Trainable Subspaces for Low Rank Tensor Completion: Model and Analysis

Zhen Long, Ce Zhu, Jiani Liu, Pierre Comon, Yipeng Liu

Trainable Subspaces for Low Rank Tensor Completion: Model and Analysis

Zhen Long, Ce Zhu, *Fellow, IEEE,* Jiani Liu, Pierre Comon, *Fellow, IEEE,* and Yipeng Liu, *Senior Member, IEEE*

***Abstract*—With the help of auxiliary data, tensor completion may better recover a low rank multidimensional array from limited observation entries. Most existing methods, including coupled matrix-tensor factorization and coupled tensor rank minimization, mainly focus on how to extract and incorporate subspace or directly use auxiliary data for tensor completion. They are either sensitive to a given rank or lack of physical interpretations of subspace information. In addition, the shared subspace information receives little attention in current tensor completion methods, especially there is no analysis of its impact on sample complexity. In this paper, we propose to separately explore and exploit shared subspaces for tensor completion. Specifically, dictionary learning takes the subspace from auxiliary data in the first step. Then a low rank optimization model for tensor completion is provided to incorporate the trained subspace by assuming that the recovered tensor is composed of two low rank components where one shares the subspace information with auxiliary data and the other is outside the shared space. Based on this optimization model, we make a quantitative analysis to illustrate the effect of subspace information on sample complexity, and provide theoretical insights into the usefulness of subspace information. Finally, experiments on simulated data are conducted to validate the theoretical analysis on the impact of subspace information. Experiments in two real-world applications including color image and multispectral image recovery show that the proposed method outperforms state-of-the-art ones in terms of prediction accuracy and CPU time.**

***Index Terms*—tensor completion, subspace information, sample complexity, low rank optimization, coupled tensor decomposition, dictionary learning**

I. Introduction

TENSOR completion concerns a problem that recovers a low-rank multidimensional array from a limited number of observations, which has a great number of applications including collaborative filtering [1], [2], [3], image recovery [4], [5], multi-task learning [6], [7], subspace clustering [8] and computer network traffic analysis [9].

The recovery performance bound largely depends on the sample complexity which is the least observed entries required to successfully recover the missing elements. The sample complexity is related to the low rank structure of data. Specifically, a matrix $\mathbf{X} \in \mathbb{R}^{I\times I}$ of rank R can be recovered from $O(RI\ln^2 I)$ observed entries with a high probability when the observed entries are uniformly sampled [10]. Based on it, for a Dth-order tensor $\mathcal{X} \in \mathbb{R}^{I_1\times I_2\times\cdots\times I_D}$, Huang et al. [11] proposed a multilinear rank minimization model to exactly recover missing entries. In this case, tensor completion problem is first transformed to a linear combination of several matrix completion problems. Considering matrix incoherence conditions over each unfolding matrix, at least $O(RI^{D-1}\ln^2 I^{D-1})$ observed entries are needed to recover a tensor with multilinear rank $(R_1,\cdots,R_D)$ where $R_1=\cdots=R_D=R$ and $I_1=\cdots=I_D=I$. Following this way, Huang et al. [12] showed exact tensor completion based on tensor ring decomposition requires no less than $O(I^{\lceil\frac{D}{2}\rceil}R^2\ln^7(I^{\lceil\frac{D}{2}\rceil}))$ known entries where R is the tensor ring rank. Jain et al. [13] proved that exact recovery of a 3rd-order symmetric tensor $\mathcal{X}\in\mathbb{R}^{I\times I\times I}$ with canonical polyadic (CP) rank R at least needs $O(I^{\frac{3}{2}}R^5\ln^4(I))$ samples using symmetric tensor incoherence condition with orthogonal decomposition. Zhang et al. [14] proposed a tubal rank minimization model which can guarantee exact recovery with high probability when the number of samples is $O(RI^2\ln I^2)$, given a tensor $\mathcal{X}\in\mathbb{R}^{I\times I\times I}$ with tubal rank R.

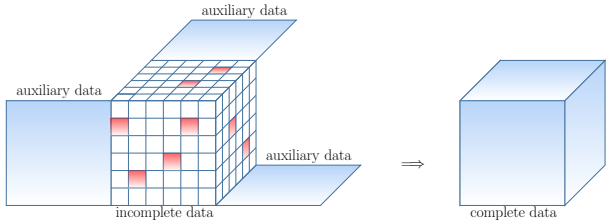


Fig. 1. An illustration of tensor completion with subspace information for a 3rd-order tensor.

As mentioned above, the number of known samples required for recovery is related to the data size I^D, no matter which tensor decomposition is used. A large number of observations will be in need for recovery when I increases, significantly limiting its applications on real world data. In particular, if the number of known entries is smaller than that required to successfully recover the data, tensor completion would fail. To address this issue, considering the usefulness of subspace information in hyperspectral image super-resolution [15], [16], some works [17], [18], [19] consider employing it into tensor completion to improve recovery performance when the observation entries are of inadequate, especially under the cold start setup [20]. For example, in a recommendation system, auxiliary data such as film types or user similarity based on social network data are also valuable in recommending movies

This research is supported by National Natural Science Foundation of China (NSFC, No. 62020106011, No. U19A2052, No. 62171088), and the China Scholarship Council (No.202006070056)

Z. Long, C. Zhu, J. Liu and Y. Liu are with School of Communication and Information Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China. (email: eczhu@uestc.edu.cn; yipengliu@uestc.edu.cn).

P. Comon is with Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-Lab, 38000 Grenoble, France.

to some users, in addition to the user's previous ratings for films [21]. In this case, the latent shared parts of auxiliary data and users ratings are known as subspace information. Fig. 1 shows an example of 3rd-order tensor completion with subspace information where the red cubes are missing entries and the three matrices contain auxiliary data.

On the use of subspace information for low rank tensor completion, three questions naturally arise, namely,

- How to efficiently explore the subspace from the auxiliary data and the incomplete data?
- What optimization model should we use to better incorporate the subspace information for tensor completion?
- How to measure the effect of subspace information on tensor completion?

Considering the fact that the subspace information can be fully or partially shared, which means the latent factors of the auxiliary data and the recovered data are the same or similar, a series of coupled matrix-tensor factorization methods [22], [23], [24], [25], [26] are proposed for tensor completion. They assume the auxiliary data and the incomplete data fully or partially share the same latent factors and simultaneously optimize the factors of the auxiliary data and the incomplete data with predefined ranks. However, this category needs a predefined rank, which may not be available for real-world data. When only a few observations are available, a choice of an inappropriate high rank may lead to overfitting. The other category algorithm, known as coupled matrix-tensor rank minimization models [27], [28], [29], attempts to minimize the ranks of the auxiliary data and the incomplete data. In this case, tensor ranks can be relaxed into its corresponding nuclear norms, and the desired data will be updated at each iteration. However, such attempts fail to give physical interpretation of the subspace or analysis of how the subspace information helps to enhance recovery performance.

As shown above, the model-based methods, like advanced coupled matrix-tensor factorization [25], can explore and exploit the fully or partially shared subspace information. However it is easily biased by the prior information, e.g. the setting of rank bound. To alleviate the problem in the model-based methods, we propose a trainable subspaces for low rank tensor completion (TS4LRTC) model by exploring and incorporating subspace separately, as shown the architecture in Fig. 2. Specifically, auxiliary and incomplete data are assumed to share some latent subspaces. In the first step, a data-driven method is considered on the auxiliary data to explore its subspace $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{A}_3$. To incorporate trained subspace, we design a low rank optimization model where the recovered tensor $\mathcal{X}$ is composed of two low rank components, one sharing the same subspace information with the auxiliary data while the other outside from the known subspaces. It is worth noting that we can choose any data-driven methods to learn subspace information, such as dictionary learning, deep learning, etc. In this paper, dictionary learning is considered to learn the subspace information.

In addition, we study two ways to measure the effect of subspace information on tensor completion. One is on sample complexity, and the other is the recovery performance. When the auxiliary data and the unknown data share the partial subspace information, we can obtain the sample complexity with $O(\max(R_{\mathcal{G}}\log K, R_{\mathcal{Y}}\log I)$ for a Dth-order tensor with size $I_1 \times I_2 \times \cdots \times I_D$, where $R_{\mathcal{G}}, R_{\mathcal{Y}}$ represents Tucker rank[1] of core tensors $\mathcal{G}$ and $\mathcal{Y}$, respectively. $I_1 = I_2 = \cdots = I_D = I$ and $K_1 = K_2 = \cdots = K_D = K$ denote the size of feature spaces. When the subspace information is fully shared and $K \ll I$, the sample complexity is $O(R_{\mathcal{G}}\log K)$, showing great advantage on tensor completion.

Finally, numerical experiments on simulated data are conducted to show that the subspace information enhance the recovery performance via reducing sample complexity. Besides, in real-world applications including color image and multispectral image recovery, our proposed method is superior to state-of-the-art ones in terms of prediction accuracy and CPU execution time.

A. Contributions

Compared with existing methods, our work mainly makes contributions in two aspects:

1) We propose a trainable subspaces for low rank tensor completion model, where the subspace information can be partially or fully shared for recovered data. Our model provides a new insight to explore and incorporate subspace for tensor completion.
2) We examine the relationship of sample complexity and subspace information, which quantitatively gives guidelines on how the subspace information helps to enhance the recovery performance.

B. Paper organization

The rest of this paper is organized as follows. In Section II, we give the notations and preliminaries on tensor decomposition. In Section III, we illustrate how to explore the subspace and what optimization model we choose to incorporate subspace for tensor completion. In addition, the solutions to optimization model and the effect of subspace information on sample complexity are also given in the section. In section IV, the sample complexity bound on synthetic data and numerical recovery results on real data are demonstrated. The conclusion is drawn in Section V.

II. Notations and preliminaries

A. Notations

For clarity, we list some frequently used notations in TABLE I.

B. Preliminaries on tensor operations and tensor decompositions

Definition *1:* (**Mode-d product**) The mode-d product of a given tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_d \times \cdots \times I_D}$ and a matrix $\mathbf{Y} \in \mathbb{R}^{I_d \times J}$ can be denoted as

$$\mathcal{Z} = \mathcal{X} \times_d \mathbf{Y} \in \mathbb{R}^{I_1 \times \cdots \times J \times \cdots \times I_D},$$

[1]Tucker rank is the Duple of ranks $(R_1, \ldots, R_D)$ of unfolding matrices. It is also referred to as *multilinear rank* in the literature.

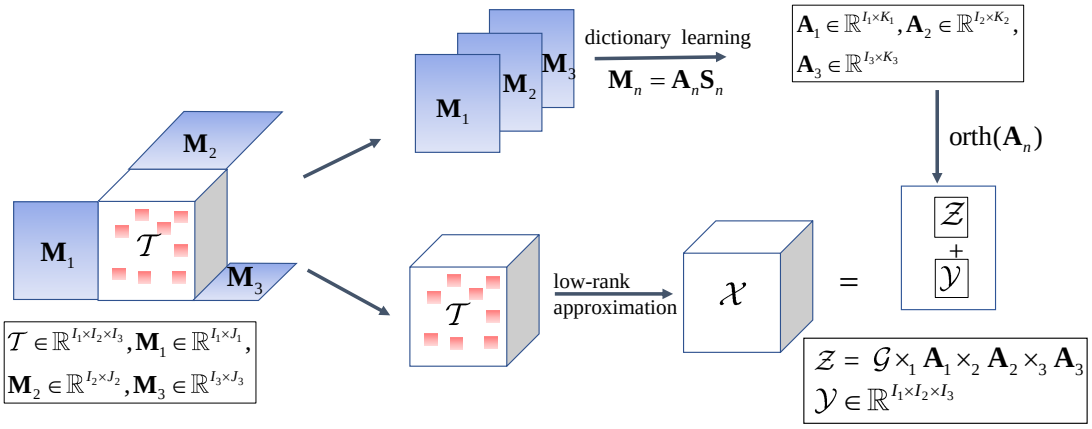


Fig. 2. An illustration of the trainable subspaces for low rank tensor completion model .

TABLE I
SUMMARY OF NOTATIONS IN THIS PAPER.

Symbol	Description
x, $\mathbf{x}$, $\mathbf{X}$, $\mathcal{X}$	scalar, vector, matrix, tensor
D	tensor order
I_d	tensor size along d-th mode
$\mathcal{X} \in \mathbb{R}^{I_1\times\cdots\times I_D}$	a D-th tensor with size $I_1\times\cdots\times I_D$
$\mathcal{X}(i_1,\cdots,i_D)$	$i_1,\cdots,i_D$-th entry of $\mathcal{X}$
$\mathbb{O}$	index set of observed entries
col($\mathbf{X}$)	column space of $\mathbf{X}$
$\|\mathbf{X}\|_*$	trace norm (nuclear norm)
$\otimes$	tensor product
$\boxtimes$	Kronecker product
$\langle\mathcal{X},\mathcal{Y}\rangle$	tensor inner product
$\|\mathcal{X}\|_{\mathrm{F}}$	tensor Frobenius norm

where $\times_d$ stands for the contraction between the dth tensor index and the second matrix index.

Definition *2:* (**Mode-d unfolding**). The mode-d unfolding matrix of $\mathcal{X}$ is defined by $\mathbf{X}_{(d)} \in \mathbb{R}^{I_d\times I_{d+1}\cdots I_D I_1\cdots I_{d-1}}$, and its opposite operation 'fold$_d$' is defined as $\mathrm{fold}_d(\mathbf{X}_{(d)}) = \mathcal{X}$.

Definition *3:* (**Tucker decomposition**) [30] For a Dth-order tensor $\mathcal{X} \in \mathbb{R}^{I_1\times\cdots\times I_D}$, the goal of Tucker decomposition is to decompose a tensor $\mathcal{X}$ into a core tensor multiplied by D factor matrices along corresponding modes:

$$\mathcal{X} = \mathcal{G}\times_1 \mathbf{A}_1\cdots\times_D \mathbf{A}_D,$$

where $\mathcal{G} \in \mathbb{R}^{R_1\times\cdots\times R_D}$ is the core tensor and $\mathbf{A}_d \in \mathbb{R}^{I_d\times R_d}, d = 1,\cdots,D$ are factor matrices. For simplicity, $\mathcal{X} = [\![\mathcal{G};\mathbf{A}_1,\cdots,\mathbf{A}_D]\!]$ is also subsequently used to denote this contraction. Next, $R_d = \mathrm{rank}(\mathbf{X}_{(d)}), d = 1,\cdots,D$, are Tucker ranks – also called *mode ranks* or *multilinear rank* – where $\mathbf{X}_{(d)}$ is the mode-d matrix of $\mathcal{X}$.

