sample size. The reader may also note that, for a fixed value of true rank, when one increases the true sparsity level, the relative error increases, which is in accordance with the theoretical finding in Lemma 3.3. For example, with sample size 800 and tubal-rank 3 (Rank of the block-circulant matrix 24), when one increases s^* from 1 to 2, the relative error increases from 0.37 to 0.42. The same argument follows when the true rank is increased for a fixed level of imposed sparsity.

The remaining tables (Table 2, Table 3 and Table 4) display the results under the remaining three scenarios. 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.

TABLE 4

Performance Evaluation under Scenario 4 : $d_1 = 20, d_2 = 20, d_3 = 18$; Relative Error, Rank of the estimated Block-Circulant matrix, Specificity and Sensitivity are reported for four different combinations of true rank and true sparsity level.

	r = 2, s* = 1, R = 36				r = 2, s* = 2, R = 36			
Sample Size	RE	R	SP	SN	RE	R	SP	SN
500	0.81	38	0.87	0.90	0.84	38	0.86	0.92
900	0.68	37	0.92	1	0.69	38	0.92	0.97
1200	0.60	34	0.94	1	0.64	36	0.94	0.97

	r = 3, s* = 1, R = 54				r = 3, s* = 2, R = 54			
Sample Size	RE	R	SP	SN	RE	R	SP	SN
500	0.88	55	0.89	0.90	0.92	56	0.87	0.90
900	0.71	51	0.94	1	0.77	52	0.97	0.94
1200	0.67	50	0.97	1	0.74	51	0.97	0.94

TABLE 5

Positive predictive values and Negative predictive values for Scenario 1 : There are no cases of very low positive predictive values even if specificity and sensitivity values are high.

	r = 2, s* = 1, R = 16				r = 2, s* = 2, R = 16			
Sample Size	SP	SN	PPV	NPV	SP	SN	PPV	NPV
400	0.90	1	0.70	1	0.94	1	0.72	1
800	0.99	1	0.89	1	0.94	1	0.72	1
1100	0.99	1	0.89	1	0.95	1	0.74	1

	r = 3, s* = 1, R = 24				r = 3, s* = 2, R = 24			
Sample Size	SP	SN	PPV	NPV	SP	SN	PPV	NPV
400	0.97	1	0.84	1	0.91	0.94	0.71	0.99
800	1	1	1	1	0.92	1	0.71	1
1100	1	1	1	1	1	1	1	1

Remark 4.1. *There can be situations, where only good values of the specificity and sensitivity may not reveal the actual underlying scenario in terms of support recovery. For example, suppose there are 10 non-zeros and 9990 zero elements and an algorithm predicts 1000 non-zeros (including 10 true ones) and 9000 zero elements. Then the sensitivity is 1 and the specificity is about 0.9. However, the positive predictive value*[1] *is as low as 0.01 and the negative predictive value is 1. To address this point, we obtained the positive predictive values and negative predictive values in addition to the specificity and sensitivity. Table 5 provides the values for the Scenario-1. From the table, it can be seen that, both the positive predictive values and the negative predictive values are fairly good. Thus, it seems that, there are no cases of very low positive predictive value (as 0.01 in the example given above) even if specificity and sensitivity are high. For example, in sub-case 1 of Table 5 (upper left part) with sample size* 800, *we have* 80 *true non-zero values* (10 *true non-zero values in each of the* 8 *slices) and* 720

[1] see https://www.medcalc.org/calc/diagnostic_test.php

($= 800 - 80$) true zero values. Our method estimates the sparse component with 90 non-zero values (that means, 80 true non-zeros and 10 false positive) and no false negative. Thus, the specificity and sensitivity values are 0.99 and 1 respectively, along with positive predictive value as 0.89 and negative predictive value as 1.

Predictive Performance: To assess the out-of-sample predictive performance of our model, we split the data into two parts. While we fit the model based on the first part of the data (training data), the second part (test data) is used to assess the performance of the model. We use the Root Mean Square Error (RMSE) as the measure, which is defined as $\sqrt{\frac{\sum_{i=1}^{n_{test}}(y_i - \hat{y_i})^2}{n_{test}}}$, where n_{test} is the number of observations in the test data, y_i is the i^{th} actual observation in the test data and $\hat{y_i}$ is the fitted value, using the model based on the train data. Thus, lower RMSE values imply better performance. We compare our model with four benchmarks, which are relevant in the literature: 1) *Vectorized Lasso (**La-vec**)*: in this case, the third order tensor predictor $\boldsymbol{\mathcal{X}}_i \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ is vectorized and the resulting vector of length $d_1 d_2 d_3$ is then used to fit a Lasso regression with ℓ_1 norm penalization. 2) *Vectorized Elastic Net (**EN-vec**)*: this is similar to benchmark 1, except for the fact that here we employ the Elastic Net regularization on the vectorized $\boldsymbol{\mathcal{X}}$, instead of Lasso. 3) *Sparse CP regression (**Sp-CP**)*: this is based on the method developed in [42], that uses the CP decomposition of the coefficient tensor, that is, $\boldsymbol{\mathcal{B}} = \sum_{r=1}^{R} \beta_1^{(r)} \circ \beta_2^{(r)} \circ \beta_3^{(r)}$ and imposes ℓ_1 norm penalization on the components $\beta_n^{(r)}$ as $\sum_{r=1}^{R} \sum_{n=1}^{3} |\beta_n^{(r)}|$. [42] developed a Block Relaxation algorithm to solve the problem, which is implemented in a Matlab toolbox, `TensorReg`[2] and we use the toolbox to generate the results. 4) *Sparse Tucker regression (**Sp-Tu**)*: this is similar to benchmark 3, except for the fact that [21] applies a Tucker decomposition ([16]) on the coefficient tensor, instead of a CP decomposition. As in benchmark 3, in this case too, we use toolbox `TensorReg` in order to generate the results. Note that, competitors (1) and (2) are oblivious to the tensor nature of the problem and treat it as a large size regularized regression one.

Table 6 summarizes the RMSE values for our Low Tubal Rank model (Low TR) and for the four benchmarks, La-vec, EN-vec, Sp-CP and Sp-Tu. The first column specifies the true data generating procedure, wherein, the first four entries, TR($r = 3, s = 2$), TR($r = 3, s = 1$), TR($r = 2, s = 2$), and TR($r = 2, s = 1$) characterize the four sub-cases (based on the values of tubal rank and the number of non-zero columns) and the relevant data generating procedures discussed in Section 4 of our paper. The last entry of the first column, Sp-data, corresponds to the data generating procedure in the sparse CP tensor regression [42], which is also provided in their toolbox documentation. For each of the true data generating processes, Table 6 reports the RMSE values for our model and the benchmarks for three different sample sizes $n = 400, 800, 1100$. While calculating the RMSEs, the number of test data, n_{test} was taken as 100 and the sizes d_1, d_2 and d_3 of the tensor predictor were fixed at $10, 10$ and 8

[2] https://hua-zhou.github.io/TensorReg/

Table 6
RMSE values for our model (Low TR) and the four benchmarks, La-vec, EN-Vec, Sp-CP and Sp-Tu, under different true data generating procedures

		Low TR	*La-vec*	*En-vec*	*Sp-CP*	*Sp-TU*
TR(r=3, s=2)	$n = 400$	4.45	5.11	5.16	7.41	7.45
	$n = 800$	3.87	5.07	5.11	6.71	6.79
	$n = 1100$	3.26	4.86	4.89	5.24	5.27
TR(r=3, s=1)	$n = 400$	4.32	5.04	5.12	6.57	6.59
	$n = 800$	3.39	4.88	4.92	5.83	5.87
	$n = 1100$	2.66	3.35	3.52	4.76	4.83
TR(r=2, s=2)	$n = 400$	4.29	5.03	5.09	6.55	6.58
	$n = 800$	3.37	4.65	4.73	5.79	5.81
	$n = 1100$	2.58	3.27	3.43	4.68	4.75
TR(r=2, s=1)	$n = 400$	3.98	4.93	4.96	6.22	6.31
	$n = 800$	3.21	4.32	4.41	5.44	5.48
	$n = 1100$	2.31	3.12	3.17	4.22	4.25
Sp-data	$n = 400$	2.71	3.32	3.38	2.35	2.37
	$n = 800$	1.85	2.23	2.31	1.96	1.98
	$n = 1100$	1.33	2.12	2.19	1.72	1.78

respectively. As depicted in Table 6, our model outperforms the benchmarks for the first four data generating procedures, as expected, since the posited low tubal rank plus sparse model corresponds to the true data generating mechanism. For the Sp-data generating mechanism, though the RMSE values from our model are initially slightly higher or on par with the Sp-CP and Sp-Tu ones (that are in accordance to the true generating mechanism), they start getting smaller than Sp-CP and Sp-Tu, as we increase the sample size. This is probably due to the fact that the optimization problem for our model is convex, as opposed to the sparse CP/Tucker decomposition and also the small number of parameters to be estimated. Both these features are advantageous in settings with not enormous sample sizes.

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

AIC: We define AIC as $n\log(\frac{RSS}{n}) + 2\ \text{Rank}\ (Circ(\hat{\mathcal{L}})) + 2k$, where, $RSS = \sum_{i=1}^{n}(y_i - \langle\hat{\mathcal{L}} + \hat{\mathcal{S}}, \mathcal{X}_i\rangle)^2$ and k is the number of non-zero elements in $\hat{\mathcal{S}}$. This formulation is quite common in the literature, which essentially rewards goodness of fit and at the same time penalizes overfitting. Below we provide a numerical analysis that justifies the performance of the posited AIC criterion.

