THE UNIVERSITY OF CHICAGO

STATISTICAL METHODS FOR HIGHER ORDER TENSOR CLUSTERING AND COMPLETION

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY
JIACHENG WANG

CHICAGO, ILLINOIS
JUNE 2022

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First of all, I would like to express my sincere gratitude to my advisor, Professor Dan Nicolae, for his continuous support, encouragement, invaluable expertise and generous guidance over the past five years. Professor Nicolae has been my role model by all means since I first came to University of Chicago. As an advisor, his vision and insightful feedback always enlighten me; beyond an advisor, he cares about every student and stands at my back during covid-19 tough times. It is my best luck and greatest honor to have worked with Professor Nicolae, without whom I would never be able to finish these exciting projects.

I would like to thank Professors Lek-Heng Lim and Matthew Stephens for granting me the honor to be member of my committee and for their time to attend my defense and enlightening advice and comments in my research.

My sincere thanks also goes to Professor Soudeep Deb and Marco Morales, who provided me an opportunity to join their team as intern and who gave access to abundant resources for internship. Without they precious support it would not be possible to conduct interesting internship projects.

Throughout my PhD study, I have received a great deal of support and assistance. My thanks to all of my cohorts and staff at Department of Statistics University of Chicago, for being part of this journey.

Finally, I want to send my special thanks to my parents, Limin Wang and Hong Xu. It is their tremendous love, unwavering support and belief in me in the past years that continuously empowered me to learn, work and explore till the completion of this study. This thesis is dedicated to them.

To my parents.

ABSTRACT

Statistical learning for tensors has gained increasing attention over the recent years. This thesis consists of methods for two major tensor-related statistical problems, tensor clustering and tensor completion.

Tensor multi-mode clustering is a generalization of biclustering for matrices, and successfully achieves better clustering performance by dealing with interactions among different modes in a more refined manner. Chapter 2 introduces a multi-mode tensor data clustering method which implements a fused version of alternating least squares algorithm (Fused-Orth-ALS) to perform tensor factorization and clustering simultaneously. The theory of Fused-Orth-ALS algorithm for statistical convergence rates of recovering and clustering is established, where the data form a noise contaminated tensor with latent low rank CANDECOMP/PARAFAC (CP) decomposition structure. In particular, we show that orthogonality adding to alternating least squares algorithm can provably recover the true latent low rank factorization structure when the data are from an asymmetric tensor with perturbation. Moreover, clustering consistency is well established for the Fused-Orth-ALS algorithm, ensuring its reliable clustering assignments.

Fibers in tensor missing not at random frequently appear in many applications, where typical tensor completion methodology that relies on entries being revealed uniform randomly may lead to underestimated mean squared error when evaluating specific imputation mechanisms. In Chapter 3, we propose to use propensity scores to remove selection bias for revealed fibers, which yields a reliable estimate for the underlying true tensor given noisy partially observed tensor fibers. Finite sample error bounds are established for the accuracy of the proposed max-norm rank constraint maximum likelihood approach estimate propensity score and underlying true tensor. The corresponding algorithms based on blockwise relaxation are constructed via low rank Tucker decomposition of the tensor.

Chapter 4 studies the problem of recovering low rank tensors with partially observed

entries using auxiliary information. Utilizing auxiliary information, or side information, such as the feature covariates for corresponding modes of tensor, is a plausible way to achieve compression and decomposition of high-dimensional low-rank tensors and thus reduce the sample complexity of tensor completion. We propose a new model for learning incomplete sparse tensor with auxiliary information (LISTAI) which employs interaction among different modes as well as sparsity structure. In particular, we prove the tensor can be fully recovered under perfect auxiliary information setting with a sample complexity lower bound. We also study the recovery performance when auxiliary information are corrupted by measuring the quality of noise perturbation. These results provide theoretical insight into the relationship among recovery performance, tensor dimension and usefulness of auxiliary information.

The efficacy and fast implementation of algorithms proposed for each topic are validated through both synthetic and real datasets, and the comparison results with alternative methods are included for each chapter.

CHAPTER 1

INTRODUCTION

1.1 Overview

Tensors, which are also known as multi-dimensional arrays, generalize matrices to more than two dimensions and appear frequently in applications and in modern machine learning research. From deep neural networks to videos and neuroimage data, the structures of higher order tensors are crucial for fully exploiting the underlying interactions within multi-dimensional data, and thus tensors are gaining accelerated attention in statistics and machine learning communities. Recent development of tensor-related statistical methods have been successfully implemented in a wide variety of scientific and business applications, including but not limited to

- *Neuroscience.* Research investigations in neuroscience are greatly facilitated by neuroimaging technologies, including anatomical magnetic resonance imaging (MRI), functional magnetic resonance imaging (fMRI), electroencephalography (EEG), diffusion tensor imaging, and positron emission tomography (PET), etc. These neuroimages often appear in tensor format and thus stimulate the exploration of statistical and computational methods that study ultrahigh dimensionality as well as complex structures. Many interesting papers related to the regression framework for tensor covariates [Zhou et al., 2013], brain functional connectivity (FC) evaluation [Zhu et al., 2017], analysis of EEG response to transcranial magnetic stimulation etc, shed light on using neuroimages for disease diagnosis and prediction, for characterization of drug effects and subjective human experience, and for understanding relationship between brain regions and cognitive outcomes.

- *Biomedical science.* Owing to quick progression of high-throughput sequencing technologies, multi-tissue experiments collect gene expression profiles from different par-

ticipants in a number of different tissues, where it naturally admits a tensor format data. For example, Hore et al. [2016] proposed a Bayesian method uncovering gene networks linked to genetic variation and successfully employed it on TwinsUK cohort with gene expression measured via RNA sequencing in adipose, LCLs and skin. Martino et al. [2021] revealed patterns driving differences in microbial composition across phenotype by analyzing tensor format data with three modes representing different microbial sequences, sampled host (or subject), and time or space.

- *Recommendation systems.* With rapid development of e-commerce, combining multiple sources of information provides more targeted potential items that users may be interested in; applications include selecting movies, restaurants, or online shopping. The accompanying challenges for these tensor format data have been alleviated using tensor factorization techniques, such as developing dynamic recommendation system techniques to capture time-dependency features [Zhang et al., 2021], generalizing the collaborative filtering method [Rendle and Schmidt-Thieme, 2010] and addressing the "cold-start" issue in the absence of information from new customers, new products or new contexts through sub-group information [Bi et al., 2018].

- *Deep learning.* The framework of deep neural networks is naturally a mapping between higher-order tensors. The typical structure of the most widely used convolutional neural networks is a large number of convolutional layers, followed by a few fully-connected layers. Lebedev et al. [2014], Kim et al. [2015] implemented compression on deep learning convolutional networks. Ye et al. [2020], Kossaifi et al. [2017] adopted different strategies on compressing fully-connected layers. Those trials successfully improved deep learning model robustness, from implicit (low-rank structure) or explicit (tensor dropout) regularization, leading to parsimonious models with a large reduction in the number of parameters, and achieved computational speed-ups by operating directly and efficiently on factorized tensors.

Motivated by the broad applications of tensors, we focus on two categories of higher order tensor statistical problems. We first study the theoretical and empirical performance of tensor multi-modes clustering. Inspired by biclustering theory for matrices, tensor multi-modes clustering aims to cluster multiple modes for tensor simultaneously, i.e., identify the subsets of each modes that are similar expressed while taking the underlying association with other modes into account. Multi-modes clustering serves as an important tool for revealing latent structures in the above mentioned applications. Brain imaging analysis can utilize the brain nodes clustering to understand how different brain regions are functionally separated and densely connected. Recommendation systems can provide more targeted recommendations, or more precise forecasting with the advantages of known users clustering structure that share similar preferences, or item clustering groups that have analogous characteristics.

The second category is tensor completion given partially observed noisy entries. High proportion of missing data is a common issue and poses huge challenges in tensor applications. For instance, in gene expression studies, multiple tissues are difficult to be accessed. Different challenges occur when working with images. Contaminated with some types of artifacts or noise, images may have poor quality or even missing parts. In some companies, due to the budgets on resources or latency requirements, only limited rating/purchasing records with respect to different items will be collected for users and no information could be attained for items that are not recommended. Thus, it is remarkable that, without delicate handling the missing data, not only the traditional tensor factorization technique will fail, but the final analysis results might lead to unacceptable consequences, such as neuroimage-based misdiagnosis, misinterpretations of gene expression traits and other cellular phenotypes, losing clients due to unsatisfactory recommendation experience, and possible financial loss.

Due to the high dimensional nature of tensors, solving these two categories of problems requires the principles of high dimensional statistics: although the model could potentially be of great complexity and the parameters might come from a very large space, the true model

should be parsimonious and the underlying parameters must belong to a small subset within the whole search space. For tensors, low rank decomposition structure has been proved to be a powerful dimension reduction approach. Considering the different nature of applications, different types of tensor low rank decomposition might be useful, which is part of the reason that there are different notions of tensor rank being studied. Two types of tensor low rank decomposition structures will be implemented in this thesis, Canonical Polyadic (CP) and Tucker.

- CP low rank decomposition stemmed independently from Carroll and Chang [1970] and Bro [1997] where the same method was separately named as Canonical decomposition (CANDECOMP) and Parallel Factors (PARAFAC). CP low rank decomposition is closely related to the concept of tensor rank where a tensor has rank 1 if it can be expressed as an outer product of D vectors. The CP low rank decomposition factorizes a tensor into a sum of component rank-one tensors, i.e., for an order D tensor $\mathcal{X} \in \mathbb{R}^{d_1 \times ... \times d_D}$, CP low rank decomposition provides an approximation of $\mathcal{X}$, i.e.,

$$\mathcal{X} \approx \sum_{r=1}^{R} \mathbf{U}_{1,(:,r)} \circ \mathbf{U}_{2,(:,r)} ... \circ \mathbf{U}_{D,(:,r)}$$

  where R is a positive integer, denoting the CP rank of $\mathcal{X}$ and $\mathbf{U}_i \in \mathbb{R}^{d_i \times R}$, $\mathbf{U}_{i,(:,r)}$ represents rth column of matrix $\mathbf{U}_i, \forall i \in \{1, 2, ..., D\}$.

- Tucker low rank decomposition was first introduced in Tucker [1963] and is a form of higher order principal component analysis (PCA). It decomposes a tensor into a core tensor multiplied by a matrix along each mode, i.e., for an order D tensor $\mathcal{X}$,

$$\mathcal{X} \approx \mathcal{C} \times_1 \mathbf{U}_1 ... \times_D \mathbf{U}_D = \sum_{i_1=1}^{r_1} ... \sum_{i_D=1}^{r_D} \mathcal{C}_{i_1,...,i_D} \mathbf{U}_{1,(:,i_1)} \circ ... \circ \mathbf{U}_{D,(:,i_D)}$$

  where $\mathcal{C} \in \mathbb{R}^{r_1 \times ... \times r_D}$ is called the core tensor, $\mathbf{U}_i \in \mathbb{R}^{d_i \times r_i}, \forall i \in \{1, 2, ..., D\}$ are

the factor matrices and can be thought of as the principal components for each mode. We use $\mathbf{U}_{i,(:,j)}$ to denote the jth column of matrix $\mathbf{U}_i$. The tuples $\mathbf{r} = (r_1, ..., r_D)$ is referred to as the Tucker rank.

Without assumption of low rankness, the total number of entries in an order D tensor is $\prod_{i=1}^{D} d_i$. With careful choice of R and $r_1, ..., r_D$, CP low rank and Tucker low rank successfully reduce that number to $\sum_{i=1}^{D} d_i R$ and $\prod_{i=1}^{D} r_i + \sum_{i=1}^{D} d_i r_i$ respectively.