When the core tensor $\mathcal{G}$ is cubical diagonal and contains only ones, $\mathcal{G}$ can be omitted, which leads to the notation $\mathcal{X} = [\![\mathbf{A}_1,\cdots,\mathbf{A}_D]\!]$. In such a case, we face the so-called CP decomposition. CP stands either for "Canonical Polyadic" or for "Candecomp/Parafac" [31]. This decomposition is also sometimes called *Kruskal decomposition* or *rank-revealing* decomposition.

Definition *4:* (**tensor trace norm**) [32] The trace norm of tensor $\mathcal{X} \in \mathbb{R}^{I_1\times\cdots\times I_D}$ is defined as:

$$\|\mathcal{X}\|_* = \sum_{d=1}^{D} \alpha_d \|\mathbf{X}_{(d)}\|_*, \tag{1}$$

where $\alpha_d, d = 1,\cdots,D$ are constants, $\alpha_d \geq 0$ and $\sum_{d=1}^{D}\alpha_d = 1$, the trace norm (nuclear norm) of matrix $\mathbf{X}_{(d)} \in \mathbb{R}^{I_d\times I_{\neq d}}$ is denoted as $\|\mathbf{X}_{(d)}\|_* = \sum_{r_d=1}^{R_d}\sigma_r(\mathbf{X}_{(d)})$, where R_d is the rank of $\mathbf{X}_{(d)}$, $\sigma_r(\mathbf{X}_{(d)})$ is the r-th singular value of matrix $\mathbf{X}_{(d)}$, and $i_{\neq d} = i_{d+1}\cdots i_D i_1\cdots i_{d-1}$. In practice, we often set $\alpha_d = \frac{w_d}{\sum_{d=1}^{D} w_d}, w_d = \min(\mathrm{size}(\mathbf{X}_{(d)}))$ or $\alpha_d = \frac{1}{D}, d = 1,\cdots,D$..

According to the fact $\|\mathbf{X}_{(d)}\|_* \leq \sqrt{R_d}\|\mathbf{X}_{(d)}\|_{\mathrm{F}}$ and $\|\mathcal{X}\|_{\mathrm{F}} = \|\mathbf{X}_{(d)}\|_{\mathrm{F}}, d = 1,\cdots,D$, we have :

$$\|\mathcal{X}\|_* \leq \sum_{d=1}^{D}\alpha_d\sqrt{R_d}\|\mathbf{X}_{(d)}\|_{\mathrm{F}} = \sum_{d=1}^{D}\alpha_d\sqrt{R_d}\|\mathcal{X}\|_{\mathrm{F}}. \tag{2}$$

Letting $R = \max(R_d), d = 1,\cdots,D$ and if $\|\mathcal{X}\|_{\mathrm{F}}$ is bounded by X, then we obtain $\|\mathcal{X}\|_* \leq \sqrt{R}X$.

Theorem *1:* Let $\mathcal{X} = [\![\mathcal{G};\mathbf{A}_1,\cdots,\mathbf{A}_D]\!]$, where $\mathcal{X} \in \mathbb{R}^{I_1\times\cdots\times I_D}$, $\mathbf{A}_d, d = 1,\cdots,D$ are of size $I_d\times K_d$. If $\mathbf{A}_d$ are orthonormal matrices and $\mathrm{col}(\mathbf{X}_{(d)}) \subseteq \mathrm{col}(\mathbf{A}_d), d = 1,\cdots,D$, then $\|\mathcal{X}\|_* = \|\mathcal{G}\|_*$.

Proof *1:* See Appendix A.

C. Related works

Recently, incorporating subspace information into matrix completion has attracted much attention [33], [34], [35], [36]. As a generalization of matrix completion, tensor completion with subspace information has been a topic of interest in the fields of community detection [37], collaborative filtering [38] and linked prediction [39]. Current tensor completion with subspace information methods can be divided into two groups. The first completion model is based on matrix-tensor factorization, and the second is based on coupled tensor rank minimization.

1) Coupled matrix-tensor factorization completion model: Assuming a Dth-order tensor $\mathcal{T} \in \mathbb{R}^{I_1\times I_2\times\cdots\times I_D}$ and a matrix $\mathbf{M} \in \mathbb{R}^{I_1\times J}$ coupled in the first mode of each, Acar et al. [23]

first proposed a CP based coupled matrix-tensor factorization model, as follows

$$\min_{\mathbf{A}_1,\cdots,\mathbf{A}_D,\mathbf{W}} \frac{1}{2}\|\mathcal{P}_{\mathbb{O}}(\mathcal{T}-[\![\mathbf{A}_1,\cdots,\mathbf{A}_D]\!])\|_{\mathrm{F}}^2+\frac{1}{2}\|\mathbf{M}-\mathbf{A}_1\mathbf{W}\|_{\mathrm{F}}^2, \quad (3)$$

where $\mathbf{A}_1 \in \mathbb{R}^{I_1\times R}$ is the shared latent factor, R is the CP rank, and $\mathcal{P}_{\mathbb{O}}$ denotes the random sampling operator, which is defined by

$$\mathcal{P}_{\mathbb{O}}(\mathcal{T}) = \begin{cases} \mathcal{T}(i_1,\cdots,i_D) & , \ i_1,\cdots,i_D \in \mathbb{O} \\ 0 & , \ i_1,\cdots,i_D \notin \mathbb{O} \end{cases}. \quad (4)$$

All factors can be updated by a gradient descent optimization. Based on it, Chen [40] proposed a collective robust tensor completion model with multiple heterogeneous side information, which is formulated as:

$$\begin{aligned} &\min_{\mathcal{L},\{\mathbf{A}_d\}_{d=1}^D,\{\mathbf{U}\}_{d=1}^D,\mathcal{E}} \frac{1}{2}\|\mathcal{L}-[\![\mathbf{A}_1,\cdots,\mathbf{A}_D]\!])\|_{\mathrm{F}}^2+\beta\|\mathcal{E}\|_0 \\ &+\sum_{d=1}^{D}\lambda_d(\|\mathbf{M}_d-\mathbf{U}_d\mathbf{U}_d^{\mathrm{T}}\|_{\mathrm{F}}^2+\|\mathbf{A}_d-\mathbf{U}_d\mathbf{S}_d\|_{\mathrm{F}}^2), \\ &\text{s. t.} \quad \mathcal{X}=\mathcal{L}+\mathcal{S}, \mathcal{X}_{\mathbb{O}}=\mathcal{T}_{\mathbb{O}}, \end{aligned} \quad (5)$$

where β and $\lambda_d, d=1,\cdots,D$ are weights, which control the sparsity of $\mathcal{E}$ and the impact of side information, respectively. $\mathbf{S}_d$ is a scale matrix of $\mathbf{U}_d$, which is defined as $\mathbf{S}_d = \mathrm{diag}(\sum_i \mathbf{U}_d(i,1), \sum_i \mathbf{U}_d(i,2), \cdots, \sum_i \mathbf{U}_d(i,R))$. In parallel, Yilmaz et al. [41] proposed a Tucker based generalized coupled tensor factorization to tackle the missing data problem. The optimization model can be formulated as

$$\min_{\mathcal{G},\mathbf{A}_1,\cdots,\mathbf{A}_D,\mathbf{W}} \frac{1}{2}\|\mathcal{P}_{\mathbb{O}}(\mathcal{T}-[\![\mathcal{G};\mathbf{A}_1,\cdots,\mathbf{A}_D]\!])\|_{\mathrm{F}}^2+\frac{1}{2}\|\mathbf{M}-\mathbf{A}_1\mathbf{W}\|_{\mathrm{F}}^2, \quad (6)$$

where the shared latent factor $\mathbf{A}_1$ is also called *shared subspace information*. Moreover, some flexible frameworks are proposed to more accurately model the coupled data [42], [22], [25], [26]. For example, Acar et al. [25] proposed a constrained optimization problem which assumes the auxiliary and the incomplete data share partial subspace information as follows.

$$\begin{aligned} &\min_{\mathbf{A}_1,\cdots,\mathbf{A}_D,\boldsymbol{\lambda},\Sigma,\mathbf{V}} \frac{1}{2}\|\mathcal{P}_{\mathbb{O}}(\mathcal{T}-[\![\boldsymbol{\lambda};\mathbf{A}_1,\cdots,\mathbf{A}_D]\!])\|_{\mathrm{F}}^2 \\ &+\frac{1}{2}\|\mathbf{M}-\mathbf{A}_1\Sigma\mathbf{V}^{\mathrm{T}}\|_{\mathrm{F}}^2, \quad \text{s. t.} \quad \|\mathbf{A}_d(:,r)\|_2=1, \\ &\|\mathbf{V}(:,r)\|_2=1, \sum_{r=1}^{R}\lambda_r\le\beta, \sum_{r=1}^{R}\sigma_r\le\beta, \sigma_r\ge 0, \lambda_r\ge 0, \\ &d=1,\cdots,D, r=1,\cdots,R, \end{aligned} \quad (7)$$

where $\beta>0$ is a user-defined parameter, λ_r and σ_r are weights of rank-one components. In this case, shared/unshared factors can be revealed by constraining weight sparsity, e.g. $\sum_{r=1}^R \lambda_r \le \beta$ and $\sum_{r=1}^R \sigma_r \le \beta$ and $\Sigma = \mathrm{diag}([\sigma_1,\cdots,\sigma_R])$. In (7), columns of $\mathbf{V}$ are not orthogonal, but additional constraints could be appended to impose a minimal angular separation between columns, as introduced in [43], [44], and used in [25].

However, this group of methods is non-convex and can converge to a local optimum. Besides, the ranks of the matrix and the tensor are assumed known in advance, but may not be available in some applications. Yet, the choice of a too large rank may lead to overfitting when only a few observations are available.

2) Coupled tensor rank minimization model: To alleviate rank determination problem, some works directly minimize the rank. For example, in [27], the authors proposed a Tucker based coupled rank minimization model for tensor completion as follows

$$\min_{\mathcal{X},\hat{\mathbf{M}}} \frac{1}{2}\|\mathcal{P}_{\mathbb{O}}(\mathcal{X}-\mathcal{T})\|_{\mathrm{F}}^2+\frac{1}{2}\|\hat{\mathbf{M}}-\mathbf{M}\|_{\mathrm{F}}^2+\lambda\|\mathcal{X},\hat{\mathbf{M}}\|_{\mathrm{cn}}, \quad (8)$$

where $\lambda\ge 0$ is the regularization parameter, $\mathcal{T}\in\mathbb{R}^{I_1\times I_2\times\cdots\times I_D}$ is the observed tensor, and $\mathbf{M}\in\mathbb{R}^{I_1\times J_1}$ is the auxiliary matrix, which is coupled with the first mode of $\mathcal{T}$. $\|\mathcal{X},\hat{\mathbf{M}}\|_{\mathrm{cn}}$ is the coupled overlapped Schatten 1-norm, which can be defined as

$$\|\mathcal{X},\hat{\mathbf{M}}\|_{\mathrm{cn}} \overset{\mathrm{def}}{=} \|[\mathbf{X}_{(1)};\hat{\mathbf{M}}]\|_* + \sum_{d=2}^{D}\|\mathbf{X}_{(d)}\|_*, \quad (9)$$

where $\|[\mathbf{X}_{(1)};\hat{\mathbf{M}}]\|_*$ represents the trace norm of matrix $[\mathbf{X}_{(1)};\hat{\mathbf{M}}]$.

In addition, a CP based coupled rank minimization model was proposed in [29]. The model can be represented by

$$\min_{\mathcal{X}}\|\mathcal{X},\mathbf{M}\|_{\mathrm{ccp}} \quad \text{s. t.} \quad \mathcal{X}_{\mathbb{O}}=\mathcal{T}_{\mathbb{O}}, \quad (10)$$

where $\|\mathcal{X},\mathbf{M}\|_{\mathrm{ccp}}$ is the coupled nuclear norm using CP decomposition, which is defined as

$$\begin{aligned} &\|\mathcal{X},\mathbf{M}\|_{\mathrm{ccp}} \overset{\mathrm{def}}{=} \{\|\mathcal{X}\|_*\le C_1, \|\mathbf{M}\|_*\le C_2| \\ &\mathcal{X}=\sum_{r=1}^{R}\lambda_r\mathbf{a}_1^r\otimes\cdots\otimes\mathbf{a}_d^r, \mathbf{M}=\sum_{r=1}^{R}\sigma_r\mathbf{a}_1^r(\mathbf{v}^r)^{\mathrm{T}}\}, \end{aligned} \quad (11)$$

where C_1 and C_2 are constants, tensor $\mathcal{X}$ and matrix $\mathbf{M}$ are assumed to have the same rank R.

However, this group directly optimizes the recovered tensor which lacks the physical interpretation about the subspace information. In addition, for solving coupled nuclear norm in Tucker decomposition format, it needs to perform SVD at each iteration, which has high computational complexity for large size matrices.

In addition to incorporate subspace information from coupled data, exploring structural information from the dataset with the same properties is also useful to improve recovery performance in tensor completion model. For instance, Yang [45] proposed a low TT rank and fiber-wise sparsity based tensor completion model, where a low rank prior is used to exploit the inter correlations and a sparsity prior with pre-trained dictionaries is used to exploit intra correlations from other datasets, as follows

$$\begin{aligned} \min_{\mathcal{X},\alpha_d} \ &\sum_{d=1}^{D-1} w_d\|\mathbf{X}_{[d]}\|_*+\gamma\sum_{d=1}^{D}\|\alpha_d\|_1 \\ \text{s. t.} \ &\mathbf{X}_{(d)}=\Phi_d\alpha_d, d=1,\cdots,D, \quad \mathcal{X}_{\mathbb{O}}=\mathcal{T}_{\mathbb{O}}, \end{aligned} \quad (12)$$

where w_d and γ are weights, which control the low rankness of $\|\mathbf{X}_{[d]}\|_*$ and the impact of information from other datasets, respectively. Φ_d is a pre-trained dictionary and $\mathbf{X}_{[d]}\in\mathbb{R}^{\prod_{l=1}^{d} I_l\times\prod_{l=d+1}^{D} I_l}$ is a balanced unfolding matrix of $\mathcal{X}$.

III. Proposed method

As shown in Section II-C, most existing methods consider using the auxiliary data in a coupled way, which simultaneously extracts and utilizes the subspace information or directly uses the auxiliary data for tensor completion. In this section, we propose a trainable subspaces for low rank tensor completion model by separating extraction and utilization of the shared subspace information.