We consider the synthetic data generated in scenario 1 - sub-case 1. Recall that for this data set, the true rank of the block-circulant matrix is 16 and there is only one non-zero column in each frontal slice of the sparse component. The goal of this experiment is to check whether the values of rank and sparsity that we get after minimizing AIC, match closely to the true rank and sparsity level or not. To that end, we obtain the values of AIC for different grids of the pair (λ_L, λ_S) with sample size $n = 800$. Figure 5 depicts the relevant part of the grids that contains the minimum AIC value (see additional tables in Appendix D for the AIC values). The pair (λ_L, λ_S) corresponding to this minimum AIC in

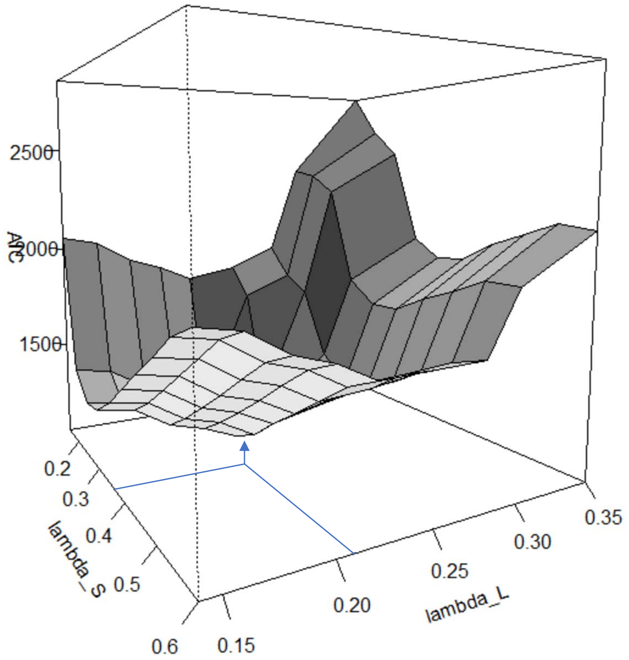


Fig 5. *AIC plot: Rank and Sparsity corresponding to minimum AIC match well with the truth*

Figure 5, produces $\hat{\mathcal{L}}$ with R = 17 ans $\hat{\mathcal{S}}$ comprising a frontal slice with two non-zero columns and remaining frontal slices with only one non-zero column in the desired positions. Hence, the rank and sparsity, decided by AIC, matches quite well the true values. Also, these AIC based rank and sparsity pattern are exactly in line with the ones obtained in the simulations (see the rank, SP and SN for $n = 800$ in scenario 1-subcase 1, depicted in Table 1). To gain more assurance on the parity between AIC based results and the simulation results obtained earlier, we recalculate the estimation results for all the sub-cases of Scenario 1, with the AIC based tuning parameters and summarize them in Table 7. As it can be seen, the AIC based results in Table 7 are quite in line with the scenario 1 simulation results in Table 1.

5. Application to educational data

In this section, we use our proposed method on educational data from an Intelligent Tutoring System (ITS). The ITS under consideration is an online video based tutoring program launched in year 2013 to help prepare students for an End-of-Course basic algebra test. The platform offers videos on various algebra topics, recorded by different tutors. The students can assess their progress by taking practice test. Also, the platform offers a monitored discussion area where the students can pose questions to peers and volunteer tutors. The data

Table 7
Performance Evaluation under Scenario 1 with the AIC based tuning parameters

	r = 2, s* = 1, R = 16				r = 2, s* = 2, R = 16			
Sample Size	RE	R	SP	SN	RE	R	SP	SN
400	0.46	22	0.90	1	0.59	18	0.95	1
800	0.35	17	0.99	1	0.39	17	0.94	1
1100	0.28	16	0.99	1	0.37	17	0.95	1

	r = 3, s* = 1, R = 24				r = 3, s* = 2, R = 24			
Sample Size	RE	R	SP	SN	RE	R	SP	SN
400	0.61	25	0.96	1	0.65	26	0.91	0.94
800	0.37	24	1	1	0.42	24	0.94	1
1100	0.32	24	1	1	0.39	22	1	1

we consider in this section, consists of records of the students for four consecutive academic years, starting from 2014-15 to 2017-18. Students who logged in the tutoring platform for at least five times in a particular academic year, were considered as users of the platform for that year. For each year, we gather information on the following variables for each user.

- Socioeconomic variables: 1) Ethnicity: Hispanic/Latino or not, 2) FRL: Reduced-price (or Free) meals at schools or not, 3) Gender: Male or Female
- Score in Maths Tests: 1) Pre-Score: Score in state standard assessment maths test, that determines the maths preparedness of the students, 2) EoC Score: Score in the end of course maths test. The students must take and pass this test to establish their maths proficiency.
- Platform Usage variables: 1) Video: Number of videos watched by the user, 2) TYS: Number of "Test yourself" questions completed by the user, 3) Logins: Number of times the user has logged into the platform, 4) Wall Post: The number of posted comments on the discussion wall by the user.

In the next step, we convert this user level data into school level. To that end, for each of the schools, we process the data as follows:

- For each academic year, we first make three categories of the teachers, based on their overall teaching experience, namely E1, E2 and E3. While E1 is the group of teachers with the least experience (teaching experience of at most 5 years), E3 is the most experienced group (at least 14 years of teaching experience).
- Thus, for each school, we arrive at a third-order tensor predictor of dimension $8 \times 4 \times 3$ as illustrated in Figure 6. The four lateral slices correspond to four academic years, where in each lateral slice, 8 variables are captured across the 3 levels of teaching experience. It is worth mentioning that these variables are now measured at school level, by averaging over the relevant student level data. As an example, the very first element of 2014-15 slice represents the average number of videos watched by the students of that

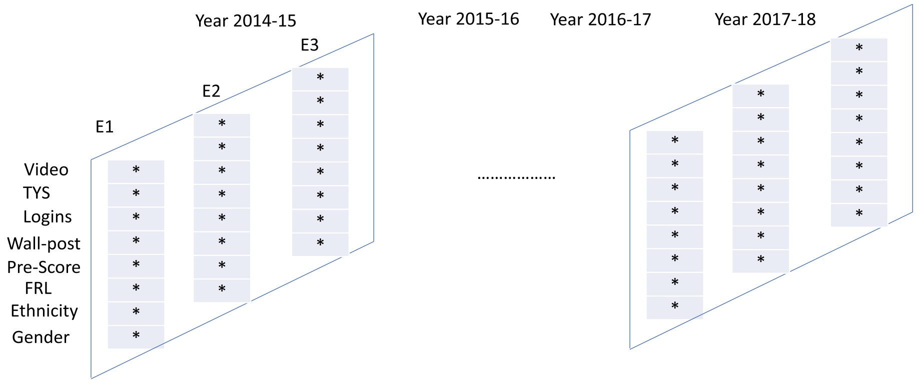


Fig 6. *Structure of the tensor predictor $\mathcal{X}_i$ for i^{th} school: Variables, Academic Year and Teaching Experience are the three dimensions*

particular school in that year, who were taught by the least experienced group of teachers. All the other usage variables along with Pre-Score have similar interpretation. For any cell, the variables Female-pr, FRL-pr and Eth-pr are defined as the proportion of the female students, proportion of the students that avails free meals at school and the proportion of the Hispanic/Latino students in that cell.

- Finally, the response variable is the average EoC score obtained by the students of that school in 2017-18. Specifically, for $i = 1, 2, \cdots, n$, y_i, the average EoC score by the students of the i^{th} school in 2017-18, is the scalar response variable corresponding to the i^{th} school. On the other hand, the $(l, m, k)^{th}$ element of the tensor predictor $\mathcal{X}_i \in \mathbb{R}^{8\times4\times3}$ captures the value of the l^{th} variable, in the m^{th} year, with the k^{th} level of teaching experience experience, $l = 1, 2, \cdots, 8$, $m = 1, 2, 3, 4$ and $k = 1, 2, 3$.

Due to lack of information for many schools, we restrict our analysis to $n = 38$ schools. Using the observed data $\{(y_i, \mathcal{X}_i)\}_{i=1}^n$, we fit the tensor regression model, given by (2). We assume that $\mathcal{B} = \mathcal{L} + \mathcal{S}$, where $\mathcal{L}$ is the low tubal-rank component and $\mathcal{S}$ is a tensor, whose frontal slices are elementwise sparse. In other words, $\mathcal{L}$ captures the baseline effects, which are shared across the academic years. In addition to that, there are some specific combinations of variables and experience level, which may show additional effect in some specific year. The sparse component $\mathcal{S}$ is devoted to determining such additional effects.

We first select suitable values of λ_L and λ_S using the AIC criteria discussed in Section 4 and then employ the Alternating Block Minimization Algorithm (Algorithm 1) to estimate the low tubal rank and sparse components of the coefficient tensor. Figure 7 provides the estimated low tubal-rank component, for which the tubal-rank is 1. Thus, there is only one academic year, for which the corresponding lateral slice is t-linearly independent. In all the remaining academic years, the effects of the variables are t-linearly dependent

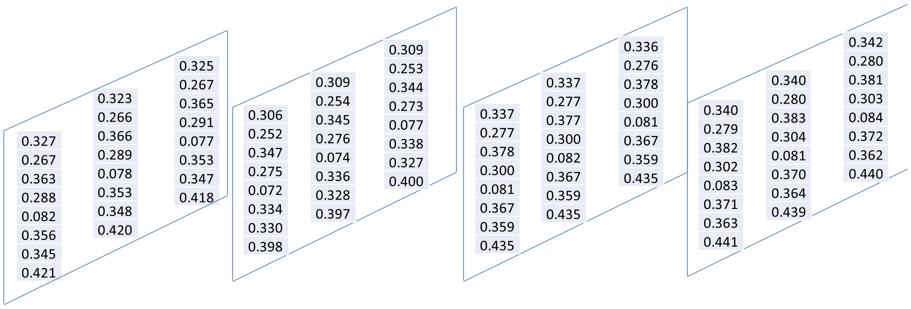


FIG 7. *Estimate of Low Tubal-Rank Component: Four slices are for four academic years. In each slice, three columns correspond to three different level of teaching experience. In each column, the variables are in the following order: Video, TYS, Logins, Wall-post, Pre-Score, FRL-pr, Eth-pr, Female-pr*

on the former one. In fact, for most of the variables, the effects are quite similar across the years. For instance, the estimated coefficients for the variable Pre-Score (across three levels of teaching experience) for different academic years are $(0.082, 0.078, 0.077)^T$, $(0.072, 0.074, 0.077)^T$, $(0.081, 0.082, 0.081)^T$ and $(0.083, 0.081, 0.084)^T$. Prior work on different models based on this data have also portrayed similar behavior of the Pre-Score coefficients. In the context of *Causal Invariance Prediction*, [30] analyzed an educational attainment data and found that the effects of the school students' prior scores on their BA degree attainment, are similar across two different "experimental" groups. The first group corresponds to the students who live within 10 miles of the nearest 4-year college. On the other hand, the second group of students live at least 10 miles away from the nearest 4-year college. Similarly in our study, the effect of students' preparedness or prior knowledge, on their future academic achievements, are similar across different years. In addition to the Pre-Score or prior knowledge, the effects of the socioeconomic variables are also similar for different years.