Even though low rank decompositions provide a powerful dimension reduction tool for tensor data, they are insufficient for capturing the intrinsic properties in real data applications. For multi-modes clustering in tensors, implementing low rank decomposition structure solely will not allow us to reveal the underlying clustering structures along each mode. For tensor completion, the imputation task can sometime be more complicated than purely recovering the missing entries (missing completely at random). For example, in neuroimages, missing data or low image quality problem may be concentrated within a specific region. Obviously, it's more reasonable to assume a different missingness mechanism for the perturbed region from the whole image. In biomedical science, multi-tissue experiments can hardly collect gene expressions for participants in all interested tissues. Researchers might have difficulty in accessing brain tissues for obvious reasons, resulting in the whole fibers for brain tissues along the gene mode unrevealed. Furthermore, apart from the tensor dataset, some valuable auxiliary information might also be provided. For instance, while scanning for MRI/fMRI images or inviting donors to join multi-tissue experiments, donors' demographics and clinical phenotypes might also be collected, like sex, age, race, life habits, medication, and previous surgery records etc. Auxiliary information collection are also prevalent for recommendation systems. Often, user's age, preference and demographic information are collected when they create an account. As users interact more, social network analysis can be conducted via user actions. Analogously, feature information for items can be wrapped up as well according to their intrinsic characteristics, such as release date, director, starring, dis-

tribution company, awards nomination records in a movie recommendation system. Ignoring the valuable auxiliary information could lead to a huge loss for obtaining more reliable and accurate imputation for missing parts in tensor. Due to the inadequacy of low rank decomposition, more sophisticated structures will be added to tackle those specific problems and the detailed methodology, theoretical guarantees and empirical performance will be provided in Chapters 2, 3, and 4.

Chapter 2 addresses the problem of tensor multi-modes clustering. First, a CP low rank decomposition with regularized row-pairwise difference model is proposed to reveal the underlying clustering structure along each mode. The clustering structure is determined by the similarity among rows in each decomposed factor matrices; under the assumption of Gaussian distributed noise, this model can be proved as a general case of tensor Gaussian mixture model. Second, the corresponding fused orthogonal alternating least square (Fused-Orth-ALS) algorithm is proposed and its statistical convergence rate of recovering and clustering are established. Third, we show that this method enjoys more generality compared to previous results. Compared to dynamic clustering which requires that the time mode clustering structure varies smoothly, our method allows more flexible clustering structure including smooth change or block structures after randomized permutation. We also prove that the orthogonality added to algorithm can avoid convergence to poor local optima for an asymmetric tensor. In addition, multiple experiments on synthetic and real datasets results are provided to show the performance of our method in different scenarios.

Chapters 3 and 4 focus on the second category of problems, tensor completion. Although tensor completion has been investigated recently, existing methods, relying on entries that are missing uniformly, may fail when considering special missing data structures in tensors. In Chapter 3, the major goal is to impute fibers that are missing not at random. Fibers are the higher-order analogue of matrix rows and columns, and are defined by fixing every index but one. Missing fibers are common in tensor datasets and we narrow down our investigation

to fibers missing not at random which require a special solution. On account of selection bias introduced by missingness not at random, a Tucker low rank model with fiber-based propensity score is constructed and the theoretical guarantees for completion performance are provided under two different situations, an experimental setting where researchers can manipulate the probability of fibers revealed (i.e., a propensity score is given), and an observational setting where unobservable factors affect the probability of fiber revealed, resulting in unknown propensity scores. A lower bound sample complexity is also derived, which provides guidelines on the minimum number of observations required for the tensor to be fully recovered. Moreover, some numerical experimental results are provided for the study of the consistency of the theoretical convergence in terms of dimension, rank, as well as perturbation level and the efficiency and efficacy of algorithm on different types of tensor.

Chapter 4 describes our work on tensor completion with auxiliary information. The proposed model is named learning incomplete sparse tensor with auxiliary information (LISTAI). LISTAI employs interaction among different modes, incorporating the auxiliary information into the latent structure constitution of tensor. Sparsity structure is implemented on account of the fact that not all the interactions among different modes facilitate the underlying tensor recovery. We studied the model performance under both perfect auxiliary information setting and corrupted setting. In addition to the theoretical guarantees, an efficient nested double ADMM algorithm is developed with convergence rate constructed. In addition, we also validate the outstanding performance of this algorithm compared to other alternative methods on both simulation experiments and real dataset.

For each chapter, we developed R functions that implement the proposed algorithms, which are publicly available through:

Chapter 2: `https://github.com/Jiacheng-Wang/Fused-Orth-ALS_clustering`

Chapter 3: `https://github.com/Jiacheng-Wang/TensorCompletionFiberMNAR`

Chapter 4: `https://github.com/Jiacheng-Wang/LISTAI`

1.2 Notation and tensor algebra

In this section, we introduce some notations and classical tensor algebra. Throughout the thesis, calligraphic fonts are employed to represent tensors, e.g., $\mathcal{X}, \mathcal{Y} \in \bigotimes_{i=1}^{D} \mathbb{R}^{d_i} := \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_D} (D > 2)$. The order, D, is the number of ways or modes of a tensor. For simplicity, we denote the component $(i_1, ..., i_D)$ of order D tensor $\mathcal{X}$ by $\mathcal{X}_{i_1,...,i_D}$, where $i_k \in \{1, 2, ..., d_k\}, \forall k \in \{1, 2, ..., D\}$. A subarray of a tensor can be extracted by fixing a subsets of its indices. For matrices, it is obvious that that by fixing the first index to be i, we can extract the ith row of the matrix and by fixing the second index to be j, the jth column of the matrix can be obtained. By convention, we use a colon to indicate all elements of a mode. Thus, the ith row and jth column of a matrix $\mathbf{A}$ can be denoted by $\mathbf{A}_{i:}$ and $\mathbf{A}_{:j}$. Similar to matrix theory, fibers of tensor are defined to be subarrays obtained by fixing all but one of its indices and slices are two dimensional subarrays of a tensor obtained by fixing all but two indices. For example, an order three ($D = 3$) tensor $\mathcal{X} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ has three sets of fibers denoted as $\mathcal{X}_{i,j,:}, \mathcal{X}_{i,:,k}, \mathcal{X}_{:,j,k}$ and three sets of slices denoted as $\mathcal{X}_{i,:,:}, \mathcal{X}_{:,j,:}, \mathcal{X}_{:,:,k}$. Besides, matrices are represented by bold capital letters such as $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n_1 \times n_2}$ and vectors by boldface lower case letters $\mathbf{e}, \mathbf{g}$. Non-bold letters are implemented to denote scalars, e.g., D, d. In particular, we use blackboard capital letters $\mathbb{R}$ to represent the set of real numbers, $\mathbb{E}, \mathbb{P}$ to represent expectation and probability. For an event $\mathcal{Q}$, $\mathbf{1}_{\mathcal{Q}}$ is the indicator function for that event, i.e., $\mathbf{1}_{\mathcal{Q}} = 1$ if $\mathcal{Q}$ happens and 0 otherwise.

Different types of norm are extensively implemented in this thesis. For tensors, the inner product between two order D tensors $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_D}$ is defined as

$$\langle \mathcal{X}, \mathcal{Y} \rangle = \sum_{i_1=1}^{d_1} \sum_{i_2=1}^{d_2} ... \sum_{i_d=1}^{d_D} \mathcal{X}_{i_1,...i_D} \mathcal{Y}_{i_1,...i_D}$$

and the Frobenious norm of tensor is the natural generalization of the Frobenius norm of a matrix, $\|\mathcal{X}\|_F = \sqrt{\langle \mathcal{X}, \mathcal{X} \rangle}$. We use $\|\mathcal{X}\|_\infty$ to denote the elementwise ℓ_∞ norm. The tensor

spectral norm is defined as

$$\|\mathcal{X}\| = \max_{\mathbf{u}_j \in \mathbb{R}^{d_j} : \|\mathbf{u}_j\|=1} \langle \mathcal{X}, \mathbf{u}_1 \circ \mathbf{u}_2 \circ ... \circ \mathbf{u}_D \rangle$$

and $\circ$ denotes the outer product among vectors. Another norm widely implemented in chapter 4 is nuclear norm, which is defined in Friedland and Lim [2018],

$$\|\mathcal{X}\|_* = \inf \Big\{ \sum_{r=1}^{R} |\lambda_r| : \mathcal{X} = \sum_{r=1}^{R} \lambda_r \mathbf{U}_{1,(:,r)} \circ \mathbf{U}_{2,(:,r)} \circ ... \circ \mathbf{U}_{d,(:,r)}, \|\mathbf{U}_{i,(:,r)}\| = 1, \forall i \in \{1, 2, ..., D\} \Big\}$$

Moreover, let $\|\mathcal{X}\|_{\min} = \min_{i_1,...,i_D} |\mathcal{X}_{i_1,...,i_D}|$ denote the element-wise minimum norm of $\mathcal{X}$. For vector $\mathbf{r} = [r_1, ..., r_D]$, we use $r_{\max}$ to represent the maximum element in $\mathbf{r}$, i.e., $r_{\max} = \max_{i \in \{1,2,...,D\}} r_i$. For simplicity of notation, $\|\mathbf{v}\| = (\sum_i v_i^2)^{1/2}$ denotes the Euclidean ℓ_2 norm of a vector v, $\|\mathbf{A}\|$ denotes the spectral norm of a matrix $\mathbf{A}$, $\|\mathbf{A}\|_\infty$ denotes the ℓ_∞ element-wise norm of matrix $\mathbf{A}$ and $\|\mathbf{A}\|_1 = \sum_{i,j} |\mathbf{A}_{i,j}|$ denotes the ℓ_1 element-wise norm of matrix $\mathbf{A}$.

Tensor algebraic rules are crucial when performing theoretical analysis. For order D tensor $\mathcal{X}$ with dimension $d_1 \times ... \times d_D$, we frequently consider its unfolding mode i matricization, $\mathcal{X}_{(i)}$, which is a matrix with dimension $d_i \times \prod_{j \neq i} d_j$ whose columns are the mode i fibers of $\mathcal{X}$. We take the vectorization of $\mathcal{X}$, denoted $\text{vec}(\mathcal{X})$, to be the column-major vectorization of the mode-1 matricization of $\mathcal{X}$, namely $\text{vec}(\mathcal{X}) = \text{vec}(\mathcal{X}_{(1)}) \in \mathbb{R}^{\prod_{i=1}^{D} d_i}$. Furthermore, tensor marginal multiplication with matrices are extensively utilized in later analysis. Let $\mathbf{X}_1$ be a matrix of size $n_1 \times d_1$. The 1st mode (matrix) product of the tensor $\mathcal{X}$ with the matrix $\mathbf{X}_1$ yields a tensor of size $n_1 \times d_2 \times ... \times d_D$, which is

$$(\mathcal{X} \times_1 \mathbf{X}_1)_{i_1,...,i_D} = \sum_{i'_1=1}^{d_1} \mathbf{X}_{1,(i_1,i'_1)} \mathcal{X}_{i'_1,...,i_D}$$

where $\mathbf{X}_{1,(i_1,i'_1)}$ represents the (i_1, i'_1) entry in $\mathbf{X}_1$. In a similar way, we can define the

other mode i marginal product of tensor $\mathcal{X}$ with $\mathbf{X}_i \in \mathbb{R}^{n_i \times d_i}, \forall i \in \{2, 3, ..., D\}$ and by performing these tensor multiplication with matrices, we obtain an order D tensor of size $n_1 \times n_2 \times ... \times n_D$,

$$\mathcal{X} \times_1 \mathbf{X}_1 \times_2 \mathbf{X}_2 ... \times_D \mathbf{X}_D := [\mathcal{X}; \mathbf{X}_1, ..., \mathbf{X}_D] \in \mathbb{R}^{n_1 \times n_2 \times ... \times n_D}$$

It's easy to show the vectorization of tensor multiplication can be expressed as

$$\text{vec}([\mathcal{X}; \mathbf{X}_1, ..., \mathbf{X}_D]) = (\mathbf{X}_D \otimes ... \otimes \mathbf{X}_1)\text{vec}(\mathcal{X})$$