A. Model development

Let $\mathcal{T} \in \mathbb{R}^{I_1 \times \cdots \times I_D}$ be a Dth-order incomplete tensor with each mode coupled with an auxiliary matrix $\mathbf{M}_d \in \mathbb{R}^{I_d \times J_d}$. In the first step, dictionary learning [46] is considered on matrices $\mathbf{M}_d \in \mathbb{R}^{I_d \times J_d}, d = 1, \cdots, D$ to obtain well-trained subspace information $\mathbf{A}_d \in \mathbb{R}^{I_d \times K_d}, d = 1, \cdots, D$, where K_d is the size of feature space. To better recover missing data with the help of subspace information, the recovered low-rank tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_D}$ is assumed to be composed by two low rank components $\mathcal{Z}$ and $\mathcal{Y}$, where $\mathcal{Z}$ is linked to shared information and $\mathcal{Y}$ lies outside feature spaces. A 3rd-order model is shown in Fig. 2. Then the optimization model can be formulated as:

$$\begin{aligned} &\min_{\mathcal{Z},\mathcal{Y},\mathcal{G}} \ \lambda_{\mathcal{G}} \operatorname{rank}(\mathcal{Z}) + \lambda_{\mathcal{Y}} \operatorname{rank}(\mathcal{Y}) \\ &\text{s. t. } \mathcal{X} = \mathcal{Z} + \mathcal{Y}, \mathcal{X}_{\mathbb{O}} = \mathcal{T}_{\mathbb{O}}, \end{aligned} \tag{13}$$

where $\mathcal{Z} = \mathcal{G} \times_1 \mathbf{A}_1 \times \cdots \times_D \mathbf{A}_D$. $\lambda_{\mathcal{G}}$ and $\lambda_{\mathcal{Y}}$ are crucial, since their ratio controls the contributions from $\mathcal{Z}$ and $\mathcal{Y}$ for recovering low rank tensor $\mathcal{X}$. For example, when $\lambda_{\mathcal{G}} = \infty$, $\mathcal{Z}$ will have no influence, and model (13) will be degenerated as a standard tensor completion model [32]. Instead, when $\lambda_{\mathcal{Y}} = \infty$, $\mathcal{Y}$ will have no influence. In this case, the incomplete tensor and auxiliary matrix share full subspace information, the optimization model (13) will be degenerated as

$$\min_{\mathcal{G},\mathcal{Z}} \ \lambda_{\mathcal{G}} \operatorname{rank}(\mathcal{Z}) \quad \text{s. t. } \mathcal{X} = \mathcal{Z}, \mathcal{X}_{\mathbb{O}} = \mathcal{T}_{\mathbb{O}}. \tag{14}$$

with $\mathcal{Z} = \mathcal{G} \times_1 \mathbf{A}_1 \times \cdots \times_D \mathbf{A}_D$.

Note that the focus is on subspace information and its effect on tensor completion. Therefore, the tensor ranks of $\mathcal{Y}, \mathcal{Z}$ can be of any kinds, e.g. Tucker rank [30], tensor tree rank [47], tensor train rank [48], tensor ring rank [49] and so on. In our model, Tucker rank and its convex trace norm are considered. In this case, model (13) can be relaxed to its convex one, as follows.

$$\begin{aligned} &\min_{\mathcal{Z},\mathcal{Y},\mathcal{G}} \ \lambda_{\mathcal{G}} \|\mathcal{Z}\|_* + \lambda_{\mathcal{Y}} \|\mathcal{Y}\|_* \\ &\text{s. t. } \mathcal{X} = \mathcal{Z} + \mathcal{Y}, \mathcal{X}_{\mathbb{O}} = \mathcal{T}_{\mathbb{O}}, \end{aligned} \tag{15}$$

with $\mathcal{Z} = \mathcal{G} \times_1 \mathbf{A}_1 \times \cdots \times_D \mathbf{A}_D$. $\|\mathcal{Z}\|_*$ is the tensor trace norm in definition (4). According to Theorem 1, $\|\mathcal{Z}\|_* = \|\mathcal{G}\|_*$ with orthogonal $\mathbf{A}_d, d = 1, \cdots, D$.

Utilizing definition (4), an equivalent optimization model for model (15) can be formulated as

$$\begin{aligned} &\min_{\mathcal{G},\mathcal{Y}} \ \lambda_{\mathcal{G}} \sum_{d=1}^{D} \alpha_d \|\mathbf{G}_{(d)}\|_* + \lambda_{\mathcal{Y}} \sum_{d=1}^{D} \gamma_d \|\mathbf{Y}_{(d)}\|_* \\ &\text{s. t. } \mathcal{X} = \mathcal{G} \times_1 \mathbf{A}_1 \times \cdots \times_D \mathbf{A}_D + \mathcal{Y}, \mathcal{X}_{\mathbb{O}} = \mathcal{T}_{\mathbb{O}}, \end{aligned} \tag{16}$$

where $\alpha_d, \gamma_d, d = 1, \cdots, D$ are constants, $\alpha_d \geq 0, \gamma_d \geq 0$ and $\sum_{d=1}^{D} \alpha_d = 1, \sum_{d=1}^{D} \gamma_d = 1$.

In the following, we give detailed solutions and theoretical results on the sample complexity for problem (16).

B. Solutions

The objective function in problem (16) can be further converted into the following optimization model:

$$\begin{aligned} \min_{\mathcal{G},\mathcal{Y},\Lambda} \quad & \lambda_{\mathcal{G}} \sum_{d=1}^{D} \alpha_d \|\mathbf{G}_{(d)}\|_* + \lambda_{\mathcal{Y}} \sum_{d=1}^{D} \gamma_d \|\mathbf{Y}_{(d)}\|_* \\ & + \langle \Lambda, \mathcal{X} - \mathcal{G} \times_1 \mathbf{A}_1 \times \cdots \times_D \mathbf{A}_D - \mathcal{Y} \rangle \\ & + \frac{\beta}{2} \|\mathcal{X} - \mathcal{G} \times_1 \mathbf{A}_1 \times \cdots \times_D \mathbf{A}_D - \mathcal{Y}\|_{\mathrm{F}}^2, \end{aligned} \tag{17}$$

under the constraint $\mathcal{X}_{\mathbb{O}} = \mathcal{T}_{\mathbb{O}}$, where $\Lambda \in \mathbb{R}^{I_1 \times \cdots \times I_D}$ is the Lagrange multiplier and β is a positive penalty scalar. Problem (17) can be split into several simple subproblems within the alternating direction method of multipliers (ADMM) [50] framework.

1) Update $\mathcal{G}$: The subproblem with respect to $\mathcal{G}$ is

$$\min_{\mathcal{G}} \quad \frac{\lambda_{\mathcal{G}}}{\beta} \sum_{d=1}^{D} \alpha_d \|\mathbf{G}_{(d)}\|_* + \frac{1}{2} \|\mathcal{G} - \mathcal{S}\|_{\mathrm{F}}^2, \tag{18}$$

where $\mathcal{S} = (\mathcal{X} - \mathcal{Y} + \frac{\Lambda}{\beta}) \times_1 \mathbf{A}_1^T \times \cdots \times_D \mathbf{A}_D^T$. It can be further divided into D subproblems as follows

$$\min_{\mathbf{G}_{(d)}} \quad \tau \|\mathbf{G}_{(d)}\|_* + \frac{1}{2} \|\mathbf{G}_{(d)} - \mathbf{S}_{(d)}\|_{\mathrm{F}}^2, \tag{19}$$

where $\tau = \frac{\alpha_d \lambda_{\mathcal{G}}}{\beta}$. The solution of $\mathbf{G}_{(d)}$ can be updated by

$$\mathbf{G}_{(d)} = \mathrm{SVT}(\mathbf{S}_{(d)}, \tau), \tag{20}$$

where $\mathrm{SVT}(\mathbf{X}, \tau) = \mathbf{U} \operatorname{sth}_\tau(\Sigma) \mathbf{V}^{\mathrm{T}}$, $[\mathbf{U}, \Sigma, \mathbf{V}] = \mathrm{SVD}(\mathbf{X})$, sth_τ is the well-known soft thresholding operator as follows

$$\operatorname{sth}_\tau(x) = \operatorname{sgn}(x) \max(|x| - \tau, 0). \tag{21}$$

2) Update $\mathcal{Y}$: With the other variables fixed, the subproblem on $\mathcal{Y}$ is given by

$$\min_{\mathcal{Y}} \quad \frac{\lambda_{\mathcal{Y}}}{\beta} \sum_{d=1}^{D} \gamma_d \|\mathbf{Y}_{(d)}\|_* + \frac{1}{2} \|\mathcal{Y} - \mathcal{W}\|_{\mathrm{F}}^2, \tag{22}$$

where $\mathcal{W} = \mathcal{X} - \mathcal{G} \times_1 \mathbf{A}_1 \times \cdots \times_D \mathbf{A}_D + \frac{\Lambda}{\beta}$. Similarly, this problem can be divided into D subproblems and each subproblem can be solved by operator $\mathrm{SVT}(\mathbf{W}_{(d)}, \tau)$ with $\tau = \frac{\gamma_d \lambda_{\mathcal{Y}}}{\beta}$.

3) Update $\mathcal{X}$: The subproblem with respect to $\mathcal{X}$ is

$$\begin{aligned} \min_{\mathcal{X}} \quad & \langle \Lambda, \mathcal{X} - \mathcal{G} \times_1 \mathbf{A}_1 \cdots \times_D \mathbf{A}_D - \mathcal{Y} \rangle \\ & + \frac{\beta}{2} \|\mathcal{X} - \mathcal{G} \times_1 \mathbf{A}_1 \cdots \times_D \mathbf{A}_D - \mathcal{Y}\|_{\mathrm{F}}^2 \\ & \text{s. t. } \mathcal{X}_{\mathbb{O}} = \mathcal{T}_{\mathbb{O}}. \end{aligned} \tag{23}$$

The objective function is smooth and differentiable, so we can update $\mathcal{X}$ as follows

$$\mathcal{X}(i_1, \cdots, i_D) = \begin{cases} \mathcal{D}(i_1, \cdots, i_D), \ i_1, \cdots, i_D \notin \mathbb{O} \\ \mathcal{T}(i_1, \cdots, i_D), \ i_1, \cdots, i_D \in \mathbb{O}. \end{cases} \tag{24}$$

with $\mathcal{D} = \mathcal{G} \times_1 \mathbf{A}_1 \cdots \times_D \mathbf{A}_D + \mathcal{Y} - \frac{\Lambda}{\beta}$.

4) Update Λ: According to ADMM, the dual variable can be updated by

$$\Lambda = \Lambda + \beta(\mathcal{X} - \mathcal{G} \times_1 \mathbf{A}_1 \times \cdots \times_D \mathbf{A}_D - \mathcal{Y}). \tag{25}$$

To be clearer, we summarize the detailed solutions in algorithm 1, namely Trainable Subspaces for Low Rank Tensor Completion (TS4LRTC). Noted that all sub-problems in equation (17) are convex and can be solved with guaranteed convergence. In fact, according to the convergence conditions in chapter 3 of ADMM [50], the convergence of the algorithm 1 can be guaranteed. The convergence condition is reached when the relative error between two successive tensors $\mathcal{X}$ is smaller than a threshold ϕ.

Algorithm 1: Trainable Subspaces for Low Rank Tensor Completion

Input: $\mathcal{T} \in \mathbb{R}^{I_1 \times \cdots \times I_D}$ and index set $\mathbb{O}$,
$\mathbf{M}_d \in \mathbb{R}^{I_d \times J_d}, d = 1, \cdots, D$
Initialization maxiter =100, $\phi = 10^{-6}$, $\mathcal{X}_{\mathbb{O}} = \mathcal{T}_{\mathbb{O}}$
For $d = 1, \cdots, D$ **do**
  update $\mathbf{A}_d$ via dictionary learning on $\mathbf{M}_d$
  Orthogonalize $\mathbf{A}_d$
End for
For iter $= 1, \cdots,$ maxiter **do**
  $\tilde{\mathcal{X}} = \mathcal{X}$
  update $\mathcal{G}$ via (18)
  update $\mathcal{Y}$ via (22)
  update $\mathcal{X}$ via (24)
  update Λ via (25)
  If $\|\mathcal{X} - \tilde{\mathcal{X}}\|_{\mathrm{F}}^2 / \|\tilde{\mathcal{X}}\|_{\mathrm{F}}^2 < \phi$
    break
  End if
End for
Output: recovered tensor $\mathcal{X}$.

C. Computational complexity analysis

The main computational complexity of the TS4LRTC algorithm lies in the update of $\mathcal{G}$ and $\mathcal{Y}$. For each updated tensor, an SVD of the D unfolding matrices is needed. Assuming that $K_d = K$ and $I_d = I$ for $d = 1, \cdots, D$, the size of $\mathbf{G}_{(d)}$ is $K \times K^{D-1}$ and that of $\mathbf{Y}_{(d)}$ is $I \times I^{D-1}$. Consequently, the computational complexities of updating $\mathcal{G}$ and $\mathcal{Y}$ are $O(DK^{D+1})$ and $O(DI^{D+1})$, respectively. Therefore, the overall complexity is $O(\max(PDK^{D+1}, PDI^{D+1}))$, where P is the number of iterations. Especially, with fully shared subspace information, the computation on $\mathcal{Y}$ is not needed, and the computational complexity is dominated by the update of $\mathcal{G}$, resulting in $O(PDK^{D+1})$. In this case, the complexity of the proposed algorithm becomes very attractive when $K \ll I$.

D. Effect of subspace information on sample complexity

In this section, we analyze how subspace information impacts the sample complexity of tensor completion from a theoretical perspective. The idea is to bound the generalization fitting error of observations with the help of subspace information.

Assuming $\|\mathcal{G}\|_{\mathrm{F}} \leq G, \|\mathcal{Y}\|_{\mathrm{F}} \leq Y$, according to equation (2), problem (16) is relaxed to the following form:

$$\begin{aligned} \min_{\mathcal{G},\mathcal{Y}} & \sum_{i_1,\cdots,i_D \in \mathbb{O}} \frac{1}{2} \| \mathcal{X}(i_1, \cdots, i_D) \\ & - \mathcal{G} \times_1 \mathbf{A}_1(i_1) \times \cdots \times_D \mathbf{A}_D(i_D) - \mathcal{Y}(i_1, \cdots, i_D) \|_{\mathrm{F}}^2 \\ & \text{s. t. } \|\mathcal{G}\|_* \leq \sqrt{R_{\mathcal{G}}} G, \|\mathcal{Y}\|_* \leq \sqrt{R_{\mathcal{Y}}} Y, \end{aligned} \tag{26}$$

where $R_{\mathcal{G}}$ and $R_{\mathcal{Y}}$ are the maximal Tucker rank of $\mathcal{G}$ and $\mathcal{Y}$, respectively. For simplicity, we use $\mathrm{L}(\mathcal{F}_{\mathcal{Q}}(i_1, \cdots, i_D), \mathcal{X}(i_1, \cdots, i_D))$ to represent the loss function.