However, as opposed to the Pre-Score and socioeconomic variables, the effects of the ITS platform usage variables are not quite similar across the years. There are some usage variables for which some additional effects are captured by the estimated sparse component $\hat{\mathbf{S}}$. The binary heatmap in Figure 8 depicts such variables for which there are non-zero values in $\hat{\mathbf{S}}$, whereas the actual values of the estimates are tabulated in Appendix (Table 9). As shown in Figure 8, the two key variables for which additional effects are present, are the number of videos viewed and the number of logins to the platform. Although it is difficult to discover the ground truth behind this, intuition suggests that the level of platform usage may vary significantly across different segments and thus leads to additional effects in the estimated coefficient. For instance, as the platform gains more and more popularity with time, more students are expected to get acquainted with the platform and thus watch tutorial videos. Consequently, additional effects of videos are expected to become more prominent in Year-3 and Year-4, as compared to Year-1 and Year-2, which is reflected in the heatmap.

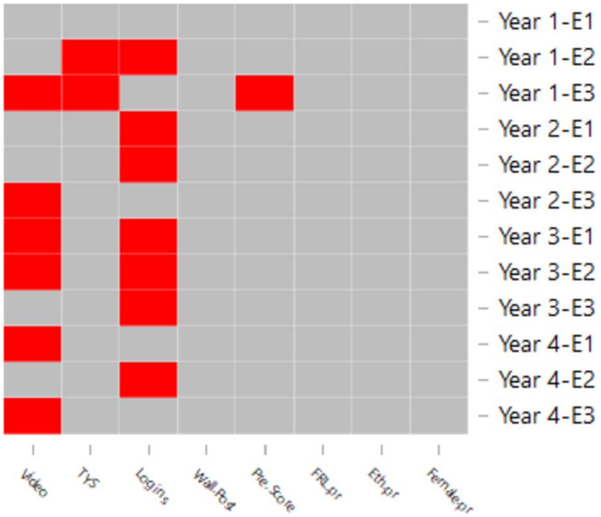


Fig 8. *Binary Heatmap of the estimate of sparse component: The combinations that are colored in red, have non-zero coefficients. All the remaining coefficients are zeros*

Regarding the other tensor regression models, as mentioned earlier, our work is the first one to explore the decomposition of the total effect into baseline (low tubal-rank tensor) and idiosyncratic effect (sparse tensor). This decomposition is a necessary intrinsic pattern of the type of educational data that we consider here. The effects of prior knowledge (Pre-Score) and socioeconomic variables are similar or shared across the years ([30]), which is the baseline component. In addition to that, the platform usage variables (number of logins to the platform, number of tutorial videos watched and so on) display some additional effects as the platform gains more and more popularity. No other tensor regression models in the literature have developed an algorithm to estimate such decomposed effects. As an example, we apply the Sparse CP regression model proposed by [42] to our data. As previously discussed, a sparse CP regression first uses CP decomposition to represent the coefficient tensor and then imposes sparsity assumption on the components of the CP decomposition. The sole purpose of this method is to reduce the dimensionality and it provides a sparse estimate of the coefficient tensor. Figure 9 depicts the heatmap of the estimated sparse coefficient tensor (represented in matrix form) using sparse-CP regression. As it can be seen, the sparsity pattern in the estimated coefficient tensor is random and hardly reveals any meaningful interpretation of the underlying effects.

6. Discussion

In this paper, we propose a tensor regression model with scalar response and third-order tensor predictor and assume that the third-order coefficient tensor is decomposed into two components. The first one corresponds to a low tubal rank

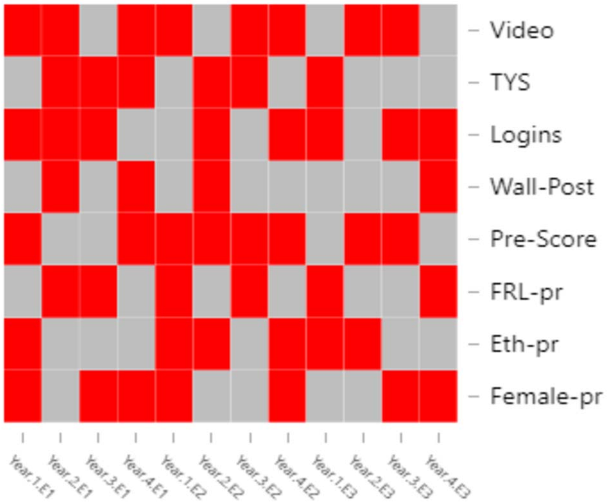


Fig 9. *Binary heatmap of the sparse CP estimate*

tensor, which captures baseline effects, shared across the lateral (or, horizontal) slices. On the other hand, the second component is a third-order tensor, whose frontal slices are either elementwise or columnwise sparse. The role of the sparse component is to capture the additional idiosyncratic effects. This decomposition of the coefficient tensor, as opposed to the Canonical Polyadic (CP) decomposition used in related literature, expands the scope of exploring more complex structure 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 extend the work in the literature of multivariate regression [1] to third-order tensor and establish a non-asymptotic interpretable upper bound to the estimation error. The efficacy of the methodology is illustrated on synthetic and real education data.

7. 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.

Appendix A: Proofs

In this section, we prove the results presented in Section 3. We start by establishing a simple proposition, followed by Lemmas 3.1 and 3.2 and then a simple inequality, termed as the, *Basic Inequality*. Based on these results we then prove

Lemma 3.3, which provides an upper bound to the estimation error in case of deterministic noise and predictors. Finally, using Lemma 3.3, we prove Theorem 3.4 and Corollary 3.4.1, which provide the upper bound in case of random realizations of errors and predictors.

Proposition A.1. $r \leq R \leq rd_3$

Proof. Note that, a block-circulant matrix of a third-order tensor in $\mathbb{R}^{d_1 \times d_2 \times d_3}$, can be expressed as $[B_1|B_2|\cdots|B_{d_2}]$, where the j^{th} block B_j is a matrix of dimension $d_1d_3 \times d_3$, $j = 1, 2, \cdots d_2$. In each B_j, the first column is the vectorized version of the j^{th} lateral slice of the tensor and the remaining $(d_3 - 1)$ 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 d_2 blocks such that:

- any column of any of the remaining $d_2 - r$ 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 $r \times d_3$ 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. □

Proof of Lemma 3.1

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 [33], it is possible to decompose $Circ(\hat{\Delta}_{\mathcal{L}})$ as $Circ(\hat{\Delta}^A_{\mathcal{L}}) + Circ(\hat{\Delta}^B_{\mathcal{L}})$, such that, Rank$(Circ(\hat{\Delta}^A_{\mathcal{L}})) \leq 2$ Rank $(Circ(\boldsymbol{\mathcal{L}}^*)) = 2R$ and $Circ(\boldsymbol{\mathcal{L}}^*)^T Circ(\hat{\Delta}^B_{\mathcal{L}}) = 0$, $Circ(\boldsymbol{\mathcal{L}}^*) Circ(\hat{\Delta}^B_{\mathcal{L}})^T = 0$. The reader may visit [33] to know the details on how to derive such decomposition. It is worth mentioning that, [1] uses the same tool while proving their Lemma 1. However, as Lemma 2.3 of [33] proves, the last two equalities are essentially the sufficient condition of the additivity of nuclear norm. In other words, these imply

$$\left\|Circ(\boldsymbol{\mathcal{L}}^*) + Circ(\hat{\Delta}^B_{\mathcal{L}})\right\|_* = \|Circ(\boldsymbol{\mathcal{L}}^*)\|_* + \left\|Circ(\hat{\Delta}^B_{\mathcal{L}})\right\|_* \tag{15}$$

We use the above finding later in this proof. It now remains to show that inequality (6) holds for such decomposition. Note that,

$$C(\boldsymbol{\mathcal{L}}^* + \hat{\Delta}_{\mathcal{L}}, \boldsymbol{\mathcal{S}}^* + \hat{\Delta}_{\mathcal{S}})$$

$$= \frac{1}{d_3} \left\|Circ(\boldsymbol{\mathcal{L}}^*) + Circ(\hat{\Delta}_{\mathcal{L}})\right\|_* + \frac{\lambda_S}{\lambda_L} \left\|\boldsymbol{\mathcal{S}}^*_{Mat} + \hat{\Delta}_{\mathcal{S}Mat}\right\|_{2,1},$$

by the definition of $C(\boldsymbol{\mathcal{L}}, \boldsymbol{\mathcal{S}})$ and the fact that $Circ(\cdot)$ is additive

$$=\frac{1}{d_3}\left\|Circ(\boldsymbol{\mathcal{L}}^*)+Circ(\hat{\Delta}^A_{\boldsymbol{\mathcal{L}}})+Circ(\hat{\Delta}^B_{\boldsymbol{\mathcal{L}}})\right\|_* + \frac{\lambda_S}{\lambda_L}\left\|\boldsymbol{\mathcal{S}}^*_{Mat}+\hat{\Delta}^{\mathbb{M}}_{\boldsymbol{\mathcal{S}}Mat}+\hat{\Delta}^{\mathbb{M}^\perp}_{\boldsymbol{\mathcal{S}}Mat}\right\|_{2,1},$$