$\otimes$ represents Kronecker product among matrices. In addition, given matrices $\mathbf{A} \in \mathbb{R}^{d_1 \times K}$ and $\mathbf{B} \in \mathbb{R}^{d_2 \times K}$, their Khatri-Rao product, denoted by $\odot$, is defined as

$$\mathbf{A} \odot \mathbf{B} \in \mathbb{R}^{d_1 d_2 \times K} = [\mathbf{A}_1 \otimes \mathbf{B}_1 \;\; \mathbf{A}_2 \otimes \mathbf{B}_2 \;\; ... \;\; \mathbf{A}_K \otimes \mathbf{B}_K]$$

For two matrices $\mathbf{A} \in \mathbb{R}^{d_1 \times d_2}, \mathbf{B} \in \mathbb{R}^{d_1 \times d_2}$, the Hadamoard product is defined as the entry-wise multiplication of the matrices,

$$\mathbf{A} * \mathbf{B} \in \mathbb{R}^{d_1 \times d_2} = \begin{bmatrix} \mathbf{A}_{1,1}\mathbf{B}_{1,1} & \mathbf{A}_{1,2}\mathbf{B}_{1,2} & ... & \mathbf{A}_{1,d_2}\mathbf{B}_{1,d_2} \\ \mathbf{A}_{2,1}\mathbf{B}_{2,1} & \mathbf{A}_{2,2}\mathbf{B}_{2,2} & ... & \mathbf{A}_{2,d_2}\mathbf{B}_{2,d_2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_{d_1,1}\mathbf{B}_{d_1,1} & \mathbf{A}_{d_1,2}\mathbf{B}_{d_1,2} & ... & \mathbf{A}_{d_1,d_2}\mathbf{B}_{d_1,d_2} \end{bmatrix}$$

Finally, we may use $[d_i] := \{1, 2, ..., d_i\}$ denotes the whole index set. For two positive sequences $\{a_n\}, \{b_n\}$, $a_n \succ b_n$ means $\frac{b_n}{a_n} \to 0$, $a_n \lesssim$ means $a_n \leq C b_n$ for some constant $C > 0$ independent of n and $a_n \asymp b_n$ means $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

CHAPTER 2

FUSED ORTHOGONAL ALTERNATING LEAST SQUARES FOR TENSOR CLUSTERING

2.1 Introduction

Higher-order tensors, or multidimensional arrays, have been recently a popular topic in the fields of statistics and machine learning. This data format plays an essential role for a wide range of scientific and business applications. In biology, analyzing gene expression data collected for multiple time points or across different biological conditions using tensor related methods has been widely explored in Alter et al. [2000], Liu et al. [2003], Hore et al. [2016], Liu et al. [2017], Zhang et al. [2019]. There is also a growing body of literature that recognizes the frequency appearance of tensor data in social network and commercial analysis. Bruce et al. [2017] analyzed the user behavior pattern through consumer engagement on advertisements over time. Nickel et al. [2011] performed relational learning for different network data based on a tensor factorization framework. Moreover, recommender systems, which aim to predict the users' preference over a large number of items, have been studied by Bi et al. [2017, 2018], Zhang et al. [2020].

As higher order tensors have been extensively used as a framework for storing and organizing massive data, the associated need to develop methods for tensor multi-modes clustering has sharply increased. Clustering, as one of the most studied machine learning problems, has been investigated through thorough and extensive research in the single-mode case. Different from traditional clustering (K-means, hierarchical clustering or principal component analysis) along each single mode of the high order tensor while concatenating samples into a large matrix, multi-modes clustering has been proved useful in revealing latent structures in high dimensional expressions. Several methods have been established for two dimensional biclustering for matrix columns and rows simultaneously in Chi and Lange [2015], Chen

et al. [2015], Chi et al. [2017], Tan and Witten [2015]. However, those methods cannot be directly applied to tensors due to the complex higher order generalizations of the matrix singular value decomposition (SVD) and principal component analysis. Motivated by the gap between low and high dimensional co-clustering, we investigate here the multi-modes clustering problem for high order tensors. We propose the Fused Orthogonal Alternating Least Squares (Fused-Orth-ALS) algorithm to unravel co-clustering structures for high order tensors.

Multiple papers are closely related to, but also clearly distinct from, our work in tensor multi-modes clustering. Foundational development of a Tucker model based tensor decomposition clustering appears in Chi et al. [2018] which tackles the convex co-clustering challenge under tensor block structures; this is further developed in Wang and Zeng [2019] which proposed a tensor block model that automatically implies low-rankness and performed a unified least-square estimation procedure for identifying the block structure. Both methods implemented one type of tensor decomposition structure, namely, the Tucker decomposition model, which decomposes the tensor into a core tensor multiplied by orthogonal matrices in each mode. A popular alternative to the Tucker model utilizes another mainstream of tensor decomposition structure, CANDECOMP/PARAFAC (CP) decomposition, by decomposing a tensor into a sum of rank-1 tensors. This approach handles heterogeneity in each mode and learns the clustering patterns across different modes of data in a more independent way, thus providing flexibility for clustering a certain mode of the tensor without being affected by correlation with other modes. We should note that single-mode clustering still accounts for a large proportion of applications, those where people would like to analyze the clustering pattern for a certain mode but still take the interactions among different modes into consideration. Our methodology adopts CP decomposition as part of the model structure to be applicable under different research requirements, either single-mode clustering or multi-modes clustering. Moreover, our method is similar to a recent series of papers, Wang

and Zeng [2019] and Sun and Li [2019] with respect to underlying CP tensor decomposition structure. In contrast, the estimation algorithm in both papers use the framework of tensor power method which was first proposed in Allen [2012]. Tensor power method decomposes the tensor via successive rank-1 approximations. After getting a rank-1 approximation, one takes the residual tensor as new input and repeats the algorithm to find the next component. Different from tensor power method, our algorithm utilizes Alternating Least Square (ALS) algorithm which is by far the most widely employed decomposition algorithm [Kolda and Bader, 2009]. ALS algorithm estimates the rank-k decomposition specified by factor matrices simultaneously, instead of column-by-column recovery for each factor matrices as the tensor power method. The ALS algorithm has been proved to be robust and computational efficient in Huang et al. [2014], Kang et al. [2012], Kossaifi et al. [2019], Bader et al. [2012]. In particular, compared to the method in Wang et al. [2019] which performs clustering over estimated factor matrices, our method encourages smoothness to the factor matrices by imposing a generalized LASSO penalty, resulting in apparent clustering structure in factor matrices since data from same cluster will possess the same value by choosing appropriate tuning parameters. Moreover, one critical assumption in Sun and Li [2019] is that samples from the same cluster are ordered consecutively, that is, the true cluster assignment looks like

$$(1, ..., 1, 2, ..., 2, ..., S, ..., S) \tag{2.1}$$

where S is the number of clusters. Under this assumption, the simple fused LASSO penalty imposed in dynamic tensor clustering algorithm still works. Generally, this assumption does not hold in real applications since data from the same cluster may appear randomly, not adjacent in the index. Thus, our approach employs a more general LASSO regularization which is based on a graphical distance intuition learned from initial observations. As we will see shortly, this generalized fusion penalty achieves better clustering performance when clus-

tering assignments are randomly distributed. Even though our Fused-Orth-ALS algorithm is motivated by Sharan and Valiant [2017], we make several advances as follows. First, Orth-ALS algorithm in Sharan and Valiant [2017] only considers tensor decomposition performance under noiseless situation. We successfully provide statistical guarantees for Fused-Orth-ALS algorithm under any error tensor with a mild constraint on error tensor spectral norm. Second, theoretical properties for Orth-ALS are only valid for symmetric tensors which assume that all the factor matrices decomposed from tensor data are the same. Fused-Orth-ALS algorithm can achieve both recovery and clustering consistency for any asymmetric tensor. In short, we prove statistical properties and convergence results for Fused-Orth-ALS algorithm under more general settings.

In summary, we propose the Fused Orthogonal Alternating Least Squares (Fused-Orth-ALS) algorithm to unravel co-clustering structures for high order tensors. This project was designed to systematically investigate and achieve the following:

1. Under mild assumptions for tensor decomposition structure, our approach encourages smoothness in the decomposed components in a more general graphical way, leading to clustering performance with higher accuracy.

2. Orthogonality is proposed in the components decomposed from our algorithm, avoiding the problem of local-minima convergence of classical ALS tensor decomposition algorithm, preventing multiple recoveries of the same factors, and achieving faster convergence rates compared to other popular tensor decomposition methods.

3. We provide theoretical guarantees for the recovery and clustering consistency with even a single tensor sample, which is difficult to achieve in vector or matrix clustering. We establish a high probability error bound for the estimators and clustering assignments as the dimension of tensor data increases.

We apply our method to both simulation and real data applications. The results demonstrate

the robustness of our approach under mild model misspecification and assumption violations, efficiency and effectiveness under high dimensional settings to extract underlying clustering structure.

The rest of the chapter is organized as follows. Section 2.2 introduces the precise model formulation of the problem. In Section 2.3, we introduce the proposed Fused-Orth-ALS algorithm. Section 2.4 establishes the recovery error bound and clustering consistency result for Fused-Orth-ALS algorithm and different simulation experiment results are provided in Section 2.5. Section 2.6 discusses the results of applying Fused-Orth-ALS algorithm to real datasets from Human Connectome Project (HCP) which captures the brain nodes connections, and a dataset that consists of political relationships among different countries. All technical proofs, supporting lemmas as well as additional experiments are provided in Appendix.

For notational convenience, we define the multilinear combination of tensor entries in this chapter as

$$\mathcal{X}(\mathbf{u}_1, ..., \mathbf{u}_D) := \mathcal{X} \times_1 \mathbf{u}_1 ... \times_D \mathbf{u}_D$$

2.2 Model

Given an order D tensor $\mathcal{Y} \in \mathbb{R}^{d_1 \times ... \times d_D}$, our goal is to uncover the underlying clustering structures for each mode. For convenience, we limit our current tensor observations to order three tensors with $D = 3$; it is straightforward to derive the analogous results for higher order tensors with $D > 3$. We would like to perform clustering over factor matrices recovered from structured tensor CP decomposition for each mode.

Given order three tensor observations $\mathcal{Y}$, we assume it comes from the model where true

underlying tensor $\mathcal{Y}^*$ is perturbed by noise tensor $\mathcal{E}$,i.e.

$$\mathcal{Y} \in \mathbb{R}^{d_1 \times d_2 \times d_3} := \mathcal{Y}^* + \mathcal{E} \tag{2.2}$$

where $\mathcal{Y}^*$ is a tensor with a rank K CP decomposition structure,

$$\mathcal{Y}^* = \sum_{i \in [K]} w_i \mathbf{A}_{:i} \otimes \mathbf{B}_{:i} \otimes \mathbf{C}_{:i} \tag{2.3}$$

$\mathbf{A} \in \mathbb{R}^{d_1 \times K}, \mathbf{B} \in \mathbb{R}^{d_2 \times K}, \mathbf{C} \in \mathbb{R}^{d_3 \times K}$ are the factor matrices with unit 1 column-wise ℓ_2 norm, i.e., $\|\mathbf{A}_{:i}\|_2 = \|\mathbf{B}_{:i}\|_2 = \|\mathbf{C}_{:i}\| = 1, \forall i \in [K]$. $\mathbf{w} = [w_1, ..., w_K] \in \mathbb{R}^K$ refers to the weight of the factor matrices.