Let $\mathcal{F}_{\mathcal{Q}}(i_1, \cdots, i_D) = \mathcal{G} \times_1 \mathbf{A}_1(i_1) \times \cdots \times_D \mathbf{A}_D(i_D) + \mathcal{Y}(i_1, \cdots, i_D)$ be the estimation function for $\mathcal{X}(i_1, \cdots, i_D)$ parameterized by $\mathcal{Q}$, $\mathcal{Q} \in \mathbb{Q} = \{(\mathcal{G}, \mathcal{Y}) | \|\mathcal{G}\|_* \leq \sqrt{R_{\mathcal{G}}} G, \|\mathcal{Y}\|_* \leq \sqrt{R_{\mathcal{Y}}} Y\}$ and $\mathbb{F}_{\mathbb{Q}} = \{\mathcal{F}_{\mathcal{Q}} | \mathcal{Q} \in \mathbb{Q}\}$ be the set of feasible functions. Before giving the analysis results, we are interested in two L-risk quantities:

Expected L-risk:

$$\mathrm{Risk}_{\mathrm{L}}(\mathcal{F}_{\mathcal{Q}}) = \mathbb{E}_{i_1,\cdots,i_D}[\mathrm{L}(\mathcal{F}_{\mathcal{Q}}(i_1, \cdots, i_D), \mathcal{X}(i_1, \cdots, i_D))]. \tag{27}$$

Empirical L-risk:

$$\widehat{\mathrm{Risk}}_{\mathrm{L}}(\mathcal{F}_{\mathcal{Q}}) = \frac{1}{|\mathbb{O}|} \sum_{i_1,\cdots,i_D} \mathrm{L}(\mathcal{F}_{\mathcal{Q}}(i_1, \cdots, i_D), \mathcal{X}(i_1, \cdots, i_D)), \tag{28}$$

where $i_1, \cdots, i_D \in \mathbb{O}$ is the observed entries.

Next, we provide a bound on the expected L-risk:

Lemma *1:* (Bound on expected L-risk [51]) Let L be a loss function with Lipschitz constant δ_{L} and bounded by B, that is

$$\sup_{\mathcal{F}_{\mathcal{Q}} \in \mathbb{F}_{\mathbb{Q}}} \mathrm{L}(\mathcal{F}_{\mathcal{Q}}(i_1, \cdots, i_D), \mathcal{X}(i_1, \cdots, i_D)) \leq B, \tag{29}$$

Let $\mathfrak{R}(\mathbb{F}_{\mathbb{Q}})$ be the Rademacher model complexity of the function class $\mathbb{F}_{\mathbb{Q}}$, defined as

$$\begin{aligned} \mathfrak{R}(\mathbb{F}_{\mathbb{Q}}) = \mathbb{E}_{\mathcal{H}}[& \sup_{\mathcal{F}_{\mathcal{Q}} \in \mathbb{F}_{\mathbb{Q}}} \frac{1}{|\mathbb{O}|} \sum_{i_1,\cdots,i_D \in \mathbb{O}} \\ & \mathcal{H}(i_1, \cdots, i_D) \mathrm{L}(\mathcal{F}_{\mathcal{Q}}(i_1, \cdots, i_D), \mathcal{X}(i_1, \cdots, i_D))], \end{aligned} \tag{30}$$

where the entries of $\mathcal{H}$ take values $\{+1, -1\}$ with equal probability. Then for any constant C where $0 < C < 1$, with probability at least $1 - C$, for all $\mathcal{F}_{\mathcal{Q}} \in \mathbb{F}_{\mathbb{Q}}$, we have

$$\mathrm{Risk}_{\mathrm{L}}(\mathcal{F}_{\mathcal{Q}}) \leq \widehat{\mathrm{Risk}}_{\mathrm{L}}(\mathcal{F}_{\mathcal{Q}}) + 2\mathbb{E}_{\mathbb{O}}[\mathfrak{R}(\mathbb{F}_{\mathbb{Q}})] + B\sqrt{\frac{\log \frac{1}{C}}{2|\mathbb{O}|}}. \tag{31}$$

Next, we need to bound $\widehat{\mathrm{Risk}}_{\mathrm{L}}(\mathcal{F}_{\mathcal{Q}})$ and $\mathbb{E}_{\mathbb{O}}[\mathfrak{R}(\mathbb{F}_{\mathbb{Q}})]$ to guarantee that $\mathrm{Risk}_{\mathrm{L}}(\mathcal{F}_{\mathcal{Q}})$ is small enough. To show that the upper bound of $\mathbb{E}_{\mathbb{O}}[\mathfrak{R}(\mathbb{F}_{\mathbb{Q}})]$ is related to the subspace information, the lemma below is needed.

Lemma *2:* Let $K_d = K, I_d = I, d = 1, \cdots, D$, the upper bound of $\mathbb{E}_{\mathbb{O}}[\mathfrak{R}(\mathbb{F}_{\mathbb{Q}})]$ is obtained by

$$\mathbb{E}_{\mathbb{O}}[\mathfrak{R}(\mathbb{F}_{\mathbb{Q}})] \leq \delta_{\mathrm{L}} C_{\mathcal{G}} G \sqrt{\frac{DR_{\mathcal{G}} \log K}{|\mathbb{O}|}} + \delta_{\mathrm{L}} C_{\mathcal{Y}} Y \sqrt{\frac{DR_{\mathcal{Y}} \log I}{|\mathbb{O}|}} \tag{32}$$

where δ_{L} is Lipschitz constant , $C_{\mathcal{G}}$ and $C_{\mathcal{Y}}$ are constants and $\|\mathcal{G}\|_{\mathrm{F}} \leq G, \|\mathcal{Y}\|_{\mathrm{F}} \leq Y$. Finally, the upper bound on $\mathbb{E}_{\mathbb{O}}[\mathfrak{R}(\mathbb{F}_{\mathbb{Q}})]$

is obtained thanks to the lower bound:

$$|\mathbb{O}| \geq \max(G^2 D R_{\mathcal{G}} \log K, Y^2 D R_{\mathcal{Y}} \log I). \tag{33}$$

Proof *2:* See Appendix B.

In this case, the sample complexity of our method is $O(\max(R_{\mathcal{G}} \log K, R_{\mathcal{Y}} \log I))$ with $\|\mathcal{G}\|_{\mathrm{F}} \leq G, \|\mathcal{Y}\|_{\mathrm{F}} \leq Y$. It is noted that the sample complexity will reduce to $O(R_{\mathcal{G}} \log K)$ with fully shared subspace information, which will greatly improve recovery performance when a few observations are available.

IV. Experimental results

To evaluate the proposed method, we conduct several groups of experiments on synthetic and real data. The experiments on synthetic data are conducted to validate the effect of subspace information on sample complexity. In addition, two experiments on Multispectral image (MSI) and color image recovery are conducted to further evaluate the recovery performance of the proposed method.

In order to compare tensor completion performance, we select the nine state-of-the-art algorithms listed below:

- HaLRTC [32]: a low rank method based on Tucker rank minimization model.
- LRTF [52]: a low rank method based on t-SVD rank minimization model by using tensor logarithmic norm to induce a sparsity-driven surrogate for rank.
- GDC [53]: a geometry-inspired approach via assuming that the recovered tensor can be approximated by a conical hull of sub-tensors in the observation space.
- TR-VBI [54]: a Bayesian low rank tensor ring method by automatically learning the low-rank structure of tensor.
- CTMNM [27]: a coupled low rank tensor completion method based on Tucker rank minimization model.
- ACMTF [25]: a coupled low rank tensor completion method based on CP factorization.
- TenHet [40]: a coupled low rank tensor completion based on CP factorization with multi-view subspace information.
- CCPNM [29]: a coupled low rank tensor completion method based on CP rank minimization.
- TranSpa [45]: a low rank and fiber-wise sparsity based method by TT rank minimization model and sparsity representation.

Among them, HaLRTC, LRTF, GDC, and TR-VBI are without auxiliary information. In addition, to evaluate the performance qualitatively, three metrics are used to evaluate the recovery performance in tensor completion task, including relative standard error (RSE), tensor completion score (TCS) [9] and CPU time. The details are as follows

$$\mathrm{RSE} = \frac{\|\mathcal{X} - \mathcal{T}\|_{\mathrm{F}}}{\|\mathcal{T}\|_{\mathrm{F}}}, \tag{34}$$

$$\mathrm{TCS} = \frac{\|(1-\mathcal{O}) * (\mathcal{X} - \mathcal{T})\|_{\mathrm{F}}}{\|(1-\mathcal{O}) * \mathcal{T}\|_{\mathrm{F}}}, \tag{35}$$

where $\mathcal{O}$ is the indication tensor of the observations, $*$ represents element-wise product, $\mathcal{X}$ is the recovered tensor, and $\mathcal{T}$ is the real tensor.

The sampling ratio (SR) of the testing data is

$$\mathrm{SR} = \frac{|\mathbb{O}|}{\prod_{d=1}^{D} I_d}. \tag{36}$$

A. Experiments on synthetic data

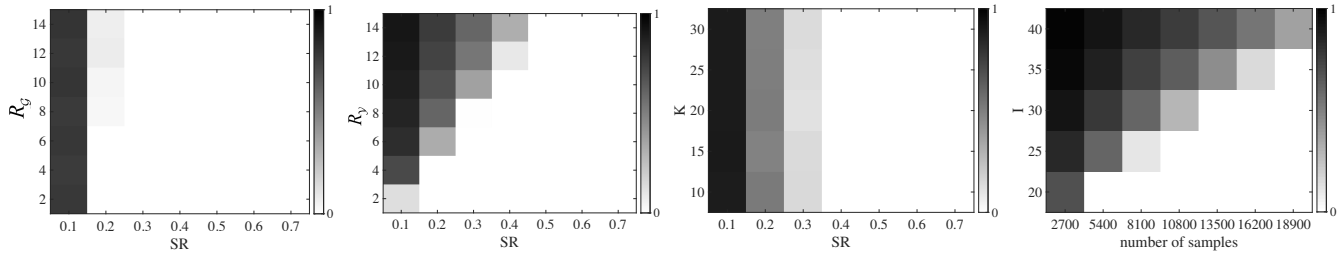
In this section, we conduct experiments on 3rd-order synthetic data $\mathcal{T} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ to verify the effect of subspace information on the sample complexity. Considering that there are four different subspace information conditions including fully shared, partially shared, both fully and partially shared, and no shared subspace information, we design four cases to generate the synthetic data.

- Case 1: partially shared subspace information
  The synthetic data $\mathcal{T}$ is generated by two low rank tensors $\mathcal{Z}$ and $\mathcal{Y}$, e.g. $\mathcal{T} = \mathcal{Z} + \mathcal{Y}$, where $\mathcal{Z}$ shares the subspace information with auxiliary data $\mathbf{M}_d \in \mathbb{R}^{I_d \times J_d}, d = 1, \cdots, 3$. The trainable subspace information matrix $\mathbf{A}_d \in \mathbb{R}^{I_d \times K_d}$ can be obtained by applying dictionary learning on $\mathbf{M}_d$, where we assume all $K_d = K$, $R_{\mathcal{G}1} = R_{\mathcal{G}2} = R_{\mathcal{G}3} = R_{\mathcal{G}}$ and $K > R_{\mathcal{G}}$. To create tensor $\mathcal{Z}$ with Tucker rank $[R_{\mathcal{G}1}, R_{\mathcal{G}2}, R_{\mathcal{G}3}]$, we need to generate a core tensor $\mathcal{B} \in \mathbb{R}^{R_{\mathcal{G}} \times R_{\mathcal{G}} \times R_{\mathcal{G}}}$, with its elements randomly sampled from a Gaussian distribution, and three matrices $\mathbf{U}_d \in \mathbb{R}^{K_d \times R_{\mathcal{G}}}$, where each column of $\mathbf{U}_d$ is sampled from $\mathbf{A}_d$ without repetition. In this case, we can generate tensors $\mathcal{Z} = \mathcal{B} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3$. Following the way to generate $\mathcal{Z}$, we can create tensor $\mathcal{Y}$ with Tucker rank $[R_{\mathcal{Y}}, R_{\mathcal{Y}}, R_{\mathcal{Y}}]$, where entries of factor matrices are randomly sampled from a Gaussian distribution.
- Case 2: both fully and partially shared subspace information
  In this case, we assume the synthetic data $\mathcal{T}$ and auxiliary data $\mathbf{M}_d \in \mathbb{R}^{I_d \times J_d}, d = 1, \cdots, 3$ share fully subspace information along mode-1 and mode-2 and partially subspace information along mode-3. Different from case 1, the core tensor of $\mathcal{Y}$ is $\mathcal{S} \in \mathbb{R}^{R_{\mathcal{G}} \times R_{\mathcal{G}} \times R_{\mathcal{Y}}}$ with its elements randomly sampled from a Gaussian distribution, and the first two factor matrices of $\mathcal{Y}$ are the same with that of $\mathcal{Z}$.
- Case 3: fully shared subspace information
  In this case, the latent factors of synthetic data $\mathcal{T}$ and auxiliary data $\mathbf{M}_d$ are the same, and $\mathcal{T} = \mathcal{Z}$ where the way to generate $\mathcal{Z}$ is same with that in case 1.
- Case 4: no shared subspace information
  Without subspace information, the synthetic data $\mathcal{T}$ is generated only by $\mathcal{Y}$, e.g. $\mathcal{T} = \mathcal{Y}$, where the details of generating $\mathcal{Y}$ can follow case 1.

For each case, we repeat experiments 10 times. The results are as follows.