by the aforementioned decomposition and the property of projection

$$\geq\frac{1}{d_3}\left\|Circ(\boldsymbol{\mathcal{L}}^*)+Circ(\hat{\Delta}^B_{\boldsymbol{\mathcal{L}}})\right\|_* - \frac{1}{d_3}\left\|Circ(\hat{\Delta}^A_{\boldsymbol{\mathcal{L}}})\right\|_*$$
$$+\frac{\lambda_S}{\lambda_L}\left\|\boldsymbol{\mathcal{S}}^*_{Mat}+\hat{\Delta}^{\mathbb{M}^\perp}_{\boldsymbol{\mathcal{S}}Mat}\right\|_{2,1} - \frac{\lambda_S}{\lambda_L}\left\|\hat{\Delta}^{\mathbb{M}}_{\boldsymbol{\mathcal{S}}Mat}\right\|_{2,1},$$

by the Triangle Inequality

$$\geq\frac{1}{d_3}\left\|Circ(\boldsymbol{\mathcal{L}}^*)\right\|_* + \frac{1}{d_3}\left\|Circ(\hat{\Delta}^B_{\boldsymbol{\mathcal{L}}})\right\|_* - \frac{1}{d_3}\left\|Circ(\hat{\Delta}^A_{\boldsymbol{\mathcal{L}}})\right\|_*$$
$$+\frac{\lambda_S}{\lambda_L}\left\|\boldsymbol{\mathcal{S}}^*_{Mat}\right\|_{2,1} + \frac{\lambda_S}{\lambda_L}\left\|\hat{\Delta}^{\mathbb{M}^\perp}_{\boldsymbol{\mathcal{S}}Mat}\right\|_{2,1} - \frac{\lambda_S}{\lambda_L}\left\|\hat{\Delta}^{\mathbb{M}}_{\boldsymbol{\mathcal{S}}Mat}\right\|_{2,1},$$

by Equation (15) and the Decomposability of $\|\cdot\|_{2,1}$

The remainder of the proof follows from the above inequality and the definition of $C(\boldsymbol{\mathcal{L}}^*, \boldsymbol{\mathcal{S}}^*)$. □

Proof of Lemma 3.2

Proof. We start by defining a function $f : \mathbb{R}^{d_1\times d_2\times d_3} \longrightarrow \mathbb{R}$ as follows:

$$\begin{aligned} f(\Delta_{\boldsymbol{\mathcal{L}}}, \Delta_{\boldsymbol{\mathcal{S}}}) = {} & L(\boldsymbol{\mathcal{L}}^*+\Delta_{\boldsymbol{\mathcal{L}}}, \boldsymbol{\mathcal{S}}^*+\Delta_{\boldsymbol{\mathcal{S}}}) - L(\boldsymbol{\mathcal{L}}^*, \boldsymbol{\mathcal{S}}^*) \\ & + \lambda_L\{C(\boldsymbol{\mathcal{L}}^*+\Delta_{\boldsymbol{\mathcal{L}}}, \boldsymbol{\mathcal{S}}^*+\Delta_{\boldsymbol{\mathcal{S}}}) - C(\boldsymbol{\mathcal{L}}^*, \boldsymbol{\mathcal{S}}^*)\} \end{aligned} \tag{16}$$

wherein as before, $L(\boldsymbol{\mathcal{L}}, \boldsymbol{\mathcal{S}})$ is used to denote the loss function given by, $\frac{1}{2n}\sum_{i=1}^n (y_i - \langle \boldsymbol{\mathcal{L}}+\boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{X}}_i\rangle)^2$. Since $f(0,0) = 0$ and $(\hat{\Delta}_{\boldsymbol{\mathcal{L}}}, \hat{\Delta}_{\boldsymbol{\mathcal{S}}})$ is the optimal error, one must have, $f(\hat{\Delta}_{\boldsymbol{\mathcal{L}}}, \hat{\Delta}_{\boldsymbol{\mathcal{S}}}) \leq f(0,0) = 0$. Recall that, we already have established a lower bound of $C(\boldsymbol{\mathcal{L}}^*+\hat{\Delta}_{\boldsymbol{\mathcal{L}}}, \boldsymbol{\mathcal{S}}^*+\hat{\Delta}_{\boldsymbol{\mathcal{S}}}) - C(\boldsymbol{\mathcal{L}}^*, \boldsymbol{\mathcal{S}}^*)$ from equation (6). Now our job is to find a lower bound to $L(\boldsymbol{\mathcal{L}}^*+\hat{\Delta}_{\boldsymbol{\mathcal{L}}}, \boldsymbol{\mathcal{S}}^*+\hat{\Delta}_{\boldsymbol{\mathcal{S}}}) - L(\boldsymbol{\mathcal{L}}^*, \boldsymbol{\mathcal{S}}^*)$. These two bounds, along with the fact that $f(\hat{\Delta}_{\boldsymbol{\mathcal{L}}}, \hat{\Delta}_{\boldsymbol{\mathcal{S}}}) \leq 0$, will prove the result.

Since $\lambda_L \frac{1}{d_3}\left\|Circ(\boldsymbol{\mathcal{L}}^*)\right\|_* + \lambda_S \left\|\boldsymbol{\mathcal{S}}^*_{Mat}\right\|_{2,1} = \lambda_L C(\boldsymbol{\mathcal{L}}^*, \boldsymbol{\mathcal{S}}^*)$, one can think C as an alternative regularizer and λ_L as the associated parameter for our problem. Now as [28] derives while proving their Lemma 1, using the convexity of the loss function and dual-norm inequality, we get the following:

$$L(\boldsymbol{\mathcal{L}}^*+\hat{\Delta}_{\boldsymbol{\mathcal{L}}}, \boldsymbol{\mathcal{S}}^*+\hat{\Delta}_{\boldsymbol{\mathcal{S}}}) - L(\boldsymbol{\mathcal{L}}^*, \boldsymbol{\mathcal{S}}^*) \geq -C^*(\nabla L(\boldsymbol{\mathcal{L}}^*, \boldsymbol{\mathcal{S}}^*))C(\hat{\Delta}_{\boldsymbol{\mathcal{L}}}, \hat{\Delta}_{\boldsymbol{\mathcal{S}}}) \tag{17}$$

Where, C^* is the dual norm associated with the regularizer C. It is easy to check that $\nabla L(\boldsymbol{\mathcal{L}}^*, \boldsymbol{\mathcal{S}}^*) = [-\boldsymbol{\mathcal{D}}, -\boldsymbol{\mathcal{D}}]$. Now, from the given conditions on the regularizer parameters, we get $\frac{\|\boldsymbol{\mathcal{D}}_{Mat}\|_{2,\infty}}{\lambda_S} \leq \frac{1}{4}$ and $\frac{1}{d_3}\left\|Circ(\boldsymbol{\mathcal{D}})\right\|_{sp} \leq \frac{\lambda_L}{4}$ and hence using the similar argument as in the proof of Lemma 1 in [1], $C^*(\nabla L(\boldsymbol{\mathcal{L}}^*, \boldsymbol{\mathcal{S}}^*))$ can be shown to be bounded above by $\frac{\lambda_L}{2}$. Also, it is easy to check that $C(\hat{\Delta}_{\mathcal{L}}, \hat{\Delta}_{\mathcal{S}}) \leq C(\hat{\Delta}^A_{\mathcal{L}}, \hat{\Delta}^{\mathbb{M}}_{\mathcal{S}}) + C(\hat{\Delta}^B_{\mathcal{L}}, \hat{\Delta}^{\mathbb{M}^\perp}_{\mathcal{S}})$. Thus (17) reduces to

$$L(\mathcal{L}^* + \hat{\Delta}_{\mathcal{L}}, \mathcal{S}^* + \hat{\Delta}_{\mathcal{S}}) - L(\mathcal{L}^*, \mathcal{S}^*) \geq -\frac{\lambda_L}{2}(C(\hat{\Delta}^A_{\mathcal{L}}, \hat{\Delta}^{\mathbb{M}}_{\mathcal{S}}) + C(\hat{\Delta}^B_{\mathcal{L}}, \hat{\Delta}^{\mathbb{M}^\perp}_{\mathcal{S}})) \quad (18)$$

Finally, the rest of the proof follows simply from (6),(16),(18) and from the fact that $f(\hat{\Delta}_{\mathcal{L}}, \hat{\Delta}_{\mathcal{S}}) \leq 0$. □

Basic Inequality

$$\begin{aligned}&\frac{1}{2n}\sum_{i=1}^n \{\langle \hat{\Delta}_{\mathcal{L}} + \hat{\Delta}_{\mathcal{S}}, \mathcal{X}_i\rangle\}^2 \\ &\quad \leq \frac{1}{n}\sum_{i=1}^n \epsilon_i \langle \hat{\Delta}_{\mathcal{L}} + \hat{\Delta}_{\mathcal{S}}, \mathcal{X}_i\rangle + \lambda_L C(\mathcal{L}^*, \mathcal{S}^*) - \lambda_L C(\mathcal{L}^* + \hat{\Delta}_{\mathcal{L}}, \mathcal{S}^* + \hat{\Delta}_{\mathcal{S}}) \quad (19)\end{aligned}$$

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

$$\begin{aligned}&\frac{1}{2n}\sum_{i=1}^n (y_i - \langle \hat{\mathcal{L}} + \hat{\mathcal{S}}, \mathcal{X}_i\rangle)^2 + \lambda_L \frac{1}{d_3}\left\|Circ(\hat{\mathcal{L}})\right\|_* + \lambda_S \left\|\hat{\mathcal{S}}_{Mat}\right\|_{2,1} \\ &\quad \leq \frac{1}{2n}\sum_{i=1}^n (y_i - \langle \mathcal{L}^* + \mathcal{S}^*, \mathcal{X}_i\rangle)^2 + \lambda_L \frac{1}{d_3}\left\|Circ(\mathcal{L}^*)\right\|_* + \lambda_S \left\|\mathcal{S}^*_{Mat}\right\|_{2,1} \quad (20)\end{aligned}$$