The clustering structure for each mode will depend on the factor matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$. Suppose we have s_1, s_2, s_3 clusters over three modes respectively and thus we can rewrite ith row of factor matrix $\mathbf{A}$ as

$$\mathbf{A}_{i:} = \sum_{j=1}^{s_1} {\boldsymbol{\mu}_{1,j}}^{*\top} \mathbf{1}_{i \in \mathfrak{A}_j^*} \tag{2.4}$$

where $\mathfrak{A}_j^*$ represents the index set, including rows belong to cluster j over first mode, and $\boldsymbol{\mu}_{1,j}^* \in \mathbb{R}^K$ is the mean value for rows in cluster j. Thus, we would expect there exists $s \in [s_1]$ such that $i \in \mathfrak{A}_s^*$, $\mathbf{A}_{i:} = \boldsymbol{\mu}_{1,s}$ and the clustering assignment for ith row of $\mathbf{A}$ is s, e.g., we would like to assign ith element in the first mode of $\mathcal{Y}^*$ into cluster s. Analogously, we can define $\boldsymbol{\mu}_{2,j}^*, \mathfrak{B}_j^*, \boldsymbol{\mu}_{3,j}^*, \mathfrak{C}_j^*$ and write ith row for $\mathbf{B}, \mathbf{C}$ respectively as

$$\mathbf{B}_{i:} = \sum_{j=1}^{s_2} \boldsymbol{\mu}_{2,j}^{*\top} \mathbf{1}_{i \in \mathfrak{B}_j^*}$$

$$\mathbf{C}_{i:} = \sum_{j=1}^{s_3} \boldsymbol{\mu}_{3,j}^{*\top} \mathbf{1}_{i \in \mathfrak{C}_j^*}$$

Similar clustering labels over second and third mode can be assigned. Given the underlying tensor decomposition structure as mentioned in (2.3), we can apply a clustering algorithm to each factor matrices to get the clustering label of each mode.

Finding the partitions $\mathfrak{A}_i^*, \mathfrak{B}_j^*, \mathfrak{C}_k^*, \forall i \in [s_1], j \in [s_2], k \in [s_3]$ is a combinatorical hard problem. Moreover, many combinatorical hard problems have been showed, in recent years, to be computational intractable. Inspired by penalized regularization term proposed in Wang et al. [2016b] and Chi et al. [2018], we impose the following full-pair fusion regularization over each factor matrices as,

$$\begin{aligned}
\mathbf{A} \in \mathcal{F}(d_1, f_1) &:= \{\mathbf{A} \in \mathbb{R}^{d_1 \times K} | \| \ {}^1\mathbf{\Delta}\mathbf{A}\|_1 \leq f_1\} \\
\mathbf{B} \in \mathcal{F}(d_2, f_2) &:= \{\mathbf{B} \in \mathbb{R}^{d_2 \times K} | \| \ {}^2\mathbf{\Delta}\mathbf{B}\|_1 \leq f_2\} \\
\mathbf{C} \in \mathcal{F}(d_3, f_3) &:= \{\mathbf{C} \in \mathbb{R}^{d_3 \times K} | \| \ {}^3\mathbf{\Delta}\mathbf{C}\|_1 \leq f_3\}
\end{aligned}$$

We use the fusion structure for $\mathbf{A}$ as an example to give a full explanation; fusion structures for $\mathbf{B}$ and $\mathbf{C}$ can be derived analogously. ${}^1\mathbf{\Delta} \in \mathbb{R}^{\binom{d_1}{2} \times d_1}$ works as pairwise difference operator over rows of $\mathbf{A}$ to yield a weighted penalty on local differences:

$$\| \ {}^1\mathbf{\Delta}\mathbf{A}\|_1 = \sum_{\mathbf{i} \in \mathcal{S}} \gamma^1_{i_1,i_2} \|\mathbf{A}_{i_1:} - \mathbf{A}_{i_2:}\|_1$$

$\mathcal{S} := \{\mathbf{i} = [i_1, i_2] | \ i_1 < i_2, \forall i_1, i_2 \in [d_1]\}$ represents all pairwise row indices and the parameters $\gamma^1_{i_1,i_2}$ are non-negative weights for pairwise difference over first mode. We will explain how to choose those weights in section 2.3.2 in detail. $\gamma^2_{i_1,i_2}, \gamma^3_{i_1,i_2}$ can be defined in a similar way for ${}^2\mathbf{\Delta}$, ${}^3\mathbf{\Delta}$. For simplicity, we assume the fusion parameter over three modes are the same, e.g., $f_1 = f_2 = f_3$ and use the same strategy as in Tibshirani et al. [2005]. In summary, we propose the following penalized constrained optimization as an approach to reveal

the latent clustering structure,

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{w}} \|\mathcal{Y} - \sum_{i\in[K]} w_i \mathbf{A}_{:i} \otimes \mathbf{B}_{:i} \otimes \mathbf{C}_{:i}\|_2 + \lambda\Big[\| {}^1\mathbf{\Delta}\mathbf{A}\|_1 + \| {}^2\mathbf{\Delta}\mathbf{B}\|_1 + \| {}^3\mathbf{\Delta}\mathbf{C}\|_1\Big]$$
$$\text{s.t. } \|\mathbf{A}_{:i}\| = \|\mathbf{B}_{:i}\| = \|\mathbf{C}_{:i}\| = 1, \forall i \in [K] \tag{2.5}$$

Taking a closer look at the regularization term, we find that as λ increases, $\mathbf{A}_{i:}$ will shrink towards each other which means the pairwise differences of the rows in $\mathbf{A}$ will become increasingly sparse. Sparsity in pairwise differences leads to the partitioning assignment $\mathfrak{A}_j^*, \forall j \in [s_1]$. Similar behavior holds for $\mathbf{B}$ and $\mathbf{C}$. For simplicity, we choose the tuning parameter λ to be the same over the three modes. In Section 2.3, we show in detail how to normalize the three regularization terms to ensure that the penalty term over three modes are on the same scale. We would like to draw the attention to the difference between the regularization term in (2.5) and that in Chi et al. [2018]. We encourage a l_1 norm penalty which will derive exact 0 difference between pairwise differences due to the feature of variable selection for lasso ℓ_1 regularization. The Frobenious norm penalty over pairwise differences of tensor slices are employed in Chi et al. [2018], which can only shrink the pairwise difference to 0 gradually. As we have mentioned, Sun and Li [2019] used fused LASSO penalty to achieve mean value estimation and cluster label assignment simultaneously. However, this simple fusion structure only works under the assumption that samples from same cluster have consecutive indices as (2.1) describes. Obviously, the fused LASSO penalty used in Sun and Li [2019] is a simple case of our weighted pairwise difference operator ${}^i\mathbf{\Delta}, \forall i \in \{1, 2, 3\}$ and simulation experiments in section 2.5 show that our method outperforms their dynamic clustering method when the cluster labels are randomly assigned.

2.3 Methods

2.3.1 Optimization Algorithm

We first introduce the classical ALS algorithm which uses alternating optimization strategy to obtain the factor matrices. As stated in Comon et al. [2009], Kolda and Bader [2009], ALS can get stuck in local optima easily, especially when the weights of the factors are non-uniform. Anandkumar et al. [2014] provided local convergence rate for rank-1 ALS updates by choosing uniformly initialization at random and global convergence through SVD based expensive initialization scheme.

Algorithm 1: Alternating Least Square Algorithm

Input: Tensor $\mathcal{Y} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$, number of the iterations N and tensor rank K
Output: $\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}$
Initialize $\hat{\mathbf{A}} \in \mathbb{R}^{d_1 \times K}, \hat{\mathbf{B}} \in \mathbb{R}^{d_2 \times K}, \hat{\mathbf{C}} \in \mathbb{R}^{d_3 \times K}$;
for *t = 1:N* **do**
  $\hat{\mathbf{A}} = \mathcal{Y}_{(1)}(\hat{\mathbf{C}} \odot \hat{\mathbf{B}})(\hat{\mathbf{C}}^\top \hat{\mathbf{C}} \times \hat{\mathbf{B}}^\top \hat{\mathbf{B}})^\dagger$;
  $\hat{\mathbf{B}} = \mathcal{Y}_{(2)}(\hat{\mathbf{C}} \odot \hat{\mathbf{A}})(\hat{\mathbf{C}}^\top \hat{\mathbf{C}} \times \hat{\mathbf{A}}^\top \hat{\mathbf{A}})^\dagger$;
  $\hat{\mathbf{C}} = \mathcal{Y}_{(3)}(\hat{\mathbf{B}} \odot \hat{\mathbf{A}})(\hat{\mathbf{B}}^\top \hat{\mathbf{B}} \times \hat{\mathbf{A}}^\top \hat{\mathbf{A}})^\dagger$
end

Sharan and Valiant [2017] proposed the Orthogonalized ALS (Orth-ALS) algorithm which solved the problem of converging to poor local optima for classical ALS algorithm. But Orth-ALS cannot be directly applied to multi-mdoes clustering for high order tensor. Inspired by the quick convergence and outstanding tensor decomposition performance of Orth-ALS, we proposed the following Fused-Orth-ALS algorithm which is described in Algorithm 2:

Compared to Algorithm 1, two extra steps are added that are designed for multi-modes clustering. First, the orthogonalization step over each factor matrix is performed before each set of ALS iterates which forces the estimate for columns in factor matrices converge to different factors. Second, a 'Fuse' operator is employed on each column of ALS estimates.

Algorithm 2: Fused Orthogonal Alternating Least Square (Fused-Orth-ALS)

Input: Tensor $\mathcal{Y} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$, number of the iterations N and tensor rank K
Output: $\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}, \hat{\mathbf{w}}$
Initialize each column of $\hat{\mathbf{A}} \in \mathbb{R}^{d_1 \times K}, \hat{\mathbf{B}} \in \mathbb{R}^{d_2 \times K}, \hat{\mathbf{C}} \in \mathbb{R}^{d_3 \times K}$ uniformly from the unit sphere;
for *t = 1:N* **do**
  Find the QR decomposition of $\hat{\mathbf{A}}$ and set $\hat{\mathbf{A}} = Q$. Orthogonalize $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$ analogously;
  $\mathbf{X} := [\mathbf{X}_{:i}]_{i=1}^K = \mathcal{Y}_{(1)}(\hat{\mathbf{C}} \odot \hat{\mathbf{B}}), \tilde{\mathbf{X}} := [\tilde{\mathbf{X}}_{:i}]_{i=1}^K, \tilde{\mathbf{X}}_{:i} = \text{Fuse}(\mathbf{X}_{:i},\ {}^1\mathbf{\Delta}, \lambda)$;
  $\mathbf{Y} := [\mathbf{Y}_{:i}]_{i=1}^K = \mathcal{Y}_{(2)}(\hat{\mathbf{C}} \odot \hat{\mathbf{A}}), \tilde{\mathbf{Y}} := [\tilde{\mathbf{Y}}_{:i}]_{i=1}^K, \tilde{\mathbf{Y}}_{:i} = \text{Fuse}(\mathbf{Y}_{:i},\ {}^2\mathbf{\Delta}, \lambda)$;
  $\mathbf{Z} := [\mathbf{Z}_{:i}]_{i=1}^K = \mathcal{Y}_{(3)}(\hat{\mathbf{B}} \odot \hat{\mathbf{A}}), \tilde{\mathbf{Z}} := [\tilde{\mathbf{Z}}_{:i}]_{i=1}^K, \tilde{\mathbf{Z}}_{:i} = \text{Fuse}(\mathbf{Z}_{:i},\ {}^3\mathbf{\Delta}, \lambda)$;
  Normalize $\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}, \tilde{\mathbf{Z}}$ and store the results in $\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}$;
end
Estimate weights $w_i = \mathcal{Y}(\hat{\mathbf{A}}_{:i}, \hat{\mathbf{B}}_{:i}, \hat{\mathbf{C}}_{:i}), \forall i \in [K]$

We define the fuse operator similar to Tibshirani et al. [2005] as,

$$\text{Fuse}(\mathbf{v}, \mathbf{\Delta}, \lambda) = \arg\min_{\mathbf{u} \in \mathbb{R}^d} \{\sum_{j=1}^{d} (v_j - u_j)^2 + \lambda \|\mathbf{\Delta}\mathbf{v}\|_1\} \tag{2.6}$$

Arnold and Tibshirani [2016], Zhu [2017] provided efficient algorithms to solve this generalized LASSO problem. To see the connection between our objective function (2.5) and the Fuse operator (2.6), the regularization term for each factor matrix in (2.5) can be viewed as the sum of Fuse operator regularization imposed on each column of factor matrix, e.g., $\| {}^1\mathbf{\Delta}\mathbf{A}\|_1 = \sum_{i=1}^K \| {}^1\mathbf{\Delta}\mathbf{A}_{:i}\|_1$ since we use elementwise ℓ_1 norm $\|\cdot\|_1$ for matrix. Thus, our original objective function (2.5) can be rewritten as

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{w}} \|\mathcal{Y} - \sum_{i \in [K]} w_i \mathbf{A}_{:i} \otimes \mathbf{B}_{:i} \otimes \mathbf{C}_{:i}\|_2 + \lambda \Big[\sum_{i=1}^{K} (\| {}^1\mathbf{\Delta}\mathbf{A}_{:i}\|_1 + \| {}^2\mathbf{\Delta}\mathbf{B}_{:i}\|_1 + \| {}^3\mathbf{\Delta}\mathbf{C}_{:i}\|_1) \Big]$$

$$\text{s.t. } \|\mathbf{A}_{:i}\| = \|\mathbf{B}_{:i}\| = \|\mathbf{C}_{:i}\| = 1, \forall i \in [K]$$

At each step we fix two factor matrices, for example, $\mathbf{B}, \mathbf{C}$, and try to estimate $\mathbf{A}$ by

solving the objective constrained function as $\min_{\mathbf{A}} \|\mathcal{Y} - \sum_{i\in[K]} w_i \mathbf{A}_{:i} \otimes \mathbf{B}_{:i} \otimes \mathbf{C}_{:i}\|_2 + \lambda\Big[\sum_{i=1}^{K} \| \ {}^1\mathbf{\Delta}\mathbf{A}_{:i}\|_1\Big]$. A three-step procedure is incorporated into Fused-Orth-ALS Algorithm: first update using ALS iterates, then impose Fuse operator on each column, and finally perform the normalization.