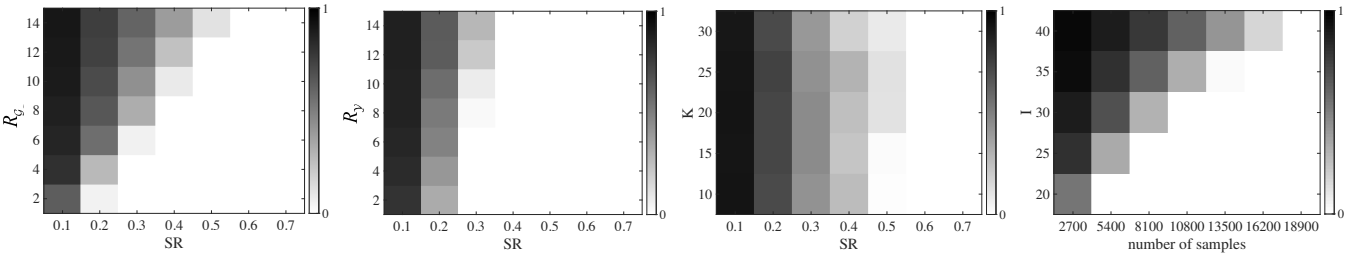
1) Partially shared subspace information: To study the relationship between partial subspace information and sample complexity, the parameters in Lemma 2 are set with one changed and others fixed as follows.

- K is selected in {10, 15, 20, 25, 30};
- $R_{\mathcal{G}}$ is selected in {2, 4, 6, 8, 10, 12, 14};
- $R_{\mathcal{Y}}$ is selected in {2, 4, 6, 8, 10, 12, 14};



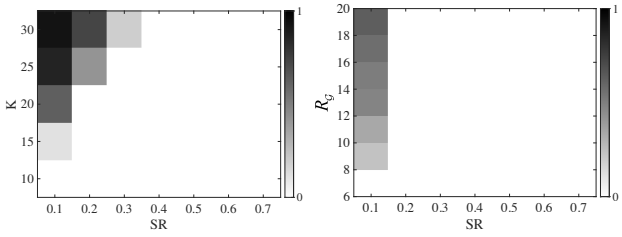
(a) RSE vs. $R_{\mathcal{G}}$ (b) RSE vs. $R_{\mathcal{Y}}$ (c) RSE vs. K (d) RSE vs. I

Fig. 3. The results for the effect of partially shared subspace information on sample complexity.



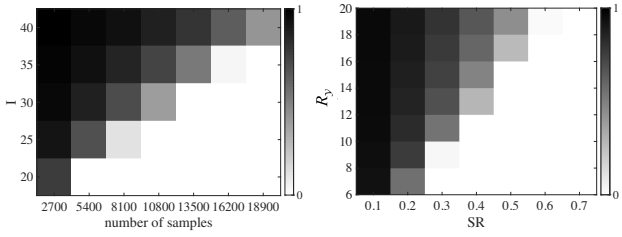
(a) RSE vs. $R_{\mathcal{G}}$ (b) RSE vs. $R_{\mathcal{Y}}$ (c) RSE vs. K (d) RSE vs. I

Fig. 4. The results for the effect of both partially and fully shared subspace information on sample complexity.



(a) RSE vs. K (b) RSE vs. $R_{\mathcal{G}}$

Fig. 5. The results for the effect of fully shared subspace information on sample complexity.



(a) RSE vs. I (b) RSE vs. $R_{\mathcal{Y}}$

Fig. 6. The results for the effect of no shared subspace information on sample complexity.

- I is selected in $\{20, 25, 30, 35, 40\}$.

When we conduct this group of experiments, other parameters are set to be: $K = 20; R_{\mathcal{G}} = 5; R_{\mathcal{Y}} = 5; I = 40$. Besides, when I is fixed, we set SR changing from 0.1 to 0.7 and when I is changing, the number of observed samples are chosen in $\{2700, 5400, 8100, 10800, 13500, 16200, 18900\}$.

Fig. 3 shows the RSE on four different conditions, in which the color of each cell reflects the recovery rate ranging from 0 to 1. A white cell means a success (RSE$< 10^{-3}$) in all experiments and a black cell means a failure. From Fig. 3 (a)-(b), we can observe when K and SR keep unchanged, the change of sample complexity is mainly depended on $R_{\mathcal{Y}}$. Meanwhile, with the number of samples fixed, the recovery performance will decrease with I keeping increased. This is because the bound of sample complexity mainly depends on $O(R_{\mathcal{Y}} \log I)$ with partially shared subspace information, where the bound will grow with I or $R_{\mathcal{Y}}$ increasing. This phenomenon is consistent with the conclusion of Lemma 2.

2) Both fully and partially shared subspace information: In this case, the parameter settings are the same in case 1. From Fig. 4 (a)-(b), we can observe the smaller $R_{\mathcal{G}}$ and $R_{\mathcal{Y}}$ are, the smaller the RSE is. Compared with Fig. 3 (a)-(b), the influence of $R_{\mathcal{G}}$ on the change of sample complexity becomes larger, with both fully and partially shared subspace information. In this case, the bound of sample complexity depends on the change of $R_{\mathcal{G}}, R_{\mathcal{Y}}$ and I.

3) Fully shared subspace information: For fully shared subspace information, the recovered tensor can be represented as $\mathcal{X} = \mathcal{Z}$. In this case, the sample complexity can be rewritten as $O(R_{\mathcal{G}} \log K)$. We can see the smaller the K or $R_{\mathcal{G}}$ is, the bound of sample complexity will be lower. To verify its reliability, we design two experiments with the values of parameters K and $R_{\mathcal{G}}$ as follows.

- K is selected in $\{10, 15, 20, 25, 30\}$;
- $R_{\mathcal{G}}$ is selected in $\{7, 9, 11, 13, 15, 17, 19\}$.

When we conduct this group of experiments, other parameters are set to be: $K = 20$; $R_{\mathcal{G}} = 10$; $R_{\mathcal{Y}} = 0$; $I = 40$. Fig. 5 illustrates the RSE on two different conditions with SR ranging from 0.1 to 0.7. From Fig. 5, it can be seen the value of RSE becomes larger with K increasing when SR $= 0.1$. And for

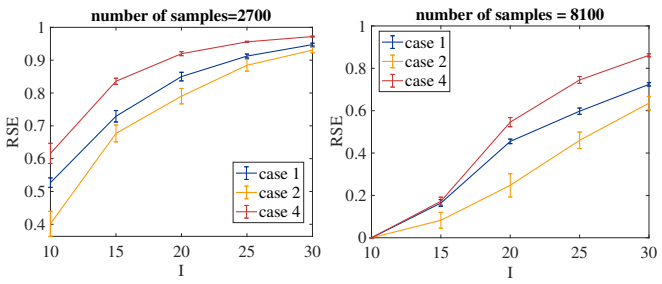


Fig. 7. Comparison results on three different cases with few samples.

(a) RGB-IR (b) RGB-Depth

Fig. 8. Testing color images with different types.

(a) Toy (b) Flowers

Fig. 9. Testing multispectral Images.

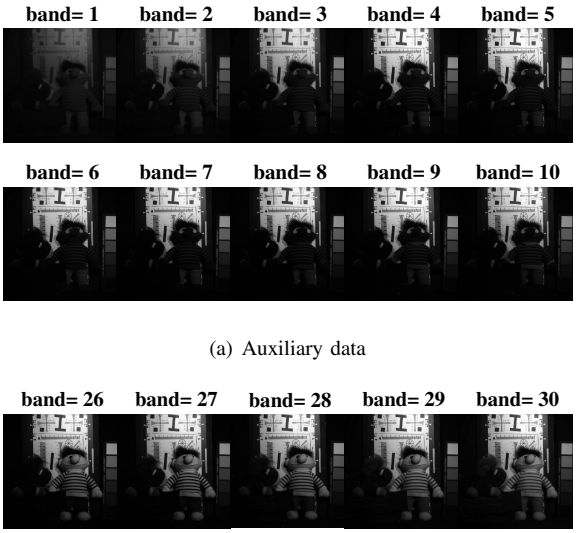
other SRs, the values of RSE are very small. This phenomenon is similar with $R_{\mathcal{G}}$ varied. Overally, fully shared subspace information has shown a great effect on sample complexity via the feature space size K.

4) No shared subspace information: Without the help of subspace information, the recovered tensor can be represented as $\mathcal{X} = \mathcal{Y}$. The key of recovering entries is based on its low-rank characteristics. In this case, the bound of sample complexity is $O(R_{\mathcal{Y}} \log I)$. Similarly, to explore the relationship between the rank and the data size, we set the parameters as follows.

- $R_{\mathcal{Y}}$ is selected in $\{7, 9, 11, 13, 15, 17, 19\}$;
- I is selected in $\{20, 25, 30, 35, 40\}$.

When we conduct this group of experiments, other parameters are set to be: $K = 20; R_{\mathcal{G}} = 5; R_{\mathcal{Y}} = 5; I = 40$. Besides, when I is changing, the number of observed samples are chosen in $\{2700, 5400, 8100, 10800, 13500, 16200, 18900\}$, and when I is fixed, we set SR changing from 0.1 to 0.7.

Fig. 6 demonstrates the change of RSE with different number of samples on different conditions. It can be seen that with number of samples fixed, RSE gets larger with the data size enlarged in Fig. 6(a). Fig. 6(b) shows that the smaller $R_{\mathcal{Y}}$, the easier the recovery of missing entries. In addition, comparing the results in Fig. 6(b) with that in Fig. 5(b), we



(a) Auxiliary data

(b) Observed data

Fig. 10. An example of sample images with b=5 and p=10.

can observe that for tensors with the same size and rank, the required number of samples to successfully recover missing entries has been greatly reduced with the help of fully shared subspace information.

In addition, Fig. 7 shows the comparison results on case 1, case 2 and case 4 with fixed number of samples to illustrate the effect of subspace information. From it, we can find the more the subspace information shares, the smaller the RSE.

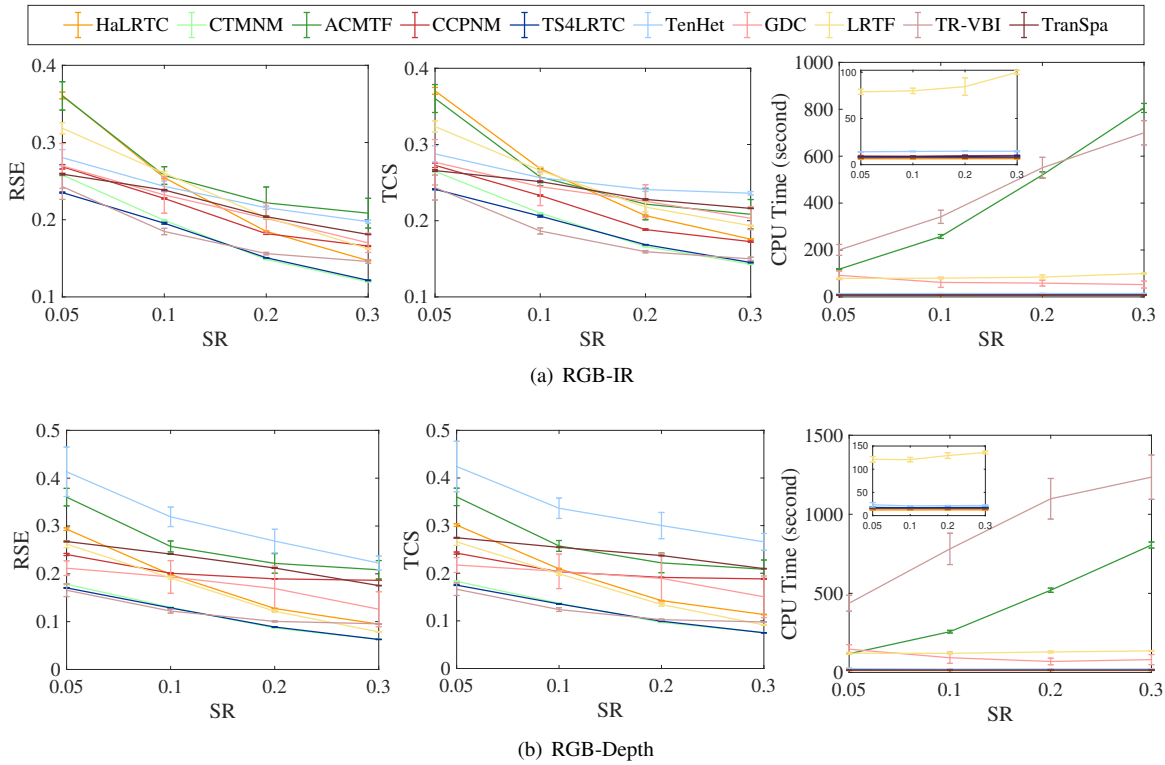
To conclude, the sample complexity of our algorithm is related to the data size I, the real rank of data R and the size of feature spaces K for partially shared subspace information. Interestingly, for fully shared subspace information, the sample complexity only relies on the real rank of data R and the size of feature spaces K. When $R < K \ll I$, the subspace information has shown a great advantage on sample complexity.

B. Experiments on real data

In this part, we use color images and MSIs as our testing data to show the effect of subspace information on tensor completion and the effect of our algorithm. All experiments are repeatedly carried 10 times.

Two types of images are chosen for testing the recovery performance of our algorithm. One is from the color remapping Gecko imagery [55], which is in Fig. 8 (a), where the RGB image $\mathcal{T} \in \mathbb{R}^{365\times488\times3}$ as our observed data and the near-infrared image $\mathbf{M} \in \mathbb{R}^{365\times488}$ as auxiliary data. The other is from NYU Depth Dataset [56], which comprises pairs of RGB and depth frame, where the RGB image $480 \times 640 \times 3$ and depth image 480×640 are as our observed data and auxiliary data, respectively, as it shows in Fig. 8 (b).

In the MSI completion task, we choose two MSIs from multispectral Image Database [57]. For each MSI, the spatial resolution is 512×512 and the number of spectral bands is 30, resulting in a 3rd order tensor $512 \times 512 \times 30$. The RGB representations of MSIs is shown in Fig. 9.



(a) RGB-IR

(b) RGB-Depth

Fig. 11. Comparison of different methods on color image recovery with different SRs from 0.05 to 0.3.