Next, from $y_i = \langle \mathcal{L}^* + \mathcal{S}^*, \mathcal{X}_i\rangle + \epsilon_i$, we will have,

$$\begin{aligned}&\sum_{i=1}^n \{y_i - \langle \hat{\mathcal{L}} + \hat{\mathcal{S}}, \mathcal{X}_i\rangle\}^2 \\ =&\sum_{i=1}^n \{y_i - \langle \mathcal{L}^* + \mathcal{S}^*, \mathcal{X}_i\rangle - \langle \hat{\mathcal{L}} + \hat{\mathcal{S}}, \mathcal{X}_i\rangle + \langle \mathcal{L}^* + \mathcal{S}^*, \mathcal{X}_i\rangle\}^2 \\ =&\sum_{i=1}^n \{\epsilon_i - \langle \hat{\Delta}_{\mathcal{L}} + \hat{\Delta}_{\mathcal{S}}, \mathcal{X}_i\rangle\}^2 \\ =&\sum_{i=1}^n \epsilon_i^2 + \sum_{i=1}^n \langle \hat{\Delta}_{\mathcal{L}} + \hat{\Delta}_{\mathcal{S}}, \mathcal{X}_i\rangle^2 - 2\sum_{i=1}^n \epsilon_i \langle \hat{\Delta}_{\mathcal{L}} + \hat{\Delta}_{\mathcal{S}}, \mathcal{X}_i\rangle\end{aligned}$$

Using the above decomposition along with (20), we arrive at the following inequality:

$$\begin{aligned}&\frac{1}{2n}\sum_{i=1}^n \epsilon_i^2 + \frac{1}{2n}\sum_{i=1}^n \langle \hat{\Delta}_{\mathcal{L}} + \hat{\Delta}_{\mathcal{S}}, \mathcal{X}_i\rangle^2 - \frac{1}{n}\sum_{i=1}^n \epsilon_i \langle \hat{\Delta}_{\mathcal{L}} + \hat{\Delta}_{\mathcal{S}}, \mathcal{X}_i\rangle + \lambda_L \frac{1}{d_3}\left\|Circ(\hat{\mathcal{L}})\right\|_* \\ &\quad + \lambda_S \left\|\hat{\mathcal{S}}_{Mat}\right\|_{2,1} \leq \frac{1}{2n}\sum_{i=1}^n \epsilon_i^2 + \lambda_L \frac{1}{d_3}\left\|Circ(\mathcal{L}^*)\right\|_* + \lambda_S \left\|\mathcal{S}^*_{Mat}\right\|_{2,1} \quad (21)\end{aligned}$$

This compeltes the proof of the Lemma. □

Proof of Lemma 3.3

Proof. To avoid complex notations, in this proof, we initially ignore the term d_3 in the definition of $C(\boldsymbol{\mathcal{L}}, \boldsymbol{\mathcal{S}})$ and adjust that later towards the end of the proof. The reader may note that Lemma 3.1, Lemma 3.2 and Basic Inequality hold good with this modification, where the earlier assumption $\lambda_L \geq 4\frac{1}{d_3} \|Circ(\boldsymbol{\mathcal{D}})\|_{sp}$ is now replaced by $\lambda_L \geq 4 \|Circ(\boldsymbol{\mathcal{D}})\|_{sp}$.

Using the Assumption 1 and (8), we get

$$\frac{1}{2n}\sum_{i=1}^{n}\{\langle \hat{\Delta}_{\boldsymbol{\mathcal{L}}} + \hat{\Delta}_{\boldsymbol{\mathcal{S}}}, \boldsymbol{\mathcal{X}}_i\rangle\}^2 \geq \frac{\gamma}{2}\left\|\hat{\Delta}_{\boldsymbol{\mathcal{L}}} + \hat{\Delta}_{\boldsymbol{\mathcal{S}}}\right\|_F^2 \tag{22}$$

We will obtain a lower bound of the right-hand side and an upper bound of the left-hand side of the above inequality. We first start with deriving a lower bound for $\left\|\hat{\Delta}_{\boldsymbol{\mathcal{L}}} + \hat{\Delta}_{\boldsymbol{\mathcal{S}}}\right\|_F^2$. It is easy to check that

$$\frac{\gamma}{2}(\left\|\hat{\Delta}_{\boldsymbol{\mathcal{L}}}\right\|_F^2 + \left\|\hat{\Delta}_{\boldsymbol{\mathcal{S}}}\right\|_F^2) - \frac{\gamma}{2}\left\|\hat{\Delta}_{\boldsymbol{\mathcal{L}}} + \hat{\Delta}_{\boldsymbol{\mathcal{S}}}\right\|_F^2 = -\gamma\langle \hat{\Delta}_{\boldsymbol{\mathcal{L}}}, \hat{\Delta}_{\boldsymbol{\mathcal{S}}}\rangle \tag{23}$$

It can be easily seen that,

$$\begin{aligned}
&\gamma|\langle \hat{\Delta}_{\boldsymbol{\mathcal{L}}}, \hat{\Delta}_{\boldsymbol{\mathcal{S}}}\rangle| \\
=&\gamma|\langle\langle MatVec(\hat{\Delta}_{\boldsymbol{\mathcal{L}}}), MatVec(\hat{\Delta}_{\boldsymbol{\mathcal{S}}})\rangle\rangle| \\
\leq&\gamma\left\|MatVec(\hat{\Delta}_{\boldsymbol{\mathcal{L}}})\right\|_{2,\infty}\left\|MatVec(\hat{\Delta}_{\boldsymbol{\mathcal{S}}})\right\|_{2,1}, \text{ using Dual-Norm Inequality} \\
=&\gamma\left\|Circ(\hat{\Delta}_{\boldsymbol{\mathcal{L}}})\right\|_{2,\infty}\left\|MatVec(\hat{\Delta}_{\boldsymbol{\mathcal{S}}})\right\|_{2,1}, \text{ since,} \\
&\left\|MatVec(\hat{\Delta}_{\boldsymbol{\mathcal{L}}})\right\|_{2,\infty} = \left\|Circ(\hat{\Delta}_{\boldsymbol{\mathcal{L}}})\right\|_{2,\infty} \\
\leq&\gamma\{\left\|Circ(\hat{\boldsymbol{\mathcal{L}}})\right\|_{2,\infty} + \|Circ(\boldsymbol{\mathcal{L}}^*)\|_{2,\infty}\}\left\|MatVec(\hat{\Delta}_{\boldsymbol{\mathcal{S}}})\right\|_{2,1} \\
\leq&\frac{2\gamma\alpha}{\sqrt{d_2}}\left\|MatVec(\hat{\Delta}_{\boldsymbol{\mathcal{S}}})\right\|_{2,1}, \text{ using Assumption 2} \\
\leq&\frac{2\gamma\alpha}{\sqrt{d_2}}\left\|\hat{\Delta}_{\boldsymbol{\mathcal{S}}Mat}\right\|_{2,1} \\
\leq&\frac{\lambda_S}{2}\left\|\hat{\Delta}_{\boldsymbol{\mathcal{S}}Mat}\right\|_{2,1}, \text{ using Assumption 3}
\end{aligned}$$

Hence, from (23) we get,

$$\frac{\gamma}{2}\left\|\hat{\Delta}_{\boldsymbol{\mathcal{L}}} + \hat{\Delta}_{\boldsymbol{\mathcal{S}}}\right\|_F^2 \geq \frac{\gamma}{2}(\left\|\hat{\Delta}_{\boldsymbol{\mathcal{L}}}\right\|_F^2 + \left\|\hat{\Delta}_{\boldsymbol{\mathcal{S}}}\right\|_F^2) - \frac{\lambda_S}{2}\left\|\hat{\Delta}_{\boldsymbol{\mathcal{S}}Mat}\right\|_{2,1}$$

$$\geq \frac{\gamma}{2}(\left\|\hat{\Delta}_{\boldsymbol{\mathcal{L}}}\right\|_F^2 + \left\|\hat{\Delta}_{\boldsymbol{\mathfrak{S}}}\right\|_F^2) - \frac{\lambda_S}{2}\left\|\hat{\Delta}_{\boldsymbol{\mathfrak{S}}Mat}\right\|_{2,1} - \frac{\lambda_L}{2}\left\|Circ(\hat{\Delta}_{\boldsymbol{\mathcal{L}}})\right\|_*$$

So, we get the following inequality,

$$\frac{\gamma}{2}\left\|\hat{\Delta}_{\boldsymbol{\mathcal{L}}} + \hat{\Delta}_{\boldsymbol{\mathfrak{S}}}\right\|_F^2 \geq \frac{\gamma}{2}(\left\|\hat{\Delta}_{\boldsymbol{\mathcal{L}}}\right\|_F^2 + \left\|\hat{\Delta}_{\boldsymbol{\mathfrak{S}}}\right\|_F^2) - \frac{\lambda_L}{2}C(\hat{\Delta}_{\boldsymbol{\mathcal{L}}}, \hat{\Delta}_{\boldsymbol{\mathfrak{S}}}) \tag{24}$$

Next, we derive an upper bound of the left-hand side of inequality (22). To that end, using the inequality (6) and (19) we get,

$$\begin{aligned}
&\frac{1}{2n}\sum_{i=1}^{n}\{\langle\hat{\Delta}_{\boldsymbol{\mathcal{L}}} + \hat{\Delta}_{\boldsymbol{\mathfrak{S}}}, \boldsymbol{\mathfrak{X}}_i\rangle\}^2 \\
&\quad \leq \frac{1}{n}\sum_{i=1}^{n}\epsilon_i\langle\hat{\Delta}_{\boldsymbol{\mathcal{L}}} + \hat{\Delta}_{\boldsymbol{\mathfrak{S}}}, \boldsymbol{\mathfrak{X}}_i\rangle + \lambda_L\{C(\hat{\Delta}_{\boldsymbol{\mathcal{L}}}^A, \hat{\Delta}_{\boldsymbol{\mathfrak{S}}}^{\mathbb{M}}) - C(\hat{\Delta}_{\boldsymbol{\mathcal{L}}}^B, \hat{\Delta}_{\boldsymbol{\mathfrak{S}}}^{\mathbb{M}^\perp})\}
\end{aligned} \tag{25}$$