From the computational complexity perspective, orthogonalization in Algorithm 2 takes $O(K^2(d_1+d_2+d_3))$ number of operations. For the factor matrices updating steps, the total ALS updates take $O(Kd_1d_2d_3)$ and the Fuse operator steps take $O(K(d_1^3+d_2^3+d_3^3))$. Generally speaking, the total computational complexity of Fused-Orth-ALS algorithm performing on order D tensor $\mathcal{Y} \in \mathbb{R}^{d_1\times\ldots\times d_D}$ is $O(KN(\max(K\sum_{j=1}^{D} d_j, D\prod_{j=1}^{D} d_j, \sum_{j=1}^{D} d_j^3)))$. When the dimension $d_j, \forall j \in [D]$ along each mode are of similar order d and tensor rank $K = O(Dd^{D-1})$, the total complexity can be simplified as $O(KD\prod_{j=1}^{D} d_j)$ which is the same order as classical ALS algorithm, meaning that the extra orthogonalization and fuse steps do not increase the computational complexity of the tensor decomposition algorithm. Compared to the dynamic tensor clustering algorithm in Sun and Li [2019], our algorithm has the same order of computational complexity but orthogonality will lead to significant speedups in terms of number of iterations required for classical ALS and tensor power methods. Comparison with regard to iterations required for convergence will be showed in section 2.5.

After obtaining estimates for factor matrices $\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}$, we can perform clustering algorithm such as k-means or agglomerative hierarchical clustering over factor matrices to get the clustering assignment estimates $\hat{\mathfrak{A}}_i, \hat{\mathfrak{B}}_j, \hat{\mathfrak{C}}_k, \forall i \in [s_1], j \in [s_2], k \in [s_3]$.

2.3.2 Regularization weight

We only introduce our choice for the non-negative weights $\gamma^1_{i_1,i_2}$ for the first mode pairwise difference operator ${}^1\mathbf{\Delta}$ since $\gamma^2_{i_1,i_2}, \gamma^3_{i_1,i_2}$ follow the same analysis strategy. The non-negative weights over the first mode $\gamma^1_{i_1,i_2}$ characterize the shrinkage of the difference between $\mathbf{A}_{i_1:}$

and $\mathbf{A}_{i_2:}$. Thus, $\gamma^1_{i_1,i_2}$ quantifies the similarity between i_1th row and i_2th row of $\mathbf{A}$ and large value of $\gamma^1_{i_1,i_2}$ indicates that there is a large probability that i_1th and i_2th row should be assigned to the same cluster. This pairwise similarity motivates the graphical view of clustering. We can treat each row of $\mathbf{A}$ as a node in an undirected graph and we connect any two nodes i_1, i_2 by an edge with weight $\gamma^1_{i_1,i_2}$. All the edges compose a edge set which is denoted by $\mathcal{S}$. One important observation is that our defined pairwise difference operator ${}^1\mathbf{\Delta}$ reduces to univariate ones (fused LASSO penalty) in the case of a chain graph where $\mathcal{S} = \{\mathbf{i} = [i, i+1] |\ i = 1, 2, ..., d_1 - 1\}$. This is exactly the fusion structure used in Sun and Li [2019].

In practice, choosing appropriate weights $\gamma^1_{i_1,i_2}$ is critical for the following two aspects. First, $\gamma^1_{i_1,i_2}$ affects the quality of clustering accuracy. Chi and Lange [2015] showed that uniform weights result in a little agglomerative clustering path, e.g., not exact sparsity in the pairwise difference except in the case when there is a large separation between two clusters. On contrast, assigning $\gamma^1_{i_1,i_2}$ with large value may erroneously cluster two rows $\mathbf{A}_{i_1:}, \mathbf{A}_{i_2:}$ from different groups into one cluster. Second, setting $\gamma^1_{i_1,i_2} = 0$ for some $\{i_1, i_2\}$ will make undirected graph constructed by the pairwise difference of rows to be sparse, resulting in computational efficiency in the Fused-Orth-ALS algorithm.

Several related papers for convex co-clustering (She et al. [2010], Chen et al. [2015], Chi and Lange [2015], Chi et al. [2018]) use a similar weights computation strategy described as follows. Since we have to calculate a 'distance' between pairs of rows for $\mathbf{A}$, our first step is to get an initial estimate $\hat{\mathbf{A}}$ from initial observations $\mathcal{Y}$. One approach is to find a rank-K CP decomposition approximation to $\mathcal{Y}$ and use the factor matrix recovered from CP decomposition as $\hat{\mathbf{A}}$. After obtaining this initial estimate, we can construct the weight estimate as

$$\gamma^1_{i_1,i_2} = \iota^k_{i_1,i_2} \exp\Big(- \tau \|\hat{\mathbf{A}}_{i_1:} - \hat{\mathbf{A}}_{i_2:}\|_2 \Big) \tag{2.7}$$

where $\iota^k_{i_1,i_2}$ is an indicator function that equals 1 if the i_2th row is among the i_1th row's k-nearest neighbors and 0 otherwise. This parameter controls how 'sparse' the undirected graph will be and as a default we can choose the smallest k that ensures the graph is still connected. The main component of weights in (2.7) is a Gaussian kernel which has been widely used in Chen et al. [2015], Chi and Lange [2015], Chi et al. [2018]. τ is a measure of scale and, in practice, we can set it to be the median Euclidean distance between the i_1th and i_2th rows that are k-nearest neighbors of each other.

Lastly, for multi-mode clustering over multiple modes of tensor, we need to normalize the sum of weights over first mode, e.g. $\sum_{[i_1,i_2]\in\mathcal{S}} \gamma^1_{i_1,i_2} = \frac{d_1}{d_1+d_2+d_3}$ to make sure that penalty term over three modes are on the same scale, and any single mode will not dominate the entire multi-modes clustering as λ increases.

2.3.3 Tuning parameter

To implement the Fused-Orth-ALS algorithm in practice, tuning parameter choice is a critical issue. Tuning parameters include λ which controls the level of regularization, rank K which controls the underlying CP decomposition structure, and the number of clusters s_1, s_2, s_3 over each mode and thus we should choose them carefully. Though Cross Validation (CV) has been widely used for choosing tuning parameter for vector-based or matrix-based data format, the computation burden in tensor problems makes CV difficult to implement. Thus, we follow the universal rule for tuning parameter choice in multi-modes clustering problem [Sun and Li, 2019, Chi et al., 2018, Wang and Zeng, 2019], which is based on the extended Bayesian Information Criterion (eBIC) proposed by Chen and Chen [2008, 2012]. The tuning parameter λ is chosen by minimizing

$$\log\Big(\frac{\|\mathcal{Y}-\sum_{i\in[K]} w_i\hat{\mathbf{A}}_{:i}\otimes\hat{\mathbf{B}}_{:i}\otimes\hat{\mathbf{C}}_{:i}\|_F^2}{\prod_{i=1}^3 d_j}\Big)+\frac{\sum_{j=1}^3\log d_j}{\prod_{j=1}^3 d_j}\times \mathrm{df}_\lambda \tag{2.8}$$

where df_λ is defining the degrees of freedom for a particular value of λ. We use the number of unique non-zero elements in $\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}$ as the estimate of df_λ since the number of unique non-zero elements in $\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}$ is the total number of cluster mean values we need to estimate (they can be considered as the parameters of our model). We will calculate the value of eBIC in (2.8) over an array $\boldsymbol{\lambda}$ of candidates and choose the optimal one which satisfies

$$\lambda^* = \arg\min_{\lambda \in \boldsymbol{\lambda}} \mathrm{eBIC}(\lambda) \tag{2.9}$$

We can choose rank K using a similar strategy as that for λ. That is, we can set a grid of candidate values for different combination of (λ, K), and select the one which minimize eBIC. However, this causes computational inefficiency when we have a large number of candidates for (λ, K). In fact, the 'elbow point' method is an easy-implementation method to choose K. After plotting the recovery error $\|\mathcal{Y} - \sum_{i\in[K]} w_i \hat{\mathbf{A}}_{:i} \otimes \hat{\mathbf{B}}_{:i} \otimes \hat{\mathbf{C}}_{:i}\|_F^2$ of classical ALS algorithm as a function of rank K, we can choose K according to the elbow point in the plot.

Several methods can be applied to choose the number of clusters. Intuitively, the 'elbow point' method is quick and efficient to choose number of cluster by choosing $s_i, \forall i \in \{1, 2, 3\}$ that minimize the intra-class variation. Alternative methods are proposed in the literature regarding the number of clusters, like gap statistics established in Tibshirani et al. [2001] and stability-based method proposed in Wang [2010], Fang and Wang [2012].

The experiments in this chapter use three steps for tuning method choice. First use 'elbow point' method to choose rank K by implementing classical ALS algorithm. Second, choose the optimal λ based on eBIC and finally, the number of clusters are determined by gap statistics.

2.4 Theoretical Properties

In this section, we provide convergence analysis for the Fused-Orth-ALS algorithm. Throughout this section, we will present the convergence analysis for order three tensors; results can be easily generalized to higher order tensors with $D > 3$. For simplicity, we set the dimension over three modes the same, e.g. $d_1 = d_2 = d_3 = d$ and choose regularization weights over all modes as uniform. Without loss of generality, we assume $w_{\max} = w_1 \geq w_2 ... \geq w_K = w_{\min} > 0$.

Through convergence result analysis for the Fused-Orth-ALS algorithm, we can derive the recovery error bound and clustering consistency directly if mild conditions are imposed on the error tensor $\mathcal{E}$ and the minimum weights signal $w_{\min}$. Thus, we will introduce the theoretical properties for recovery consistency and clustering consistency respectively in the remaining parts of this section.

2.4.1 Recovery Error

To obtain convergence results for the Fused-Orth-ALS algorithm, we need some deterministic conditions on the tensor factor matrices. We state the convergence result for factor matrix over the third mode $\mathbf{C}$ and analogous results can be derived for $\mathbf{A}, \mathbf{B}$ as well.