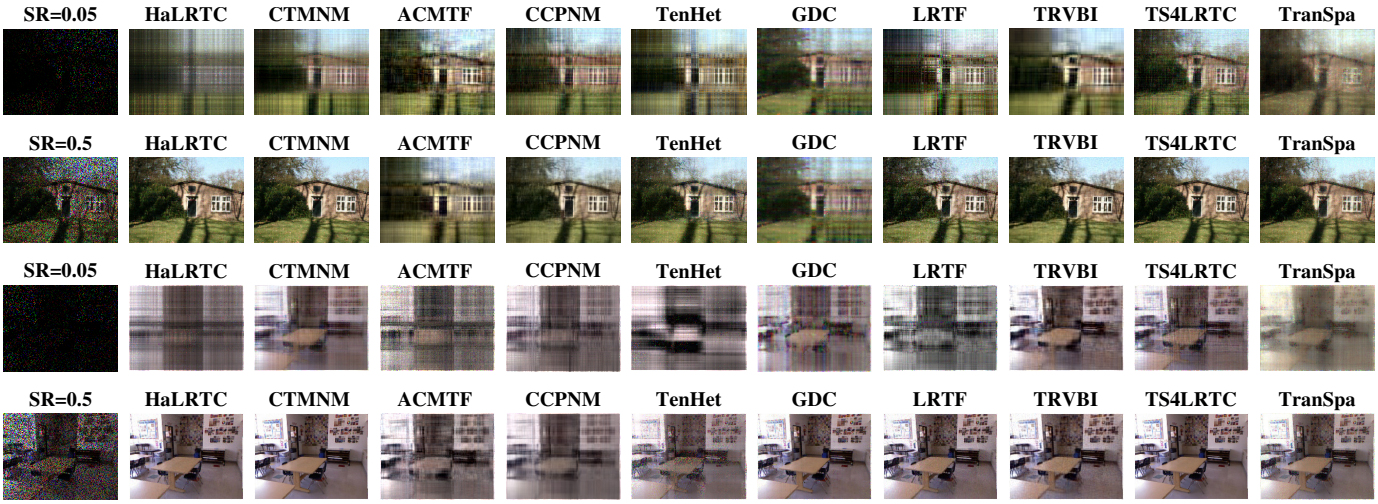


Fig. 12. Examples on image recovery with different methods when SR=0.05 and SR=0.5.

In this task, we choose the last b bands from the datasets as observed data and p bands from the rest as auxiliary data, where b is the number of observed bands and is chosen from the set $\{5,10,15\}$. p is the number of auxiliary bands and is chosen from the set $\{5,\cdots,30\text{-}b\}$. An example of b=5 and p=10 can be seen in Fig.10.

1) Color image recovery: For each observed data, we randomly choose the SR ranging from 0.05 to 0.5. And with the known auxiliary data, we can obtain its corresponding trainable subspaces. The parameter β in our algorithm is set as $0.5/\|\mathcal{T}_{\mathbb{O}}\|_{\mathrm{F}}$ and others are chosen according to section IV-C1. For HaLRTC, we set the weight $\mathbf{w} = \mathbf{a}/\|\mathbf{a}\|_1$ with $\mathbf{a} = [1;1;1\times10^{-3}]$ as in [32]. As suggested in [27], the parameters of CTMNM are set as $\mathbf{w} = \mathbf{a}/\|\mathbf{a}\|_1$ with $\mathbf{a} = [365;488;3]$ for "RGB-IR" images and $\mathbf{a} = [480;640;3]$ for "RGB-Depth" images. The CP rank for ACMTF are set as $R = 20$ by tuning for the recovery performance. Following [27], the bound for CCPNM is set to 800. Followed by [40], we set the parameters in TenHet as $\beta = 0.001;\gamma = 0.01;\lambda = 0.01$. In addition, the rank settings for GDC and LRTF are chosen from the set $\{5{:}5{:}100\}$ with different missing ratios. The initial rank for TR-VBI is set as 30. The γ in TranSpa is set to 0.01.

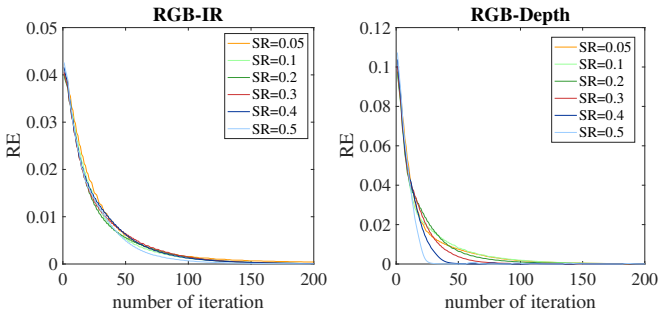


Fig. 13. The convergence performance for color image.

Fig. 11 shows the recovery performance of different methods with SR ranging from 0.05 to 0.3 in terms of RSE, TCS and CPU time. It can be observed our method outperforms state-of-the-art ones when SR = 0.05 for "RGB-IR" images. When SR = 0.1, the recovery performance of TR-VBI performs best in terms of RSE and TCS, but its computational complexity is larger than the others. As SR increases, the recovery performance of CTMNM becomes superior to others in terms of RSE and TCS, and our method performs the second-best. For "RGB-Depth" images, the performance of the TS4LRTC and CTMNM are almost the same and better than that of the others.

Fig. 12 presents the comparison of recovery performance using HaLRTC, CTMNM, ACMTF, CCPNM, TS4LRTC TenHet, GDC, LRTF and TR-VBI when SR = 0.05 and SR = 0.5, respectively. We can see for "RGB-IR" images, most methods with subspace information including CTMNM, ACMTF, CCPNM,TR-VBI, TranSpa, and TS4LRTC can recover image with SR = 0.05. Among them, our method provides a more detailed information, e.g., the framework of house. For "RGB-Depth" images, only CTMNM, TR-VBI and TS4LRTC can recover the destroyed image. It may imply that the auxiliary data of "RGB-IR" images provides more subspace information than that of "RGB-Depth" images and tensor ring rank has a good ability to exploit the low rank information in an image.

Fig.13 shows the convergence performance of the method we propose on color images with random SRs from 0.05 to 0.5. The relative error (RE) is used to measure the convergence condition, where RE=$\frac{\mathcal{X}^q-\mathcal{X}^{q-1}}{\mathcal{X}^{q-1}}$ and $\mathcal{X}^q$ is the recovered image in q-th iteration. From Fig.13 , we could observe all cases converge to 0 at last. Especially, the larger the SR is, the faster our method converges.

2) MSI recovery: Similar to color image recovery, we set the SR varying from 0.02 to 0.5 and obtain the trainable subspaces via applying dictionary learning on the auxiliary data. We set the parameter $\beta = 0.05/\|\mathcal{T}_{\mathbb{O}}\|_{\mathrm{F}}$. $\lambda_{\mathcal{G}}$ and $\lambda_{\mathcal{Y}}$ are set according to section IV-C1. The parameters of other algorithms are chosen by following the image recovery experiments.

Fig. 14 shows the performance of our method and some other state-of-the-art algorithms in terms of RSE, TCS and CPU time with b fixed and p varied when SR = 0.02, e.g. when $b = 5$ for the HSI Flowers, p varies from $\{5,10,15,20,25\}$. For these methods without auxiliary information, p is set to 0. It shows that the RSE of theses methods including CCPNM, CTMNM, TS4LRTC, TenHet and TranSpa changes along p, e.g. the larger p, the better the recovery results. It may imply that a higher reconstruction accuracy will be obtained with more auxiliary information provided. Among them, TS4LRTC performs best in terms of TCS and RSE. In addition, the recovery performance of TS4LRTC and CCPNM with p = 5 and p = 10 is better than that of other number of bands when $b = 5$ and $b = 10$, respectively. This is because the observed data and the auxiliary data are coupled in each mode when $b = p$. It may imply that the more information is shared in the subspace, the better the performance for TS4LRTC and CCPNM. Moreover, we can see TR-VBI performs well in terms of RSE and TCS, but the CPU time is very large, especially for b = 10. This may infer that low rank tensor ring factorization can explore more latent information from the observed data than traditional tensor decomposition.

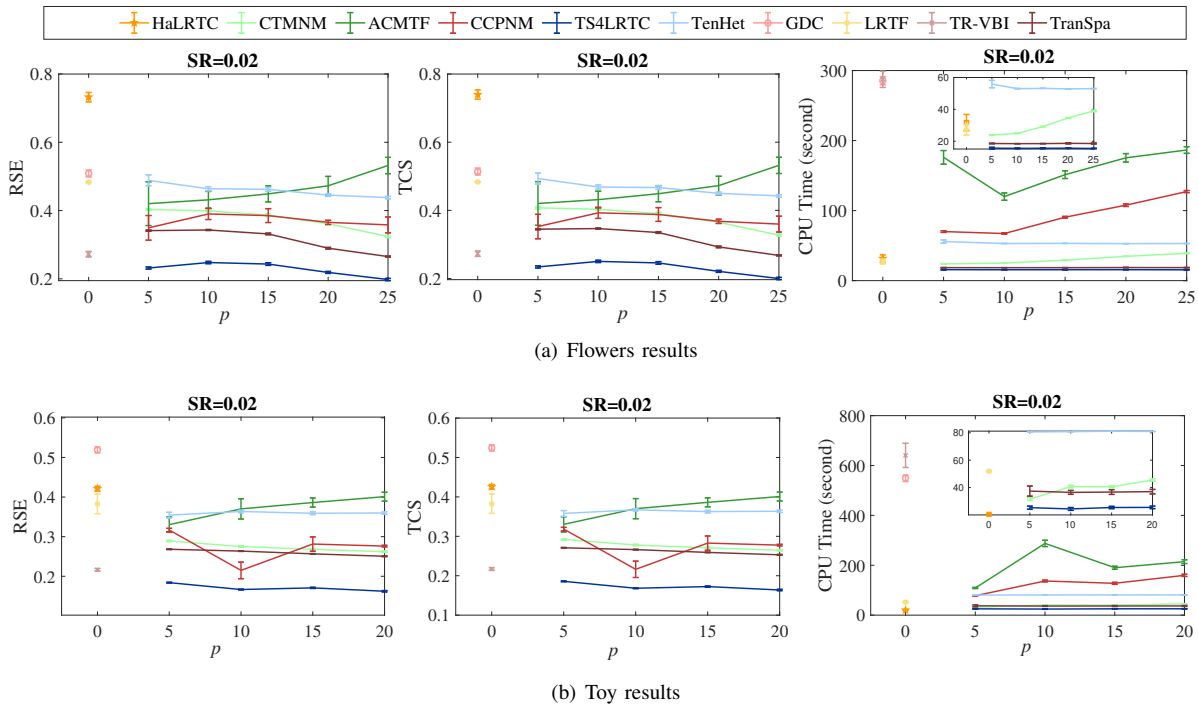
To clearly observe the recovery details, we presented the recovered MSI with different algorithms when SR = 0.02 and SR = 0.3 in Fig. 15. From this figure, we could find ACMTF, CCPNM, TenHet, TR-VBI, TranSpa and TS4LRTC can recover the outline of flowers and the shape of toy when SR = 0.02. Among these methods, the one we propose performs the best, which provides more details about the recovered data. For example, the shape of petals and the background of toy can be clearly observed. With large SR, all methods can successfully recover the missing entries.

Fig.16 shows the convergence results for "Flowers" and "Toy" with SR from 0.02 to 0.4. From it, we could observe all cases converge 0 at 100-th iteration. In addition, when SR= 0.4, our proposed one converges at 50-th iteration. It may imply our proposed one converges fast for large observed entries.

In conclusion, with the help of subspace information, the tensor recovery performance can be enhanced, especially when a few observations are available. Besides, compared color image recovery with MSI recovery, we could observe, the more information the auxiliary data provides, the better the recovery performance. Among all methods, the proposed one performs best with small SR, which may imply that the our proposed one provides a better way to explore and incorporate the subspace information.

C. Discussions

1) Discussions on $\lambda_{\mathcal{G}}$ and $\lambda_{\mathcal{Y}}$: In real-world experiments, we choose parameters from the $\{10^t\}_{t=-3}^{2}$ via brute force search and report the best recovery one. Fig.17 shows the change of RSE as a function of $\lambda_{\mathcal{G}}$ and $\lambda_{\mathcal{Y}}$ when SR=0.05. We could observe when $\lambda_{\mathcal{G}} \in [0.13, 0.59]$ and $\lambda_{\mathcal{Y}} \in [0.76, 27.82]$, the recovery performance on color images shows well. And for MSI with b =15 and p=15, it can be seen when $\lambda_{\mathcal{G}} \in [0.001, 0.59]$ and $\lambda_{\mathcal{Y}} \in [0.05, 100]$, the recovery results are satisfied. In addition, compared with Fig.17(a), we could observe the larger value $\lambda_{\mathcal{Y}}$ and the smaller value $\lambda_{\mathcal{G}}$, the better the performance. It may imply the auxiliary data of MSI shares more subspace information than that of color image, which is consistent with the results we obtained experimentally.



(a) Flowers results

(b) Toy results

Fig. 14. Comparison of different methods on MSI recovery with b fixed and p varied..

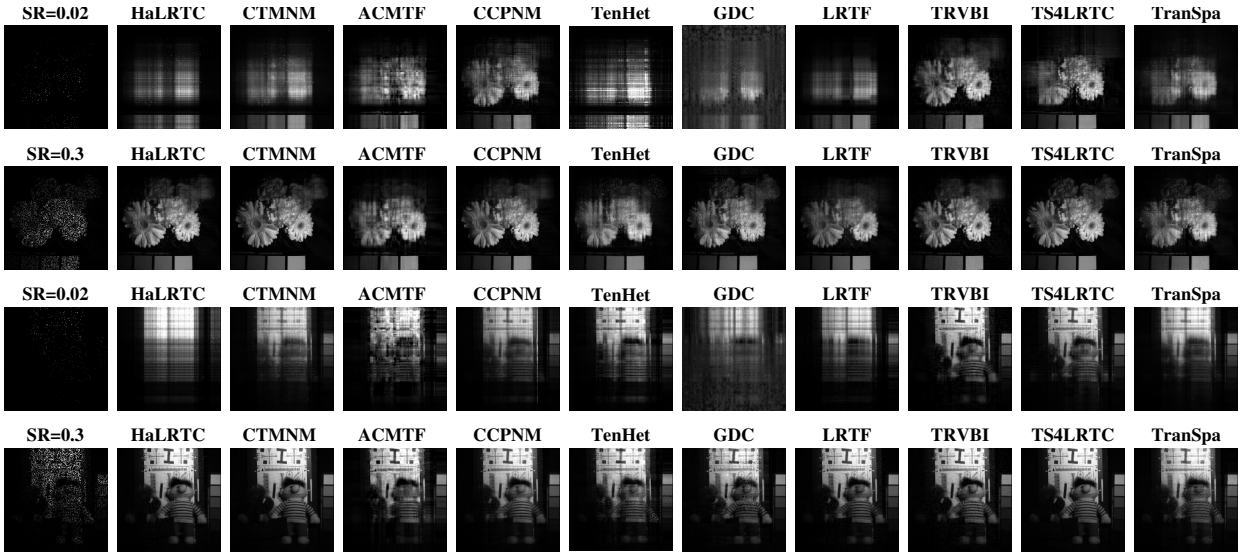


Fig. 15. Examples on MSI recovery using different methods with b=15 when SR=0.02 and SR=0.3.