It can be seen that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\langle\hat{\Delta}_{\boldsymbol{\mathcal{L}}} + \hat{\Delta}_{\boldsymbol{\mathfrak{S}}}, \boldsymbol{\mathfrak{X}}_i\rangle \\
&=\langle\hat{\Delta}_{\boldsymbol{\mathcal{L}}} + \hat{\Delta}_{\boldsymbol{\mathfrak{S}}}, \boldsymbol{\mathfrak{D}}\rangle \\
&\leq\langle Circ(\hat{\Delta}_{\boldsymbol{\mathcal{L}}}), Circ(\boldsymbol{\mathfrak{D}})\rangle + \langle\hat{\Delta}_{\boldsymbol{\mathfrak{S}}Mat}, \boldsymbol{\mathfrak{D}}_{Mat}\rangle \\
&\leq\left\|Circ(\hat{\Delta}_{\boldsymbol{\mathcal{L}}})\right\|_*\left\|Circ(\boldsymbol{\mathfrak{D}})\right\|_{sp} + \left\|\hat{\Delta}_{\boldsymbol{\mathfrak{S}}Mat}\right\|_{2,1}\left\|\boldsymbol{\mathfrak{D}}_{Mat}\right\|_{2,\infty} \\
&\leq\left\|Circ(\boldsymbol{\mathfrak{D}})\right\|_{sp}\{\left\|Circ(\hat{\Delta}_{\boldsymbol{\mathcal{L}}}^A)\right\|_* + \left\|Circ(\hat{\Delta}_{\boldsymbol{\mathcal{L}}}^B)\right\|_*\} \\
&\quad + \left\|\boldsymbol{\mathfrak{D}}_{Mat}\right\|_{2,\infty}\{\left\|\hat{\Delta}_{\boldsymbol{\mathfrak{S}}Mat}^{\mathbb{M}}\right\|_{2,1} + \left\|\hat{\Delta}_{\boldsymbol{\mathfrak{S}}Mat}^{\mathbb{M}^\perp}\right\|_{2,1}\} \\
&\leq\frac{\lambda_L}{4}\{C(\hat{\Delta}_{\boldsymbol{\mathcal{L}}}^A, \hat{\Delta}_{\boldsymbol{\mathfrak{S}}}^{\mathbb{M}}) + C(\hat{\Delta}_{\boldsymbol{\mathcal{L}}}^B, \hat{\Delta}_{\boldsymbol{\mathfrak{S}}}^{\mathbb{M}^\perp})\} \text{ ,by definition of } C(\boldsymbol{\mathcal{L}}, \boldsymbol{\mathfrak{S}}) \\
&\quad \text{and Assumption 3, both ignoring } d_3
\end{aligned}$$

Putting the above inequality in inequality (25), we arrive at the following inequality

$$\frac{1}{2n}\sum_{i=1}^{n}\{\langle\hat{\Delta}_{\boldsymbol{\mathcal{L}}} + \hat{\Delta}_{\boldsymbol{\mathfrak{S}}}, \boldsymbol{\mathfrak{X}}_i\rangle\}^2 \leq \frac{3\lambda_L}{2}C(\hat{\Delta}_{\boldsymbol{\mathcal{L}}}^A, \hat{\Delta}_{\boldsymbol{\mathfrak{S}}}^{\mathbb{M}}) \tag{26}$$

Using the inequalities (22), (24) and (26), we get the following inequality,

$$\frac{\gamma}{2}(\left\|\hat{\Delta}_{\boldsymbol{\mathcal{L}}}\right\|_F^2 + \left\|\hat{\Delta}_{\boldsymbol{\mathfrak{S}}}\right\|_F^2) \leq \frac{3\lambda_L}{2}C(\hat{\Delta}_{\boldsymbol{\mathcal{L}}}^A, \hat{\Delta}_{\boldsymbol{\mathfrak{S}}}^{\mathbb{M}}) + \frac{\lambda_L}{2}C(\hat{\Delta}_{\boldsymbol{\mathcal{L}}}, \hat{\Delta}_{\boldsymbol{\mathfrak{S}}}) \tag{27}$$

Again, using Lemma 3.2 and the fact $C(\hat{\Delta}_{\boldsymbol{\mathcal{L}}}, \hat{\Delta}_{\boldsymbol{\mathcal{S}}}) \leq C(\hat{\Delta}_{\boldsymbol{\mathcal{L}}}^{A}, \hat{\Delta}_{\boldsymbol{\mathcal{S}}}^{\mathbb{M}}) + C(\hat{\Delta}_{\boldsymbol{\mathcal{L}}}^{B}, \hat{\Delta}_{\boldsymbol{\mathcal{S}}}^{\mathbb{M}^{\perp}})$, one can easily have,

$$C(\hat{\Delta}_{\boldsymbol{\mathcal{L}}}, \hat{\Delta}_{\boldsymbol{\mathcal{S}}}) \leq 4C(\hat{\Delta}_{\boldsymbol{\mathcal{L}}}^{A}, \hat{\Delta}_{\boldsymbol{\mathcal{S}}}^{\mathbb{M}}) \tag{28}$$

Replacing the above inequality in (27), we arrive at

$$\frac{\gamma}{2}(\left\|\hat{\Delta}_{\boldsymbol{\mathcal{L}}}\right\|_F^2 + \left\|\hat{\Delta}_{\boldsymbol{\mathcal{S}}}\right\|_F^2) \leq 4\lambda_L C(\hat{\Delta}_{\boldsymbol{\mathcal{L}}}^{A}, \hat{\Delta}_{\boldsymbol{\mathcal{S}}}^{\mathbb{M}}) \tag{29}$$

Recall from Lemma 3.1 that rank of $Circ(\hat{\Delta}_{\boldsymbol{\mathcal{L}}}^{A})$ is at most $2R$. This fact, along with the concept of *Compatibility Constant* defined in [1], reveals that

$$\begin{aligned} &\lambda_L C(\hat{\Delta}_{\boldsymbol{\mathcal{L}}}^{A}, \hat{\Delta}_{\boldsymbol{\mathcal{S}}}^{\mathbb{M}}) \\ \leq &\sqrt{2R}\lambda_L \left\|Circ(\hat{\Delta}_{\boldsymbol{\mathcal{L}}})\right\|_F + \lambda_S \sqrt{s} \left\|\hat{\Delta}_{\boldsymbol{\mathcal{S}}Mat}\right\|_F \end{aligned}$$

Next, we adjust the term d_3, that we ignored in the beginning of the proof, by adding the factor $\frac{1}{d_3}$ prior to $\left\|Circ(\hat{\Delta}_{\boldsymbol{\mathcal{L}}})\right\|_F$ in the above expression. Then, using the facts that $\left\|Circ(\hat{\Delta}_{\boldsymbol{\mathcal{L}}})\right\|_F = \sqrt{d_3}\left\|\hat{\Delta}_{\boldsymbol{\mathcal{L}}}\right\|_F$, $\left\|\hat{\Delta}_{\boldsymbol{\mathcal{S}}Mat}\right\|_F = \left\|\hat{\Delta}_{\boldsymbol{\mathcal{S}}}\right\|_F$ and $R \leq rd_3$, the above inequality reduces to

$$\lambda_L C(\hat{\Delta}_{\boldsymbol{\mathcal{L}}}^{A}, \hat{\Delta}_{\boldsymbol{\mathcal{S}}}^{\mathbb{M}}) \leq \sqrt{2r}\lambda_L \left\|\hat{\Delta}_{\boldsymbol{\mathcal{L}}}\right\|_F + \sqrt{s}\ \lambda_S \left\|\hat{\Delta}_{\boldsymbol{\mathcal{S}}}\right\|_F \tag{30}$$

The reader may note that the above inequality is exactly the same as the one obtained in [1], towards the very end of the proof of their Theorem 1. Hence, as done in [1], we substitute the above inequality into inequality (29) and then following the exact same steps as in [1], we complete the proof. □

Proof of Theorem 3.4

Proof. To prove this result, we follow the same technique used by [32] while proving their Lemma 11.

First, we prove that the condition $\lambda_L \geq 4\frac{1}{d_3}\left\|Circ(\boldsymbol{\mathfrak{D}})\right\|_{sp}$ is satisfied with high probability. It is easy to note that,

$$\begin{aligned} &\left\|Circ(\boldsymbol{\mathfrak{D}})\right\|_{sp} \\ =&\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i Circ(\boldsymbol{\mathfrak{X}}_i)\right\|_{sp} \end{aligned}$$

Comparing the above expression with the one in the statement of Lemma 11 of [32], one can claim that by choosing λ_L greater than $\frac{2\sigma c_{\max}}{\sqrt{n}}\mathbb{E}[\|G\|_{sp}]$, the condition $\lambda_L \geq \left\|Circ(\boldsymbol{\mathfrak{D}})\right\|_{sp}$ holds with probability at least $1 - \exp\{-(\mathbb{E}[\|G\|_{sp}])^2\}$, where G is a matrix of order $d_1 d_3 \times d_2 d_3$ with i.i.d. $N(0,1)$ entries and $c_{\max}$ is such that $\lambda_{\max}(\Sigma_{Circ}) \leq c_{\max}^2$, with $\Sigma_{Circ} = Cov((Vec(Circ(\boldsymbol{\mathfrak{X}}_1)))^T, \cdots, (Vec(Circ(\boldsymbol{\mathfrak{X}}_n)))^T)^T$.