Assumption 1. *(Incoherence)*

$$\rho := \max_{i \neq j} \{|\langle \mathbf{A}_{:i}, \mathbf{A}_{:j} \rangle|, |\langle \mathbf{B}_{:i}, \mathbf{B}_{:j} \rangle|, |\langle \mathbf{C}_{:i}, \mathbf{C}_{:j} \rangle|\} \leq \frac{\alpha}{\sqrt{d}} \tag{2.10}$$

for some $\alpha = polylog(d)$ and $K\rho^2 = o(1)$. Furthermore, spectral norm of $\mathcal{Y}$ satisfies

$$\|\mathcal{Y}\| \leq w_{\max}\alpha \tag{2.11}$$

Assumption 1 is commonly used in the tensor decomposition literature [Anandkumar

et al., 2014, Sun et al., 2017, Sun and Li, 2019], and it relaxes the requirements on the orthogonality of columns in factor matrices. Anandkumar et al. [2014] provides detailed justification for (2.10) which is satisfied if columns of factor matrices are uniformly i.i.d drawn from the unit sphere. Notice that it is also reasonable to assume this assumption holds for some non-random matrices. The other statement (2.11) puts constraint on the spectral norm of factor matrices which can be proved to be satisfied with high probability using bounded tensor spectral norm (Tomioka and Suzuki [2014]). Due to $\rho \leq \frac{\alpha}{\sqrt{d}}$, the condition $K\rho^2 = o(1)$ is equivalent to $K = o(d)$ which puts the same order requirement on rank K. However, Sharan and Valiant [2017] requires a more strict condition on rank K which is $K = o(d^{0.25})$. Even if our rank requirement is stricter than the best known recovery guarantees for polynomial-time algorithm on random tensors, which works even under over-complete settings with $k = o(d^{1.5})$, we consider that the lower order rank shortcoming for Fused-Orth-ALS algorithm is more a property of the analysis strategy than the algorithm itself.

Define the initialization error as $\epsilon_0 = \max\{\text{dist}(\hat{\mathbf{A}}_{:i}^{(0)}, \mathbf{A}_{:i}), \text{dist}(\hat{\mathbf{B}}_{:i}^{(0)}, \mathbf{B}_{:i})\}$ and we would like to impose an initialization condition on ϵ_0 in the following assumption.

Assumption 2. *(Initialization)*
We assume the initialization error ϵ_0 satisfies

$$\epsilon_0 \leq \min\Big\{\frac{1}{6\gamma} - \rho^2(K-1), \frac{1}{12\sqrt{2}\gamma\alpha} - \frac{2\rho(K-1)}{\alpha}, \frac{(\sqrt{2}-1)\sqrt{d}/(K-1) - \alpha(1+16\sqrt{2}\gamma K)}{\sqrt{d}+16\sqrt{2}\alpha\gamma K}\Big\}$$

For simplicity, we denote $\gamma = \frac{w_{\max}}{w_{\min}}$. We recognize that assumption 2 restricts initialization for Fused-Orth-ALS algorithm to be related with weights ratio γ, rank K and dimension d. However, these assumptions are needed only for computational issues, and detailed explanation for each term can be found in Appendix A.1. We only put the initialization condition on the first two factor matrices since $\hat{\mathbf{C}}_{:i}^{(0)}$ will be calculated through Fused-Orth-ALS algorithm and $\text{dist}(\hat{\mathbf{C}}_{:i}^{(0)}, \mathbf{C}_{:i})$ will be bounded by theorem 1.

Assumption 3. *(Bounded Perturbation)*

Suppose the spectral norm of error tensor denoted by ψ is bounded as below

$$\psi \leq \min\left\{\frac{w_{\min}}{6}, \frac{w_{\max}K}{d}\right\}$$

Assumption 3 bounds the perturbation level in terms of the spectral norm of error tensor. This can be easily satisfied with high probability if we assume each element in $\mathcal{E}$ follows an i.i.d sub-Gaussian distribution as stated in Tomioka and Suzuki [2014].

Assumption 4. *(Bounded fusion)*

Let $\zeta := \max_i \| {}^3\mathbf{\Delta}\mathbf{C}_{:i}\|_1$. Suppose ζ is bounded by

$$\zeta \leq \frac{w_{\max}(\epsilon_0^2\alpha + 2\epsilon_0\rho(K-1) + \rho^2(K-1)) + \psi}{2Mw_{\min}(1-\epsilon_0^2)}$$

We employ the bounded fusion assumption 4 on the clustering structure to restrict the clustering complexity on factor matrix. To clarify this point, we use a simple example to illustrate what is the meaning of it. Suppose we have s_3 clusters over the third factor matrix, each cluster has the same size d/s_3, and we use uniform weights for pairwise difference operator ${}^3\mathbf{\Delta}$. For rows in $\mathbf{C}$ belonging to the same cluster, theoretically speaking, they should take the same value as the ith element of cluster mean under our proposed model. Then assumption 4 can be rephrased as $\| {}^3\mathbf{\Delta}\mathbf{C}_{:i}\|_1 = O(d^{1.5}(1-1/s_3))$ and in a special case where $\mathbf{C}$ has the simplest clustering structure with only one cluster, e.g. $s_3 = 1$, assumption 4 refers to $\| {}^3\mathbf{\Delta}\mathbf{C}_{:i}\|_1 = 0$. Thus, when sample size over third mode is kept fixed, bounded ψ is equivalent to bounded cluster size s_3.

Now we can demonstrate the local convergence guarantee or recovery error bound for the Fused-Orth-ALS algorithm in the following theorem.

Theorem 1. *Assume Assumption 1-4 hold, factor matrix estimate $\hat{\mathbf{C}}_{:i}, \forall i \in [K]$ of Algorithm*

2 satisfies the following error bound with high probability

$$\|\hat{\mathbf{C}}_{:i} - \mathbf{C}_{:i}\|_2^2 \lesssim \gamma\rho^2(K-1) + \frac{\psi}{w_{\min}}$$

by choosing $\lambda \geq 2M(w_{\max}(\epsilon_0^2\alpha + 2\epsilon_0\rho(K-1) + \rho^2(K-1)) + \psi)/[w_i(1-\epsilon_0^2) - w_{\max}(\epsilon_0^2\alpha + 2\epsilon_0\rho(K-1) + \rho^2(K-1)) - \psi]$ *where* $M = \max_j \| {}^3\mathbf{\Delta}_j^\dagger \|_2$. *Similar error bounds hold for the other two factor matrices* $\mathbf{A}, \mathbf{B}$.

The error bound derived in Theorem 1 reveals how the signal w_i, incoherence parameter ρ, rank K and perturbation level ψ interact with each other on affecting the convergence behavior of Fused-Orth-ALS algorithm. Lower perturbation level ψ, lower incoherence parameter ρ, rank K and large signal ratio of $w_{\max}/w_{\min}$ all result in lower error bound. Clearly, the error bound for each column in factor matrices are bounded by two parts: one is related to perturbation level ψ, the other is dependent on tensor underlying CP decomposition structure, weights ratio γ, rank K and dimension along each mode d. Thus, under high dimensional settings when $d \to \infty$, i.e., $\rho^2 \leq \alpha^2/d \to 0$, the first part will dominate the error bound while if tensor data is almost noiseless, e.g. $\psi \approx 0$, the second part will be the main source for error. The following corollary analyzes the relationship between these two parts under a special case when the error tensor follows a sub-Gaussian distribution.

Corollary 1. *Assume the conditions in Theorem 1 hold, and we further assume that each element in error tensor,* $\mathcal{E}_{ijk}$, *is independent, zero-mean and satisfies* $\mathbb{E}[e^{t\mathcal{E}_{ijk}}] \leq e^{\frac{\sigma^2 t^2}{2}}$, $\gamma \leq C$ *where* C *is positive constant and the minimal weight satisfies*

$$w_{\min} \succ \sqrt{\sigma^2\Big[3d\log\frac{6}{\log 3/2} + \log\frac{2}{\delta}\Big]d^2/(K-1)^2}$$

then updates $\hat{\mathbf{C}}_{:i},\ \forall i \in [K]$ *from Algorithm 2 satisfies the following error bound,*

$$\|\hat{\mathbf{C}}_{:i} - \mathbf{C}_{:i}\|_2 \lesssim \frac{(K-1)}{d}$$

with probability at least $1-\delta$*. Same error bounds hold for other two factor matrices* $\mathbf{A}, \mathbf{B}$*.*

By imposing the sub-Gaussian distribution assumption on error tensor elements, we can achieve recovery consistency for factor matrices estimated from Fused-Orth-ALS algorithm.

2.4.2 Clustering Consistency

We establish clustering consistency for the clustering algorithm performed on factor matrices recovered by Fused-Orth-ALS algorithm. For simplicity, we continue to set the dimension over three modes to be the same, e.g. $d_1 = d_2 = d_3 = d$. Recall that the clustering algorithm are performed on the rows of $\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}$, thus clustering error is quantified through the true mean value and its estimate $\|\boldsymbol{\mu}^*_{1,j_1} - \hat{\boldsymbol{\mu}}_{1,j_1}\|_2, \|\boldsymbol{\mu}^*_{2,j_2} - \hat{\boldsymbol{\mu}}_{2,j_2}\|_2, \|\boldsymbol{\mu}^*_{3,j_3} - \hat{\boldsymbol{\mu}}_{3,j_3}\|_2$.

Theorem 2. *Assume the conditions in Corollary 1 hold. Then we have*

$$\max_{j_1 \in \mathfrak{A}_m, m \in \{1,2,\ldots,s_1\}} \|\hat{\boldsymbol{\mu}}_{1,j_1} - \boldsymbol{\mu}^*_{1,j_1}\|_2 \lesssim \frac{K^{1.5}}{d}$$
$$\max_{j_2 \in \mathfrak{B}_n, n \in \{1,2,\ldots,s_2\}} \|\hat{\boldsymbol{\mu}}_{2,j_2} - \boldsymbol{\mu}^*_{2,j_2}\|_2 \lesssim \frac{K^{1.5}}{d}$$
$$\max_{j_3 \in \mathfrak{C}_l, l \in \{1,2,\ldots,s_3\}} \|\hat{\boldsymbol{\mu}}_{3,j_3} - \boldsymbol{\mu}^*_{3,j_3}\|_2 \lesssim \frac{K^{1.5}}{d}$$

hold with probability at least $1-\delta$*. Furthermore, if* $\min_{j_1 \in \mathfrak{A}^*_m, j_1' \in \mathfrak{A}^*_{m'}, m \neq m'} \|\boldsymbol{\mu}^*_{1,j_1} - \boldsymbol{\mu}^*_{1,j_1'}\|_2 \gtrsim \frac{K^{1.5}}{d}, \min_{j_2 \in \mathfrak{B}^*_n, j_2' \in \mathfrak{B}^*_{n'}, n \neq n'} \|\boldsymbol{\mu}^*_{2,j_2} - \boldsymbol{\mu}^*_{2,j_2'}\|_2 \gtrsim \frac{K^{1.5}}{d}, \min_{j_3 \in \mathfrak{C}^*_l, j_3' \in \mathfrak{C}^*_{l'}, l \neq l'} \|\boldsymbol{\mu}^*_{3,j_3} - \boldsymbol{\mu}^*_{3,j_3'}\|_2 \gtrsim \frac{K^{1.5}}{d}$*, we have* $\hat{\mathfrak{A}}_m = \mathfrak{A}^*_m$*,* $\hat{\mathfrak{B}}_n = \mathfrak{B}^*_n$*,* $\hat{\mathfrak{C}}_l = \mathfrak{C}^*_l$ *hold with probability at least* $1-\delta$*.*

This theorem shows that clustering consistency holds as long as $K^{1.5}/d \to 0$, which allows rank K increase with the dimension size d. At first look, the conclusion in theorem 2

seems to indicate that the cluster mean error bound does not depend on number of clusters over each mode, s_1, s_2, s_3. However, we assume that assumption 4 holds which employs an 'invisible' bound on the number of clusters. As we have analyzed before, $\| {}^3\mathbf{\Delta C}_{:i}\|_1 = O(d^{1.5}(1-1/s_i)), \forall i \in \{1,2,3\}$ which is closely related with convergence rate $K^{1.5}/d$ in theorem 2 under assumptions for Corollary 1 (we provide detailed explanation in Appendix A.2). Thus, cluster mean error bound $K^{1.5}/d$ increases with the number of clusters s_i. Note that theorem 2 assumes that true rank K is known and thus the effect of estimated rank $\hat{K}$ on clustering consistency needs to be further explored.