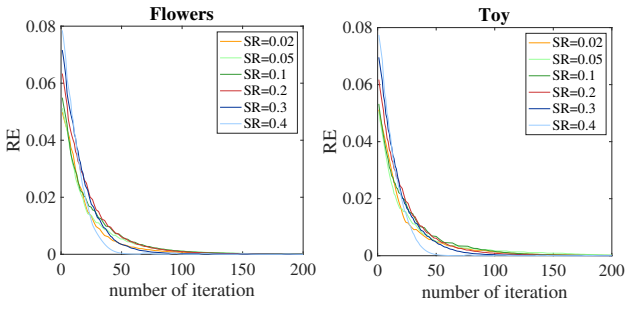
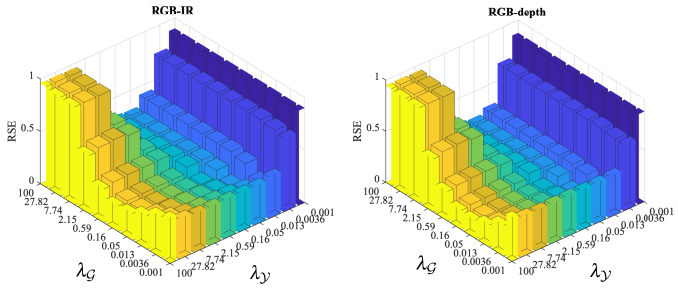


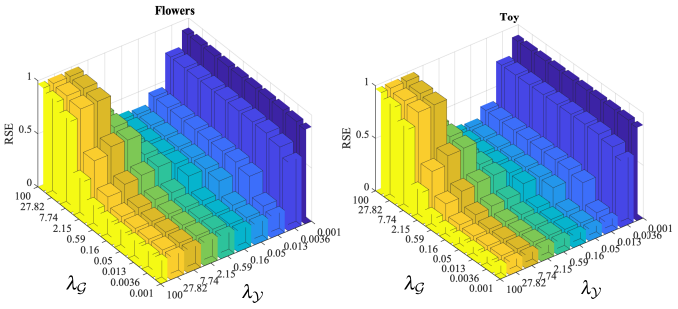
Fig. 16. The convergence performance for MSI.

2) Discussions on scalable method for TS4LRTC : The proposed method contains the Singular Value Thresholding (SVT) operator, which could be intractable with large scale data. To deal with it, Lanczos techniques [58], [59], [60], which allow to compute approximations of a restricted subset of singular triplets in a fast way, can be used for acceleration. More precisely, L dominant left singular vectors of a $M \times N$ matrix can be computed rather accurately within $O(2LMN)$ flops.

For better illustration, we have conducted a group of experiments on large size image completion, where these large



(a) RSE vs. $\lambda_{\mathcal{G}}$ and $\lambda_{\mathcal{Y}}$



(b) RSE vs. $\lambda_{\mathcal{G}}$ and $\lambda_{\mathcal{Y}}$

Fig. 17. The change of recovery performance with different parameters.

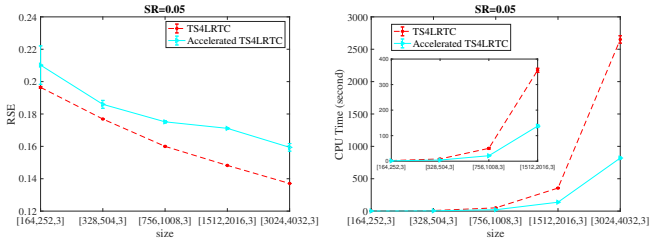
scale images with size $3024 \times 4032 \times 3$ are taken by iPhone, as shown in Fig. 18. Here the SVT operator is replaced by the fast randomized one [60]. In addition, comparison experiments on sub-sampled images of various sizes are also considered. Fig.19 and Fig. 20 show the recovery performance on images of different sizes in terms of RSE and CPU time when SR=0.05 and SR=0.1, respectively. Each experiment is repeated 10 times. We can observe from the results that the accelerated TS4LRTC performs slightly worse than TS4LRTC with an explicit improvement in CPU time consumption, especially with large scale data.

(a) observed data (b) auxiliary data

Fig. 18. Large size testing images.

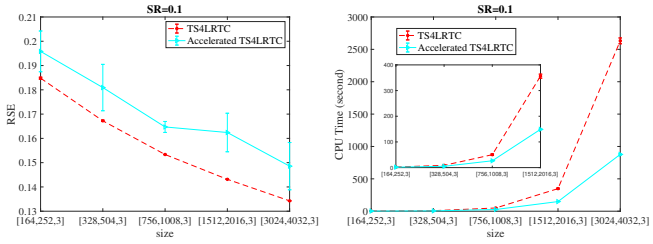
V. Conclusions

In this paper, we propose a trainable subspaces for low rank tensor completion model. Different from current tensor completion model with subspace information, the proposed one firstly considers a data-driven way to learn the subspace information and provides a new perspective to incorporate subspace in the tensor completion task. This idea may also



(a) RSE vs. size (b) CPU Time vs. size

Fig. 19. Recovery performance on different size when SR=0.05.



(a) RSE vs. size (b) CPU Time vs. size

Fig. 20. Recovery performance on different size when SR=0.1.

be helpful for other tensor processing tools, such as tensor classification, tensor robust PCA and tensor regression. In addition, the sample complexity of our method is provided to give a theoretical insight into the usefulness of subspace information. Experiments on synthetic data present that the subspace information can reduce the sample complexity for tensor completion, especially with fully shared subspace information. Besides, experiments on real world data show our algorithm is superior to state-of-the-art ones in terms of prediction accuracy and CPU time.

Appendix A
Proof of Theorem 1

Assuming $\mathcal{X} = [\![\mathcal{G}; \mathbf{A}_1, \cdots, \mathbf{A}_D]\!]$, where $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_D}$ and $\text{col}(\mathbf{X}_{(d)}) \subseteq \text{col}(\mathbf{A}_d)$, $\mathbf{A}_d, d = 1, \cdots, D$ be $I_d \times K_d$ orthonormal matrices, $\mathcal{G} \in \mathbb{R}^{K_1 \times \cdots \times K_D}$ and $R_d \leq K_d \leq I_d, d = 1, \cdots, D$, R_d is the rank of $\mathbf{X}_{(d)}$, each unfolding matrix $\mathbf{X}_{(d)}$ can be equivalent to

$$\mathbf{X}_{(d)} = \mathbf{A}_d \mathbf{G}_{(d)} (\mathbf{A}_{\boxtimes}^d)^{\mathrm{T}}, \tag{37}$$

where $\mathbf{A}_{\boxtimes}^d = \mathbf{A}_{d+1} \boxtimes \cdots \boxtimes \mathbf{A}_D \boxtimes \mathbf{A}_1 \boxtimes \cdots \boxtimes \mathbf{A}_{d-1} \in \mathbb{R}^{I_{\neq d} \times K_{\neq d}}$ with $I_{\neq d} = I_{d+1} \cdots I_D I_1 \cdots I_{d-1}$ and $K_{\neq d} = K_{d+1} \cdots K_D K_1 \cdots K_{d-1}$.

According to definition 4, the tensor trace norm of $\mathcal{X}$ is:

$$\|\mathcal{X}\|_* = \sum_{d=1}^{D} \alpha_d \|\mathbf{X}_{(d)}\|_*, \tag{38}$$

Then

$$\mathbf{G}_{(d)} = \mathbf{A}_d^{\mathrm{T}} \mathbf{X}_{(d)} \mathbf{A}_{\boxtimes}^d, \tag{39}$$

Applying SVD operator on $\mathbf{X}_{(d)}$, we can obtain

$$\mathbf{X}_{(d)} = \mathbf{U}\mathbf{S}\mathbf{V}^T. \tag{40}$$

Then the equation (39) can be rewritten as:

$$\mathbf{G}_{(d)} = \mathbf{A}_d^{\mathrm{T}}\mathbf{U}\mathbf{S}\mathbf{V}^T\mathbf{A}_{\boxtimes}^d, \tag{41}$$

According to [61], $\mathbf{A}_{\boxtimes}^d$ is also an orthonormal matrix, then $\mathbf{A}_d^{\mathrm{T}}\mathbf{U}$ and $\mathbf{V}^T\mathbf{A}_{\boxtimes}^d$ are semi-orthogonal matrices. And equation (41) can be regarded as the SVD of $\mathbf{G}_{(d)}$. In this case, the values of $\mathbf{S}$ keep the same. Therefore, $\|\mathbf{G}_{(d)}\|_* = \|\mathbf{X}_{(d)}\|_*$.

It can be easily obtained that $\|\mathcal{X}\|_* = \|\mathcal{G}\|_*$, where $\mathcal{X} = [\![\mathcal{G};\mathbf{A}_1,\cdots,\mathbf{A}_D]\!]$ with orthonormal matrices $\mathbf{A}_d, d = 1,\cdots,D$.

APPENDIX B
PROOF OF LEMMA 2

Now, we will give the bound of $\mathfrak{R}(\mathbb{F}_{\mathbb{Q}})$ as follows.

$$\begin{aligned}
&\mathfrak{R}(\mathbb{F}_{\mathbb{Q}}) = \mathbb{E}_{\mathcal{H}}[\sup_{\mathcal{F}_{\mathcal{Q}}\in\mathbb{F}_{\mathbb{Q}}} \frac{1}{|\mathbb{O}|}\sum_{i_1,\cdots,i_D\in\mathbb{O}} \\
&\mathcal{H}(i_1,\cdots,i_D)\mathrm{L}(\mathcal{F}_{\mathcal{Q}}(i_1,\cdots,i_D),\mathcal{X}(i_1,\cdots,i_D))] \\
=\ & \frac{1}{|\mathbb{O}|}\mathbb{E}_{\mathcal{H}}[\sup_{\|\mathcal{G}\|_*\le\sqrt{R_{\mathcal{G}}}G,\|\mathcal{Y}\|_*\le\sqrt{R_{\mathcal{Y}}}Y}\sum_{i_1,\cdots,i_D\in\mathbb{O}} \\
&\mathcal{H}(i_1,\cdots,i_D)\mathrm{L}(\mathcal{F}_{\mathcal{Q}}(i_1,\cdots,i_D),\mathcal{X}(i_1,\cdots,i_D))] \\
\le\ & \frac{\delta_{\mathrm{L}}}{|\mathbb{O}|}\mathbb{E}_{\mathcal{H}}[\sup_{\|\mathcal{G}\|_*\le\sqrt{R_{\mathcal{G}}}G,\|\mathcal{Y}\|_*\le\sqrt{R_{\mathcal{Y}}}Y}\sum_{i_1,\cdots,i_D\in\mathbb{O}} \\
&\mathcal{H}(i_1,\cdots,i_D)\mathcal{F}_{\mathcal{Q}}(i_1,\cdots,i_D)] \qquad (42)\\
=\ & \frac{\delta_{\mathrm{L}}}{|\mathbb{O}|}\mathbb{E}_{\mathcal{H}}[\sup_{\|\mathcal{G}\|_*\le\sqrt{R_{\mathcal{G}}}G}\sum_{i_1,\cdots,i_D\in\mathbb{O}} \\
&\mathcal{H}(i_1,\cdots,i_D)(\mathcal{G}\times_1\mathbf{A}_1(i_1)\times\cdots\times_D\mathbf{A}_D(i_D))] \\
+\ & \frac{\delta_{\mathrm{L}}}{|\mathbb{O}|}\mathbb{E}_{\mathcal{H}}[\sup_{\|\mathcal{Y}\|_*\le\sqrt{R_{\mathcal{Y}}}Y}\sum_{i_1,\cdots,i_D\in\mathbb{O}}\mathcal{H}(i_1,\cdots,i_D)\mathcal{Y}(i_1,\cdots,i_D)] \\
=\ & \frac{\delta_{\mathrm{L}}}{|\mathbb{O}|}\mathbb{E}_{\mathcal{H}}[\sup_{\|\mathcal{G}\|_*\le\sqrt{R_{\mathcal{G}}}G}\sum_{i_1,\cdots,i_D\in\mathbb{O}} \\
&\frac{1}{D}\sum_{d=1}^{D}\mathbf{H}_{(d)}(i_d,i_{\neq d})\mathbf{A}_d(i_d)\mathbf{G}_{(d)}\mathbf{A}_{\boxtimes}^d(i_{\neq d})] \\
+\ & \frac{\delta_{\mathrm{L}}}{|\mathbb{O}|}\mathbb{E}_{\mathcal{H}}[\sup_{\|\mathcal{Y}\|_*\le\sqrt{R_{\mathcal{Y}}}Y}\sum_{i_1,\cdots,i_D\in\mathbb{O}} \\
&\frac{1}{D}\sum_{d=1}^{D}\mathbf{H}_{(d)}(i_d,i_{\neq d})\mathbf{e}_{i_d}\mathbf{Y}_{(d)}\mathbf{e}_{i_{\neq d}}] \\
\le\ & \delta_{\mathrm{L}}\sup_{\|\mathcal{G}\|_*\le\sqrt{R_{\mathcal{G}}}G}\frac{1}{D}\sum_{d=1}^{D}\|\mathbf{G}_{(d)}\|_* \\
&\max_{i_1,\cdots,i_D\in\mathbb{O}}\|\mathbf{A}_{\boxtimes}^d(i_{\neq d})\mathbf{A}_d(i_d)\|\sqrt{\frac{2\log 2K_dK_{\neq d}}{|\mathbb{O}|}} \\
+\ & \delta_{\mathrm{L}}\sup_{\|\mathcal{Y}\|_*\le\sqrt{R_{\mathcal{Y}}}Y}\frac{1}{D}\sum_{d=1}^{D}\|\mathbf{Y}_{(d)}\|_* \\
&\max_{i_1,\cdots,i_D\in\mathbb{O}}\|\mathbf{e}_{i_{\neq d}}\mathbf{e}_{i_d}\|\sqrt{\frac{2\log 2I_dI_{\neq d}}{|\mathbb{O}|}} \qquad (43)
\end{aligned}$$

where δ_{L} is the Lipschitz constant of loss function L, and $i_{\neq d} = i_{d+1}\cdots i_D i_1\cdots i_{d-1}$, $\mathbf{A}_{\boxtimes}^d(i_{\neq d}) = \mathbf{A}_{d+1}(i_{d+1})\boxtimes\mathbf{A}_D(i_D)\boxtimes\mathbf{A}_1(i_1)\boxtimes\mathbf{A}_{d-1}(i_{d-1})$ and $\mathbf{G}_{(d)},\mathbf{H}_{(d)},\mathbf{Y}_{(d)}, d = 1,\cdots,D$ are the mode-n unfolding matrices of $\mathcal{G},\mathcal{H},\mathcal{Y}$, respectively. e_{i_d} is the vector with its i_d-th entry equal to 1 while all others equal to 0. The first inequality is derived by Rademacher contraction and the second one is derived from Lemma 3.