It is easy to check that with proper permutations of rows and columns of Σ_{Circ}, one can arrive at the following block matrix,

$$C = \begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1d_3} \\ C_{21} & C_{22} & \cdots & C_{2d_3} \\ \vdots & \vdots & \ddots & \vdots \\ C_{d_3 1} & C_{d_3 2} & \cdots & C_{d_3^2} \end{bmatrix}$$

where, $C_{ij} = \Sigma$ (see (11) and (12)) for all $i = 1, 2, \cdots d_3$ and for all $j = 1, 2, \cdots, d_3$. Hence C is a block matrix, whose each of the d_3^2 blocks are Σ. Thus $\lambda_{\max}(\Sigma_{Circ}) = \lambda_{\max}(\Sigma) \times d_3 \leq c_u^2 d_3$, using assumption 12. Also, using Lemma H.1 of [27] we get $\mathbb{E}[\|G\|_{sp}] \leq 12(\sqrt{d_1 d_3} + \sqrt{d_2 d_3})$. Hence, for suitably chosen constant c_1^*, λ_L should be chosen as follows:

$$\begin{aligned} \lambda_L &\geq \frac{c_1^*}{d_3} \frac{\sigma}{\sqrt{n}} c_u \sqrt{d_3} (\sqrt{d_1 d_3} + \sqrt{d_2 d_3}) \\ &= c_1^* \frac{\sigma}{\sqrt{n}} c_u (\sqrt{d_1} + \sqrt{d_2}) \end{aligned}$$

and with such a choice of λ_L, the condition $\lambda_L \geq 4 \frac{1}{d_3} \|Circ(\boldsymbol{\mathfrak{D}})\|_{sp}$ is satisfied with probability greater than $1 - \exp(-d_3(d_1 + d_2))$.

Next, we prove that the condition $\lambda_S \geq 4 \|\boldsymbol{\mathfrak{D}}_{Mat}\|_{2,\infty}$ is satisfied with high probability. We note that,

$$\begin{aligned} &\|\boldsymbol{\mathfrak{D}}_{Mat}\|_{2,\infty} \\ &= \left\| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \boldsymbol{\mathfrak{X}}_{iMat} \right\|_{2,\infty} \end{aligned}$$

where $\boldsymbol{\mathfrak{X}}_{iMat}$ is the matrix of order $d_1 \times d_2 d_3$ that is constructed by placing the frontal slices $\boldsymbol{\mathfrak{X}}_i$ side by side. As before, using Lemma 11 of [32] one can claim that, by choosing λ_S greater than $\frac{2\sigma c^*}{\sqrt{n}} \mathbb{E}[\|G^*\|_{2,\infty}]$, the condition $\lambda_S \geq \|\boldsymbol{\mathfrak{D}}_{Mat}\|_{2,\infty}$ holds with probability at least $1 - \exp\{-(\mathbb{E}[\|G^*\|_{2,\infty}])^2\}$, where G^* is a matrix of order $d_1 \times d_2 d_3$ with i.i.d. $N(0,1)$ entries and c^* is such that $\lambda_{\max}(\Sigma_{Mat}) \leq {c^*}^2$, with $\Sigma_{Mat} = Cov((Vec(\boldsymbol{\mathfrak{X}}_{1Mat}))^T, \cdots, (Vec(\boldsymbol{\mathfrak{X}}_{nMat}))^T)^T$. However, it is easy to see that Σ_{Mat} is same as Σ and thus $\lambda_{\max}(\Sigma_{Mat}) \leq c_u^2$, using assumption 12.

Next, using Lemma 16 of [32], we have $\mathbb{E}[\|G^*\|_{2,\infty}] \leq 3(\sqrt{d_1} + \sqrt{\log(d_2 d_3)})$. Hence for a suitably chosen constant c_2^*, λ_S should be chosen as $\lambda_S \geq c_2^* \frac{\sigma}{\sqrt{n}} c_u (\sqrt{d_1} + \sqrt{\log(d_2 d_3)})$ and for such a choice of λ_S, the condition $\lambda_S \geq 4 \|\boldsymbol{\mathfrak{D}}_{Mat}\|_{2,\infty}$ is satisfied with probability greater than $1 - \exp(-9 \log(d_2 d_3))$. Hence, we have shown that the regularizer parameters satisfy the conditions in Assumption 3 with high probability.

Finally, we complete the proof by using Lemma 12 of [32] to show that assumption 1 holds with high probability. That lemma is based on the assumption that for any $c > 0$, there exist an n such that $\sqrt{S}\lambda \leq c$, where S is the *compatibility constant* (see Section 3 of [32]) of the regularizer in the cone set and λ is the associated parameter. Note that in our case, Lemma 3.2 characterizes the cone set and $C(\boldsymbol{\mathcal{L}}, \boldsymbol{\mathcal{S}})$ and λ_L are the regularizer and the parameter respectively. Hence, keeping in mind the choice of λ_L that we made at the first part of the proof and following some simple algebra, it can be shown that $\sqrt{S}\lambda$ has the form *constant* $\times$ $(\frac{\sqrt{d_1}}{\sqrt{n}} + \frac{\sqrt{d_2}}{\sqrt{n}})$. Hence, we need to assume that for any $c > 0$, there exists an n, such that $(\frac{\sqrt{d_1}}{\sqrt{n}} + \frac{\sqrt{d_2}}{\sqrt{n}}) \leq c$. However, note that this requirement is in line with the choices of λ_L and λ_S we make. Thus, using Lemma 12 of [32], we prove that Assumption 1 is satisfied with high probability. Now the proof follows using Lemma 3.3 and employing similar steps as in the proof of Corollary 4 of [1]. □

Proof of Corollary 3.4.1

Proof. Since the condition on λ_L does not change, we arrive at the same choice of λ_L as we did in the proof of Theorem 3.4. Now using the result on Gaussian maxima([18]), for the matrix G^* defined in the proof of Theorem 3.4, we get $\mathbb{E}[\|G^*\|_\infty] \leq 3\sqrt{\log(d_1 d_2 d_3)}$ and thus we choose $\lambda_S \geq c_3^* \frac{\sigma}{\sqrt{n}} c_u \sqrt{\log(d_1 d_2 d_3)}$. The rest of the proof follows along the same line of the proof of Theorem 3.4. □

Appendix B: Matrix-type view of a third-order tensor

We first present 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 [13], 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. Finally, by viewing the tube fibers of the tensor from the front, one would visualize them as the elements of a matrix. Figure 2 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 $\mathbb{R}^{d_1 \times 1 \times d_3}$ and $\mathbb{R}^{1 \times d_2 \times d_3}$ respectively. Similarly, a tube fiber, although a vector, can be considered as a third-order tensor in $\mathbb{R}^{1 \times 1 \times d_3}$. [13] refer such elements in $\mathbb{R}^{1 \times 1 \times d_3}$ as *Tubal Scalar*.

Given this matrix-type view of a third-order tensor, to proceed, we discuss next *Block Circulant Matrices*.

Notation B.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 B.1. *As discussed in [9], 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 B.2. *diag(F_n Circ($\boldsymbol{a}$) F_n^*) = fft($\boldsymbol{a}$), where 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 B.2. *For any third-order tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$, let $\boldsymbol{A_1}$, $\boldsymbol{A_2}$, ..., $\boldsymbol{A_{d_3}}$ 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 $d_1 d_3 \times d_2 d_3$*

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

Fact B.3. *Similar to Fact B.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:*

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

with $\otimes$ denoting the Kronecker product.

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

Notation B.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*

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

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

Appendix C: Background on t-product and t-SVD

Equipped with the aforementioned notations in Appendix B, we present next the *t-product* between two tensors and the corresponding t-SVD decomposition. The idea of the t-product was introduced in [15] and some of its important theoretical properties used in this work were developed in [14] and summarized next.

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

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

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

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

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

Fact C.1. *When a third-order tensor in $\mathbb{R}^{d_1 \times d_2 \times d_3}$ 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 C.2. *t-product can be computed efficiently in three steps. First, apply FFT on $\mathcal{A}$ and $\mathcal{B}$ along each tube and denote the resulting tensors as $\tilde{\mathcal{A}}$ and $\tilde{\mathcal{B}}$ respectively. Then multiply each frontal slice of $\tilde{\mathcal{A}}$ by the corresponding frontal slice of $\tilde{\mathcal{B}}$. 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 C.2. *The $n \times n \times l$ Identity Tensor, denoted by $\mathfrak{I}_{\boldsymbol{nnl}}$, 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 $\boldsymbol{\mathcal{A}} * \mathfrak{I} = \boldsymbol{\mathcal{A}} = \mathfrak{I} * \boldsymbol{\mathcal{A}}$

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

Definition C.4. *Transpose of $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{d_1\times d_2\times d_3}$, 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 C.2. *Suppose $d_3 = 4$. Then from the above definition,*

$$\boldsymbol{\mathcal{A}}^T = fold\left(\begin{bmatrix} \boldsymbol{A}_{\mathbf{1}}^T \\ \boldsymbol{A}_{\mathbf{4}}^T \\ \boldsymbol{A}_{\mathbf{3}}^T \\ \boldsymbol{A}_{\mathbf{2}}^T \end{bmatrix}\right)$$

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

Definition C.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*

$$\boldsymbol{Q}_{\boldsymbol{:i:}}^T * \boldsymbol{Q_{:j:}} = \begin{cases} \alpha_i \boldsymbol{e}\mathbf{1}, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}$$

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

Fact C.3. *$\mathfrak{Q} \in \mathbb{R}^{n\times n\times l}$ 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 $\mathbb{R}^{m\times 1\times l}$ contains m elements. Comparing this framework to usual matrix algebra, it would be of great use, if one could reconstruct any element in $\mathbb{R}^{m\times 1\times l}$ from those m elements. As discussed in [13], 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 C.7. *Given d_2 lateral slices, $\boldsymbol{X_{:1:}}$, $\boldsymbol{X_{:2:}}$,$\cdots$, $\boldsymbol{X_{:d_2:}}$ of $\mathfrak{X} \in \mathbb{R}^{d_1\times d_2\times d_3}$ 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 $\boldsymbol{X_{:1:}} * \boldsymbol{c_1} + \boldsymbol{X_{:2:}} * \boldsymbol{c_2}$ +...+ $\boldsymbol{X_{:d_2:}} * \boldsymbol{c_{d_2}}$, 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 $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{d_1\times d_2\times d_3}$, denoted by $\boldsymbol{R(\mathcal{A})}$, as the set of all possible t-linear combinations of its lateral slices. Similarly, extending the notion of 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 exist a tubal scalar $\boldsymbol{c}$, such that, $\boldsymbol{A_{:j_2:}} = \boldsymbol{A_{:j_1:}} * \boldsymbol{c}$. Figure 3 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 $\mathbb{R}^{d_1 \times 1 \times d_3}$, that is required to construct any arbitrary element in $\boldsymbol{R}(\boldsymbol{\mathcal{A}})$. As described in [13], 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 C.1. *An f-diagonal tensor, denoted by $\boldsymbol{\mathcal{F}}$, is a third-order tensor, whose each frontal slice is a diagonal matrix. In terms of notation, $\mathcal{F}_{ijk} = 0$, for $i \neq j, \forall k$.*