2.5 Numerical Experiments

To investigate the performance of Fused-Orth-ALS algorithm on multi-modes tensor clustering, we perform simulation experiments. We consider the Fused-Orth-ALS algorithm on finite sample performance and compare the recovery error and clustering error with other alternative tensor-based clustering methods. We use two criteria to assess the performance of tensor recovery and clustering. Recovery error is defined as $\frac{\|\hat{\mathcal{Y}}-\mathcal{Y}^*\|_F}{\|\mathcal{Y}^*\|_F}$ where $\hat{\mathcal{Y}}$ is the estimation from Fused-Orth-ALS algorithm. In general, clustering error measuring the probability of mismatches between the estimated clustering assignments $\hat{\mathfrak{M}}$ and the true one $\mathfrak{M}$ over sample $x_1, ..., x_n$ is

$$\binom{n}{2}^{-1}\left|\{(i,j): \mathbf{1}_{\hat{\mathfrak{M}}(x_{j_1})=\hat{\mathfrak{M}}(x_{j_2})} \neq \mathbf{1}_{\mathfrak{M}(x_{j_1})=\mathfrak{M}(x_{j_2})}, j_1 < j_2, j_1, j_2 \in [n]\}\right|$$

Naturally, it's simple to compute the clustering error along each mode by employing this definition. To assess the quality of clustering performance, the above clustering error is similar to other two common used measures in clustering literature, like adjusted Rand index [Hubert and Arabie, 1985] and variation of information [Meilă, 2007].

In particular, elements in error tensor $\mathcal{E}$ are i.i.d generated from Gaussian distribution

with mean 0 and variance σ^2. In each simulation study, the summary statistics are based on 50 replications and we report the corresponding standard deviation (in parenthesis).

2.5.1 Finite sample performance

We start from evaluating the Fused-Orth-ALS algorithm performance on finite sample and the relationship between recovery error, clustering error with different parameters including perturbation level ψ, dimension d_i as shown in theorem 1 and theorem 2. Theorem 1 reveals that the convergence bound for each mode are in the same form with respect to parameters over each mode. Thus, our first experiment takes an order three tensor example and we would like to do clustering over the third mode. Assume order three tensor is generated under CP decomposition structure in (2.2) and (2.3) with rank $K = 2$ and we set the dimension of the first two matrices the same e.g. $d_1 = d_2 = d$ and their unnormalized columns are

$$\begin{aligned}
\mathbf{A}_{:1} = \mathbf{B}_{:1} &= (\mu, -\mu, 0.5\mu, -0.5\mu, \underbrace{0, ..., 0}_{d-4})^\top \\
\mathbf{A}_{:2} = \mathbf{B}_{:2} &= (0, 0, 0, 0, \alpha\mu, -\alpha\mu, 0.5\alpha\mu, -0.5\alpha\mu, \underbrace{0, ..., 0}_{d-8})^\top
\end{aligned} \tag{2.12}$$

The third factor matrix with unnormalized & unshuffled column is generated by

$$\begin{aligned}
\mathbf{C}_{:1} &= (\underbrace{\mu, ..., \mu}_{\lfloor d_3/2 \rfloor}, \underbrace{-\mu, ..., -\mu}_{\lfloor d_3/2 \rfloor}) \\
\mathbf{C}_{:2} &= (\underbrace{-\mu, ..., -\mu}_{\lfloor d_3/4 \rfloor}, \underbrace{\mu, ..., \mu}_{\lfloor d_3/2 \rfloor}, \underbrace{-\mu, ..., -\mu}_{\lfloor d_3/4 \rfloor})
\end{aligned} \tag{2.13}$$

Then we shuffle the row of $\mathbf{C}$ to make the samples from same cluster not necessarily in consecutive order. There are four clusters over third mode with cluster means as $(\mu, -\mu), (\mu, \mu)$, $(-\mu, \mu), (-\mu, -\mu)$ respectively. After normalizing the columns of $\mathbf{A}, \mathbf{B}, \mathbf{C}$, we can calculate the weights w_i.

The above simulation example is inspired from Sun and Li [2019] but with the added permutation to the rows of factor matrix $\mathbf{C}$ which is the mode we are going to perform clustering on. This imposes an additional challenge on clustering since we need a more sophisticated penalty fusion structure to capture the clustering features. Parameters in this example characterize different aspects of model. μ represents the clustering difficulty since it quantifies the distance between different clusters. The smaller the value μ takes, more complicated the clustering task will be. Different value of α controls different level of signal max/min ratio γ. For simplicity, the factor matrices we used in this example, $\mathbf{A}, \mathbf{B}, \mathbf{C}$, are orthogonal, satisfying Assumption 1.

Moreover, we would like to see how recovery error and cluster error change as the noise level σ, the sample size over third mode d_3, weights max/min ratio γ change. Results for multiple experiments are shown in Figure 2.1. Top left panel in Figure 2.1 indicates that recovery error increases linearly as we change the noise level σ from 0 to 2. As we increases sample size d_3 from 20 to 200, recovery error decreases roughly at the rate of $1/d_3$. The trend in these two figure is consistent with our theoretical result in theorem 1. Note in top right panel of Figure 2.1 that clustering error increases as noise level σ increases. As reflected by bottom right panel of Figure 2.1, clustering error decreases at a rate of $1/d_3$ as d_3 increases which validates the clustering error bound provided in theorem 2.

Theoretical results show that a large number of clusters s_3 gives rise to recovery error and clustering error. To test this conclusion, we modify third mode factor matrix $\mathbf{C}$ to increase the number of cluster from 2 to 8. Remember that we repeat the row shuffling for $\mathbf{C}$ to make sure that rows from same cluster are not necessarily adjacent to each other. The detailed choice of cluster mean value for different number of clusters are illustrated in Table 2.1. Corresponding recovery error and clustering error for experiments with different number of clusters are provided in Table 2.2, which is in agreement with previous analysis. Next, we continue to use the simulation setting defined in (2.12) and (2.13) and we would like to

s_3	cluster mean
2	$(\mu,\mu),(\mu,-\mu)$
4	$(\mu,\mu),(\mu,-\mu),(-\mu,\mu),(-\mu,-\mu)$
6	$(\mu,\mu),(\mu,-\mu),(-\mu,\mu),(-\mu,-\mu),(0,\mu),(\mu,0)$
8	$(\mu,\mu),(\mu,-\mu),(-\mu,\mu),(-\mu,-\mu),(0,\mu),(\mu,0),(0,-\mu),(-\mu,0)$

Table 2.1: Cluster center mean choice for $\mathbf{C}$ with different number of clusters s_3

	$s_3=2$	$s_3=4$	$s_3=6$	$s_3=8$
Recovery Error	0.504(0.0871)	0.507(0.0829)	0.555(0.0604)	0.614(0.1109)
Clustering Error	0.019(0.0328)	0.025(0.0490)	0.109(0.0326)	0.121(0.0301)

Table 2.2: Recovery error and clustering error with different number of cluster s_3. (Model setting: $\mu=1, d_1=d_2=8, d_3=40, \sigma=1, \alpha=1$)

show effect of pairwise difference operator ${}^{3}\mathbf{\Delta}$ on tensor recovery and clustering compared to the simple fusion structure ${}^{\text{Fuse}}\mathbf{\Delta}$ which is defined as

$$ {}^{\text{Fuse}}\mathbf{\Delta} \in \mathbb{R}^{(d_3-1)\times d_3} = \begin{bmatrix} 1 & -1 & 0 & ... & 0 \\ 0 & 1 & -1 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & ... & 1 & -1 \end{bmatrix} $$

As noted above, μ controls the difficulty level of clustering. We set $d_1 = d_2 = 8, \sigma = 0.001, d_3 = 20$ and vary the value for $\mu \in \{0.1, 0.3, 0.5, 0.7\}$. By shrinking the value μ from 0.7 to 0.1, the 'signal' level becomes weak and, at the same time, the boundary between different clusters will be less apparent, which increases the difficulty for the clustering task. Table 2.3 summarizes recovery error, clustering error and computational running time comparison of Fused-Orth-ALS algorithm by imposing different fusion structure, ${}^{3}\mathbf{\Delta}$ and ${}^{\text{Fuse}}\mathbf{\Delta}$. Obviously, ${}^{3}\mathbf{\Delta}$ outweighs ${}^{\text{Fuse}}\mathbf{\Delta}$ in terms of the fact that it leads to smaller recovery error and clustering error especially when clustering signal μ becomes weak. Still, the computational time increases for ${}^{3}\mathbf{\Delta}$ are still acceptable and even comparable to ${}^{\text{Fuse}}\mathbf{\Delta}$ when μ is small.

		${}^{3}\mathbf{\Delta}$	${}^{\text{Fuse}}\mathbf{\Delta}$
$\mu = 0.1$	Recovery error	**0.8198(0.2034)**	1.0538(0.0224)
	Clustering error	**0.1902(0.1475)**	0.3847(0.0394)
	Running time(sec)	4.3557(3.5838)	2.2829(0.9545)
$\mu = 0.3$	Recovery error	**0.0142(0.0015)**	0.1991(0.2256)
	Clustering error	**0(0)**	0.0239(0.0727)
	Running time(sec)	0.3480(0.0786)	0.4659(0.8681)
$\mu = 0.5$	Recovery error	**0.0030(0.0003)**	0.0189(0.0023)
	Clustering error	**0(0)**	0(0)
	Running time(sec)	0.3719(0.1078)	0.0896(0.0314)
$\mu = 0.7$	Recovery error	**0.0011(0.0001)**	0.0069(0.0006)
	Clustering error	**0(0)**	0(0)
	Running time(sec)	0.4079(0.1456)	0.0753(0.0241)

Table 2.3: Performance comparison for ${}^{3}\mathbf{\Delta}$ and ${}^{\text{Fuse}}\mathbf{\Delta}$ under different signal level μ

2.5.2 Comparison with alternative methods

We would like to compare the performance of Fused-Orth-ALS Algorithm with the following three methods: the dynamic tensor clustering (DTC) algorithm proposed in Sun and Li [2019]; the CP-Kmeans algorithm which performs a rank K CP decomposition on the tensor observations first and then independently applies k-means algorithm clustering to the rows of factor matrices of resulting CP decomposition; and multiway clustering for tensor block models (TBM) proposed in Wang and Zeng [2019].

First, we consider the performance of single-mode clustering. We use the simulation setting defined in section 2.5.1 to compare algorithm performance in a single mode clustering scenario. Here we set the model settings as $d_1 = d_2 = 20, d_3 = 48, \mu = 1, \alpha = 1$ and vary $\sigma \in \{0, 0.25, 0.5, 0.75, 1\}$. We compare the performance of four algorithms, e.g., Fused-Orth-ALS, DTC, CP-Kmeans and TBM in the following aspects: recovery error, clustering error, average convergence running time in seconds and iterations required for convergence. Table 2.4 provides a summary for the performance of these four algorithms.