Lemma *3:* [62] Let $\mathbb{W} = \{\mathbf{W}\in\mathbb{R}^{I_1\times I_2}|\|\mathbf{W}\|_*\le W\}$and $A = \max_p\|\mathbf{A}_p\|$ where $\mathbf{A}_p\in\mathbb{R}^{I_2\times I_1}, p = 1,\cdots,P$, then

$$\mathbb{E}_{\delta}[\sup_{\mathbf{W}\in\mathbb{W}}\frac{1}{P}\sum_{p=1}^{P}\delta_p\,\mathrm{trace}(\mathbf{W}\mathbf{A}_p)] \le WA\sqrt{\frac{2\log(2I_1I_2)}{P}}. \tag{44}$$

Let $\max_{i_d}\|\mathbf{A}_d(i_d)\| = A, d = 1,\cdots D$, then

$$\max_{i_{\neq d}}\|\mathbf{A}_{\boxtimes}^d(i_{\neq d})\| = A^{D-1}, \tag{45}$$

Therefore, let $K_d = K, d = 1,\cdots,D$ and $I_d = K, d = 1,\cdots,D$, the upper bound of $\mathbb{E}_{\mathbb{O}}[\mathfrak{R}(\mathbb{F}_{\mathbb{Q}})]$ is obtained as follows.

$$\mathbb{E}_{\mathbb{O}}[\mathfrak{R}(\mathbb{F}_{\mathbb{Q}})] \le \delta_{\mathrm{L}}C_{\mathcal{G}}G\sqrt{\frac{DR_{\mathcal{G}}\log K}{|\mathbb{O}|}} + \delta_{\mathrm{L}}C_{\mathcal{Y}}Y\sqrt{\frac{DR_{\mathcal{Y}}\log I}{|\mathbb{O}|}} \tag{46}$$

where $C_{\mathcal{G}}$ and $C_{\mathcal{Y}}$ are the constants.

$$|\mathbb{O}| \ge \max(G^2DR_{\mathcal{G}}\log K, Y^2DR_{\mathcal{Y}}\log I). \tag{47}$$

with $\|\mathcal{G}\|_{\mathrm{F}}\le G, \|\mathcal{Y}\|_{\mathrm{F}}\le Y$.

REFERENCES

[1] Q. Yao, X. Chen, J. T. Kwok, Y. Li, and C.-J. Hsieh, "Efficient neural interaction function search for collaborative filtering," in *Proceedings of The Web Conference 2020*, 2020, pp. 1660–1670.

[2] H. Liu, Y. Li, M. Tsang, and Y. Liu, "Costco: A neural tensor completion model for sparse tensors," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 324–334.

[3] H. Chen and J. Li, "Neural tensor model for learning multi-aspect factors in recommender systems." in *IJCAI*, 2020, pp. 2449–2455.

[4] Y. Liu, Z. Long, and C. Zhu, "Image completion using low tensor tree rank and total variation minimization," *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 338–350, 2018.

[5] Z. Long, Y. Liu, L. Chen, and C. Zhu, "Low rank tensor completion for multiway visual data," *Signal Processing*, vol. 155, pp. 301–316, 2019.

[6] Z. Zhang, Y. Xie, W. Zhang, Y. Tang, and Q. Tian, "Tensor multi-task learning for person re-identification," *IEEE Transactions on Image Processing*, vol. 29, pp. 2463–2477, 2019.

[7] K. Wimalawarne, M. Sugiyama, and R. Tomioka, "Multitask learning meets tensor factorization: task imputation via convex optimization," in *Advances in Neural Information Processing Systems*, 2014, pp. 2825–2833.

[8] M. Ashraphijuo and X. Wang, "Clustering a union of low-rank subspaces of different dimensions with missing data," *Pattern Recognition Letters*, vol. 120, pp. 31–35, 2019.

[9] E. Acar, D. M. Dunlavy, T. G. Kolda, and M. Mørup, "Scalable tensor factorizations for incomplete data," *Chemometrics and Intelligent Laboratory Systems*, vol. 106, no. 1, pp. 41–56, 2011.

[10] E. J. Candes and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, 2010.

[11] B. Huang, C. Mu, D. Goldfarb, and J. Wright, "Provable low-rank tensor recovery," *Optimization-Online*, vol. 4252, no. 2, pp. 455–500, 2014.

[12] H. Huang, Y. Liu, J. Liu, and C. Zhu, "Provable tensor ring completion," *Signal Processing*, vol. 171, p. 107486, 2020.

[13] P. Jain and S. Oh, "Provable tensor factorization with missing data," in *Advances in Neural Information Processing Systems*, 2014, pp. 1431–1439.

[14] Z. Zhang and S. Aeron, "Exact tensor completion using t-svd," *IEEE Transactions on Signal Processing*, vol. 65, no. 6, pp. 1511–1526, 2016.

[15] C. Prévost, K. Usevich, P. Comon, and D. Brie, "Hyperspectral super-resolution with coupled tucker approximation: Recoverability and svd-based algorithms," *IEEE Transactions on Signal Processing*, vol. 68, pp. 931–946, 2020.

[16] Y. Xu, Z. Wu, J. Chanussot, P. Comon, and Z. Wei, "Nonlocal coupled tensor CP decomposition for hyperspectral and multispectral image fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 1, pp. 348–362, 2019.

[17] H. Lamba, V. Nagarajan, K. Shin, and N. Shajarisales, "Incorporating side information in tensor completion," in *Proceedings of the 25th International Conference Companion on World Wide Web*, 2016, pp. 65–66.

[18] M. Nimishakavi, B. Mishra, M. Gupta, and P. Talukdar, "Inductive framework for multi-aspect streaming tensor completion with side information," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018, pp. 307–316.

[19] T. Zhou, H. Qian, Z. Shen, C. Zhang, and C. Xu, "Tensor completion with side information: A riemannian manifold approach." in *IJCAI*, 2017, pp. 3539–3545.

[20] A. P. Singh and G. J. Gordon, "Relational learning via collective matrix factorization," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008, pp. 650–658.

[21] N. Rao, H.-F. Yu, P. K. Ravikumar, and I. S. Dhillon, "Collaborative filtering with graph information: Consistency and scalable methods," in *Advances in Neural Information Processing Systems*, 2015, pp. 2107–2115.

[22] S. Bahargam and E. E. Papalexakis, "A constrained coupled matrix-tensor factorization for learning time-evolving and emerging topics," *arXiv preprint arXiv:1807.00122*, 2018.

[23] E. Acar, T. G. Kolda, and D. M. Dunlavy, "All-at-once optimization for coupled matrix and tensor factorizations," *arXiv preprint arXiv:1105.3422*, 2011.

[24] A. S. Zamzam, V. N. Ioannidis, and N. D. Sidiropoulos, "Coupled graph tensor factorization," in *2016 50th Asilomar Conference on Signals, Systems and Computers*. IEEE, 2016, pp. 1755–1759.

[25] E. Acar, M. Nilsson, and M. Saunders, "A flexible modeling framework for coupled matrix and tensor factorizations," in *2014 22nd European Signal Processing Conference (EUSIPCO)*. IEEE, 2014, pp. 111–115.

[26] R. C. Farias, J. E. Cohen, and P. Comon, "Exploring multimodal data fusion through joint decompositions with flexible couplings," *IEEE Transactions on Signal Processing*, vol. 64, no. 18, pp. 4830–4844, 2016.

[27] K. Wimalawarne, M. Yamada, and H. Mamitsuka, "Convex coupled matrix and tensor completion," *Neural computation*, vol. 30, no. 11, pp. 3095–3127, 2018.

[28] ——, "Scaled coupled norms and coupled higher-order tensor completion," *Neural Computation*, vol. 32, no. 2, pp. 447–484, 2020.

[29] K. Wimalawarne and H. Mamitsuka, "Efficient convex completion of coupled tensors using coupled nuclear norms," in *Advances in Neural Information Processing Systems*, 2018, pp. 6902–6910.

[30] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM review*, vol. 51, no. 3, pp. 455–500, 2009.

[31] P. Comon, "Tensors: a brief introduction," *IEEE Sig. Proc. Magazine*, vol. 31, no. 3, pp. 44–53, May 2014.

[32] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 208–220, 2012.

[33] M. Xu, R. Jin, and Z.-H. Zhou, "Speedup matrix completion with side information: Application to multi-label learning," in *Advances in Neural Information Processing Systems*, 2013, pp. 2301–2309.

[34] K.-Y. Chiang, C.-J. Hsieh, and I. S. Dhillon, "Matrix completion with noisy side information," in *Advances in Neural Information Processing Systems*, 2015, pp. 3447–3455.

[35] J. Lu, G. Liang, J. Sun, and J. Bi, "A sparse interactive model for matrix completion with side information," in *Advances in Neural Information Processing Systems*, 2016, pp. 4071–4079.

[36] P. Zhao, Y. Jiang, and Z.-H. Zhou, "Multi-view matrix completion for clustering with side information," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2017, pp. 403–415.

[37] V. N. Ioannidis, A. S. Zamzam, G. B. Giannakis, and N. D. Sidiropoulos, "Coupled graph and tensor factorization for recommender systems and community detection," *IEEE Transactions on Knowledge and Data Engineering*, 2019.

[38] F. M. Almutairi, N. D. Sidiropoulos, and G. Karypis, "Context-aware recommendation-based learning analytics using tensor and coupled matrix factorization," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 5, pp. 729–741, 2017.

[39] B. Ermiş, E. Acar, and A. T. Cemgil, "Link prediction in heterogeneous data via generalized coupled tensor factorization," *Data Mining and Knowledge Discovery*, vol. 29, no. 1, pp. 203–236, 2015.

[40] H. Chen and J. Li, "Collective tensor completion with multiple heterogeneous side information," in *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019, pp. 731–740.

[41] K. Y. Yılmaz, A. T. Cemgil, and U. Simsekli, "Generalised coupled tensor factorisation," in *Advances in Neural Information Processing Systems*, 2011, pp. 2151–2159.

[42] L. De Lathauwer and E. Kofidis, "Coupled matrix-tensor factorizations-sthe case of partially shared factors," in *2017 51st Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2017, pp. 711–715.

[43] L.-H. Lim and P. Comon, "Blind multilinear identification," *IEEE Trans. Inf. Theory*, vol. 60, no. 2, pp. 1260–1280, Feb. 2014, open access. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6671475

[44] S. Sahnoun and P. Comon, "Joint source estimation and localization," *IEEE Trans. Sig. Proc.*, vol. 63, no. 10, pp. 2485–2495, May 2015, hal-01005352.

[45] J. Yang, Y. Zhu, K. Li, J. Yang, and C. Hou, "Tensor completion from structurally-missing entries by low-tt-rankness and fiber-wise sparsity," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 6, pp. 1420–1434, 2018.

[46] W. Dong, F. Fu, G. Shi, X. Cao, J. Wu, G. Li, and X. Li, "Hyperspectral image super-resolution via non-negative structured sparse representation," *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 2337–2352, 2016.

[47] J. Ballani, L. Grasedyck, and M. Kluge, "Black box approximation of tensors in hierarchical tucker format," *Linear Algebra and its Applications*, vol. 438, no. 2, pp. 639–657, 2013.

[48] I. V. Oseledets, "Tensor-train decomposition," *SIAM Journal on Scientific Computing*, vol. 33, no. 5, pp. 2295–2317, 2011.

[49] L. Yuan, C. Li, J. Cao, and Q. Zhao, "Rank minimization on tensor ring: an efficient approach for tensor decomposition and completion," *Machine Learning*, vol. 109, no. 3, pp. 603–622, 2020.

[50] S. Boyd, N. Parikh, and E. Chu, *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.

[51] P. L. Bartlett and S. Mendelson, "Rademacher and gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, no. Nov, pp. 463–482, 2002.

[52] L. Chen, X. Jiang, X. Liu, and Z. Zhou, "Logarithmic norm regularized low-rank factorization for matrix and tensor completion," *IEEE Transactions on Image Processing*, vol. 30, pp. 3434–3449, 2021.

[53] R. Zdunek and T. Sadowski, "Image completion with approximate convex hull tensor decomposition," *Signal Processing: Image Communication*, vol. 95, p. 116276, 2021.

[54] Z. Long, C. Zhu, J. Liu, and Y. Liu, "Bayesian low rank tensor ring for image recovery," *IEEE Transactions on Image Processing*, vol. 30, pp. 3568–3580, 2021.

[55] A. Toet, M. J. de Jong, M. A. Hogervorst, and I. T. Hooge, "Perceptual evaluation of color transformed multispectral imagery," *Optical Engineering*, vol. 53, no. 4, p. 043101, 2014.

[56] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.

[57] F. Yasuma, T. Mitsunaga, D. Iso, and S. Nayar, "Generalized Assorted Pixel Camera: Post-Capture Control of Resolution, Dynamic Range and Spectrum," Tech. Rep., Nov 2008.

[58] C. H. Golub, F. T. Luk, and M. L. Overton, "A block Lanczos method for computing the singular values and corresponding singular vectors of a matrix," *ACM Trans. Math. Soft.*, pp. 149–169, Jun. 1981.

[59] P. Comon and G. H. Golub, "Tracking a few extreme singular values and vectors in signal processing," *Proceedings of the IEEE*, vol. 78, no. 8, pp. 1327–1343, 1990.

[60] T.-H. Oh, Y. Matsushita, Y.-W. Tai, and I. S. Kweon, "Fast randomized singular value thresholding for low-rank optimization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 2, pp. 376–391, 2017.

[61] X. Zhang, F. X. Yu, R. Guo, S. Kumar, S. Wang, and S.-F. Chang, "Fast orthogonal projection based on kronecker product," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2929–2937.

[62] S. M. Kakade, K. Sridharan, and A. Tewari, "On the complexity of linear prediction: Risk bounds, margin bounds, and regularization," 2008.