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

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

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

Definition C.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 $\boldsymbol{\mathcal{S}}$ 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 $\boldsymbol{\mathcal{S}}$.*

Like matrix singular value decomposition, in this case too, $\boldsymbol{R}(\boldsymbol{\mathcal{A}})$ can be written unambiguously by the lateral slices of $\boldsymbol{\mathcal{U}}$. Also, the number of elements in $R^{d_1 \times 1 \times d_3}$, required to construct any element in $\boldsymbol{R}(\boldsymbol{\mathcal{A}})$, is same as the tubal rank of $\boldsymbol{\mathcal{A}}$. The reader may visit [13] 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 4 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 [14] for more details. However, [26] 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. [26] 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 C.9. In terms of the notations used here, r $= \# \{i\text{: } \boldsymbol{s_{ii:}} \neq 0\} = \# \{i\text{: } \mathcal{S}_{ii1} \neq 0\}$. 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 [26], 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 [41]. However using Definition 7 of [26] and equation 12 of [25], one can derive the following relationship between

Tensor Nuclear Norm and Block Circulant matrix.

$$\sum_{i=1}^{r} \mathcal{S}_{ii1} = \frac{1}{d_3} \left\| Circ(\boldsymbol{\mathcal{A}}) \right\|_* \tag{31}$$

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.

Appendix D: Additional tables

TABLE 8
AIC Values for different choices of λ_L ans λ_S

λ_S λ_L	0.14	0.16	0.18	0.20	0.22	0.25	0.30	0.35
0.15	2053.95	2023.95	1923.95	1873.95	1800.36	1845.90	2406.37	2843.36
0.20	1404.88	1374.88	1274.88	1224.88	1141.90	1704.99	2376.26	2633.37
0.25	1292.50	1262.50	1162.50	1112.50	1034.37	1188.46	2285.26	2501.69
0.30	1314.68	1284.68	1184.68	1134.68	1064.73	1244.81	1812.72	2024.89
0.35	1404.16	1374.16	1274.16	1224.16	1134.12	1228.47	1714.84	1927.86
0.40	1512.23	1482.23	1382.23	1332.23	1250.18	1303.01	1716.18	1944.16
0.45	1621.33	1591.33	1491.33	1441.33	1357.01	1407.52	1797.11	2032.39
0.50	1733.60	1703.60	1603.60	1553.60	1466.13	1482.57	1858.72	2098.66
0.55	1845.40	1815.40	1715.40	1665.40	1577.37	1589.02	1933.68	2157.46
0.60	1914.18	1884.18	1784.18	1734.18	1649.76	1652.96	1930.98	2137.56

TABLE 9
Values of the estimated coefficients of the sparse component

Variable	Teaching Experience Level	Academic Year	Coefficient
tys	EXP2	Year1	-1.05
logins	EXP2	Year1	-0.17
Video	EXP3	Year1	-0.48
tys	EXP3	Year1	-1.06
Pre_Score	EXP3	Year1	-0.04
logins	EXP1	Year2	-0.01
logins	EXP2	Year2	-1.15
Video	EXP3	Year2	-0.39
Video	EXP1	Year3	-0.89
logins	EXP1	Year3	-0.16
Video	EXP2	Year3	-0.15
logins	EXP2	Year3	-0.16
logins	EXP3	Year3	-1.99
Video	EXP1	Year4	-0.04
logins	EXP2	Year4	-2.13
Video	EXP3	Year4	-0.51

References

[1] Agarwal, A., Negahban, S., Wainwright, M. J. et al. (2012). Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics* **40** 1171–1197. MR2985947

[2] Ahmed, T., Raja, H. and Bajwa, W. U. (2020). Tensor regression using low-rank and sparse Tucker decompositions. *SIAM Journal on Mathematics of Data Science* **2** 944–966. MR4161310

[3] Argyriou, A., Evgeniou, T. and Pontil, M. (2008). Convex multi-task feature learning. *Machine learning* **73** 243–272.

[4] Candès, E. J., Li, X., Ma, Y. and Wright, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM)* **58** 1–37. MR2811000

[5] Carroll, J. D. and Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika* **35** 283–319.

[6] Chandrasekaran, V., Sanghavi, S., Parrilo, P. A. and Willsky, A. S. (2011). Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization* **21** 572–596. MR2817479

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

[8] De Lathauwer, L., De Moor, B. and Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications* **21** 1253–1278. MR1780272

[9] Golub, G. H. and Loan, C. F. V. (1996). *Matrix Computations (Johns Hopkins Studies in Mathematical Sciences)(3rd Edition)*. Johns Hopkins University Press. MR1417720

[10] He, L., Chen, K., Xu, W., Zhou, J. and Wang, F. (2018). Boosted sparse and low-rank tensor regression. *arXiv preprint arXiv:1811.01158*.

[11] Hoff, P. D. (2015). Multilinear tensor regression for longitudinal relational data. *The annals of applied statistics* **9** 1169. MR3418719

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

[13] Kilmer, M. E., Braman, K., Hao, N. and Hoover, R. C. (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** 148–172. MR3032996

[14] Kilmer, M. E. and Martin, C. D. (2011). Factorization strategies for third-order tensors. *Linear Algebra and its Applications* **435** 641–658. MR2794595

[15] Kilmer, M. E., Martin, C. D. and Perrone, L. (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*.

[16] Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review* **51** 455–500. MR2535056

[17] Kossaifi, J., Lipton, Z. C., Khanna, A., Furlanello, T. and Anandkumar, A. (2017). Tensor regression networks. *arXiv preprint arXiv:1707.08308*.

[18] Ledoux, M. (1991). M. Talagrand Probability in Banach spaces. *Springer-Verlag* **62** 67–69. MR1102015

[19] Li, J., Bien, J., Wells, M. and Li, M. J. (2018). Package 'rTensor'.

[20] Li, W., Lou, J., Zhou, S. and Lu, H. (2019). Sturm: Sparse tubal-regularized multilinear regression for fmri. In *International Workshop on Machine Learning in Medical Imaging* 256–264. Springer.

[21] Li, X., Xu, D., Zhou, H. and Li, L. (2018). Tucker tensor regression and neuroimaging analysis. *Statistics in Biosciences* **10** 520–545.

[22] Li, Z., Suk, H.-I., Shen, D. and Li, L. (2016). Sparse multi-response tensor regression for Alzheimer's disease study with multivariate clinical assessments. *IEEE transactions on medical imaging* **35** 1927–1936.

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

[24] Liu, X.-Y., Aeron, S., Aggarwal, V. and Wang, X. (2019). Low-tubal-rank tensor completion using alternating minimization. *IEEE Transactions on Information Theory*. MR4077514

[25] Lu, C., Feng, J., Chen, Y., Liu, W., Lin, Z. and Yan, S. (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* 5249–5257.

[26] Lu, C., Feng, J., Lin, Z. and Yan, S. (2018). Exact low tubal rank tensor recovery from gaussian measurements. *arXiv preprint arXiv:1806.02511*.

[27] Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics* 1069–1097. MR2816348

[28] Negahban, S. N., Ravikumar, P., Wainwright, M. J., Yu, B. et al. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science* **27** 538–557. MR3025133

[29] Oseledets, I. V. (2011). Tensor-train decomposition. *SIAM Journal on Scientific Computing* **33** 2295–2317. MR2837533

[30] Peters, J., Bühlmann, P. and Meinshausen, N. (2015). Causal inference using invariant prediction: identification and confidence intervals. *arXiv preprint arXiv:1501.01332*. MR3557186

[31] Rabusseau, G. and Kadri, H. (2016). Low-rank regression with tensor responses. In *Advances in Neural Information Processing Systems* 1867–1875.

[32] Raskutti, G., Yuan, M., Chen, H. et al. (2019). Convex regularization for high-dimensional multiresponse tensor regression. *The Annals of*

Statistics **47** 1554–1584. MR3911122

[33] Recht, B., Fazel, M. and Parrilo, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review* **52** 471–501. MR2680543

[34] Sun, W. W. and Li, L. (2017). STORE: sparse tensor response regression and neuroimaging analysis. *The Journal of Machine Learning Research* **18** 4908–4944. MR3763769

[35] Tomioka, R. and Aihara, K. (2007). Classifying matrices with a spectral regularization. In *Proceedings of the 24th international conference on Machine learning* 895–902.

[36] Tucker, L. R. and Tucker, L. (1964). The extension of factor analysis to three-dimensional matrices.

[37] Xu, H., Caramanis, C. and Sanghavi, S. (2010). Robust PCA via outlier pursuit. In *Advances in Neural Information Processing Systems* 2496–2504. MR2952532

[38] Yang, Y. and Zou, H. (2015). A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing* **25** 1129–1141. MR3401877

[39] Yu, R. and Liu, Y. (2016). Learning from multiway data: Simple and efficient tensor regression. In *International Conference on Machine Learning* 373–381.

[40] Zhang, Z. and Aeron, S. (2016). Exact tensor completion using t-SVD. *IEEE Transactions on Signal Processing* **65** 1511–1526. MR3604692

[41] Zhang, Z., Ely, G., Aeron, S., Hao, N. and Kilmer, M. (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* 3842–3849.

[42] Zhou, H., Li, L. and Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association* **108** 540–552. MR3174640