As summarized in table 2.4, Fused-Orth-ALS outperforms the other three methods under different values of perturbation, achieving the best recovery error and clustering error. Also,

σ		Fused-Orth-ALS	DTC	CP-Kmeans	TBM
0	Recovery	**0.002(0.0002)**	0.004(0.0003)	0.018(0.084)	1(0)
	Clustering	**0(0)**	0(0)	0(0)	0.7659(0)
	Time	0.3342(0.1007)	0.2876(0.0760)	0.1180(0.0787)	0.0256(0.0077)
	Iter	6.16(1.52)	8.86(1.73)	-	-
0.25	Recovery	0.1326(0.0068)	**0.1307(0.0068)**	0.1802(0.1628)	1.0023(0.0005)
	Clustering	0(0)	**0(0)**	0.0197(0.0677)	0.3863(0.0278)
	Time	0.5495(0.0985)	0.5875(0.0900)	0.1163(0.0699)	0.0597(0.0146)
	Iter	11.02(1.78)	19.64(2.25)	-	-
0.5	Recovery	**0.274(0.0138)**	0.282(0.0742)	0.304(0.1227)	1.0087(0.0017)
	Clustering	**0(0)**	0.005(0.0328)	0.0151(0.0605)	0.3873(0.0337)
	Time	0.9384(0.3258)	0.9788(0.4728)	0.1729(0.1126)	0.0628(0.0362)
	Iter	19.6(6.87)	36.12(17.65)	-	-
0.75	Recovery	**0.467(0.1225)**	0.476(0.1347)	0.5351(0.1851)	1.0191(0.0040)
	Clustering	**0.0192(0.0633)**	0.0262(0.0755)	0.0577(0.1061)	0.3782(0.0209)
	Time	2.0011(0.8024)	1.8025(0.7251)	0.2543(0.1123)	0.0639(0.0142)
	Iter	39.64(14.42)	62.92(24.73)	-	-
1	Recovery	**0.645(0.137)**	0.691(0.162)	0.844(0.207)	1.0340(0.0058)
	Clustering	**0.047(0.081)**	0.075(0.097)	0.156(0.124)	0.3835(0.0271)
	Time	3.5724(1.1233)	3.3227(1.0192)	0.3071(0.0848)	0.0697(0.0319)
	Iter	74.32(22.76)	118.84(34.55)	-	-

Table 2.4: Performance comparison for Fused-Orth-ALS, Dynamic tensor clustering (DTC), CP-Kmeans and Tensor block model (TBM)

Mode	Cluster Mean
1	$(\mu, 0.5\mu), (-0.5\mu, \mu), (0, -\mu), (-\mu, 0)$
2	$(0, \mu), (-\mu, 0), (0.1\mu, -\mu), (-\mu, -0.1\mu)$
3	$(\mu, \mu), (-\mu, \mu), (\mu, -\mu), (-\mu, -\mu)$

Table 2.5: Cluster mean choice for factor matrices

even though there is an orthogonal step in Fused-Orth-ALS algorithm which may increase convergence running time, the average running time is still comparable or even smaller than convergence running time for dynamic tensor clustering algorithm. In addition, we can see the advantage of quick convergence for Fused-Orth-ALS which is based on Alternating Least Squares algorithm over dynamic tensor clustering which is based on Tensor Power Method. The average convergence iterations required for Fused-Orth-ALS algorithm is less than that for dynamic tensor clustering, which validates theoretical iceproperties of quick convergence for orthogonal ALS algorithm.

We illustrate next the performance of multi-mode clustering analysis. The performance comparison among three methods are provided: Fused-Orth-ALS algorithm, tensor block model (TBM) and CP-Kmeans. The mean value for different clusters over each mode is provided in Table 2.5. As presented, each mode has 4 clusters of similar size, e.g., $s_1 = s_2 = s_3 = 4$ and cluster size are $\lfloor d_1/4 \rfloor, \lfloor d_2/4 \rfloor, \lfloor d_3/4 \rfloor$. Take the first mode as an example, there are $\lfloor d_1/4 \rfloor$ rows of the factor matrix taking value $(\mu, 0.5\mu)$, $\lfloor d_1/4 \rfloor$ rows taking value $(-0.5\mu, \mu)$, $\lfloor d_1/4 \rfloor$ rows taking value $(0, -\mu)$ and the rest rows taking value $(\mu, 0)$. Recall that these rows are shuffled randomly. Comparison result for three methods under different noise level σ can be found in Figure 2.2. It is clear that Fused-Orth-ALS algorithm outperforms the other two methods over clustering performed on three modes, especially under the noisy case when σ is large.

2.6 Real Data Analysis

We illustrate our Fused-Orth-ALS algorithm performance on two real world datasets: brain nodes structural connectivities from Human Connectome Project (HCP) [Van Essen et al., 2013] and political relationships between nations [Kemp et al., 2006].

2.6.1 Human Connectome Project (HCP)

The first real dataset is an order three tensor $\mathcal{Y} \in \mathbb{R}^{68\times68\times136}$ consisting of brain connectivity among 68 brain nodes for 136 individuals. Each entry takes on ordinal value $\{0, 1, 2\}$ which indicates the strength level of connectivity {low, moderate , high} between different brain nodes.

Details on choosing rank K and the number of clusters can be found in Appendix A.4. In particular, the brain nodes clustering result in Table 2.6 captures the spatial connectivity between hemispheres of brain. Cluster I, II mainly focus on the connectivity in either left or right hemisphere while cluster III represents the cross-section connection between left and right hemisphere. Interestingly, r.supramarginal and l.supramarginal are separately picked as smaller cluster IV,V and we infer that this is related to the fact that those regions are known to play a critical role in visual word recognition and reading.

We also compare the goodness of fit (proportion of variance explained by tensor estimation) in Table 2.8 for four different methods; almost all of them perform at around 92%, but tensor block model seems to slightly outperform other three CP-low-rank-based methods.

2.6.2 Nations

The second dataset is an order three binary tensor $\mathcal{Y} \in \mathbb{R}^{14\times14\times56}$ consisting of 56 political relationships of 14 countries between 1950-1965. Each entry represents the presence or absence of a political action, such as 'treaties', 'send tourists to ' between different nations.

Cluster	Brain Nodes
I	l.insula,l.superiortemporal(3),l.middletemporal(3) l.inferiortemporal(3),l.inferiorparietal,l.lateraloccipital(2)
II	r.insula,r.superiortemporal(3),r.middletemporal(3), r.inferiortemporal(3),r.lateraloccipital(2),r.precuneus,r.lingual
III	l.superiorfrontal(3),l.frontalpole ,l.caudalmiddlefrontal, l.parstriangularis, l.parsopercularis,l.precentral,l.temporalpole,l.postcentral, l.superiorparietal,l.medialorbitofrontal,l.isthmuscingulate,l.precuneus, l.cuneus,l.parahippocampal,l.lingual, r.superiorfrontal(3),r.frontalpole,r.caudalmiddlefrontal,r.parstriangularis, r.parsopercularis,r.precentral,r.temporalpole,,r.postcentral, r.superiorparietal,r.inferiorparietal,r.medialorbitofrontal,r.isthmuscingulate, r.cuneus,r.parahippocampal
IV	r.supramarginal(4)
V	l.supramarginal(4)

Table 2.6: Clustering result for 68 brain nodes in HCP dataset (The first alphabet in the node name indicates the left or right hemisphere. The number in the parenthesis indicates the node count with same name)

We include the details for choosing rank K and number of clusters in Appendix A.4. Since 78.9% entries are zero in this dataset, we use a ℓ_0 penalized tensor block model (denoted as 'TBM-Sparse') . The clustering result (Table 2.7) for Fused-Orth-ALS algorithm assigned the nations into 3 clusters, one representing western-bloc countries (Cluster I), one communist bloc (Cluster III), one neutral bloc (Cluster II). This result is consistent with relation structure of political environment after world war II. In addition, the goodness-of-fit for five methods are provided in Table 2.8. For this nations dataset, Fused-Orth-ALS algorithm achieves the highest variance proportion, indicating good clustering since the within cluster variance are small.

Cluster	Characteristic	Country
Cluster I	Western	Brazil, Netherlands, UK, USA
Cluster II	Neutral	Burma, Egypt, Israel, Jordan,India, Indonesia
Cluster III	Communist	China, Cuba, Poland, USSR

Table 2.7: Clustering result for 14 nations in Nations dataset

	Fused-Orth-ALS	DTC	TBM	TBM-Sparse	CP-Kmeans
HCP	0.921	0.921	0.925	-	0.921
Nations	0.522	0.458	0.439	0.433	0.324

Table 2.8: Comparison of goodness-of-fit for HCP and nations dataset

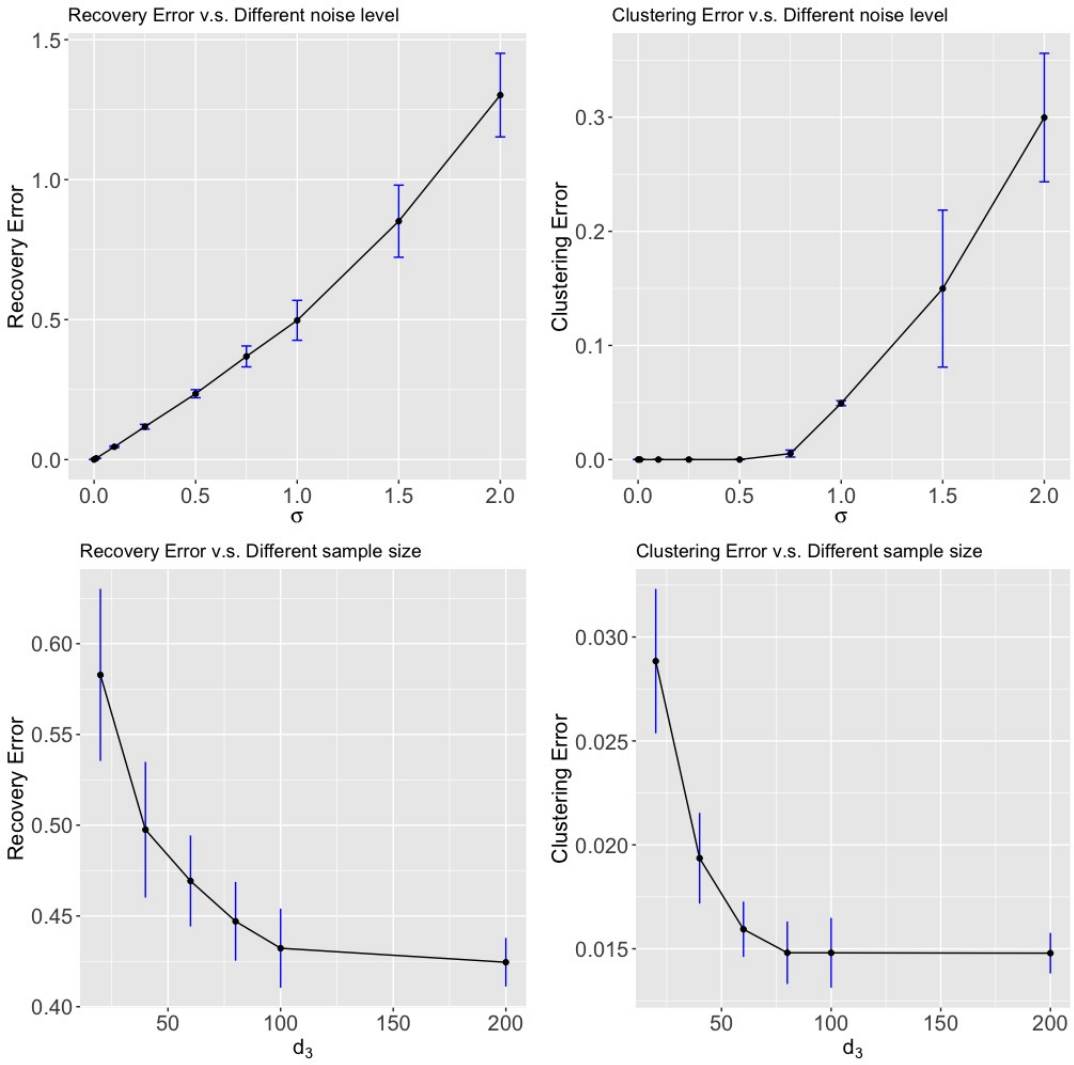


Figure 2.1: (Top left): Recovery Error under different noise level (Model setting $\mu = 1, \alpha = 1, d_1 = d_2 = 8, d_3 = 40$ and $\sigma \in \{0, 0.01, 0.1, 0.25, 0.5, 0.75, 1, 1.5, 2\}$. (Bottom left): Recovery Error under different sample size d_3 (Model setting $\mu = 1, \alpha = 1, d_1 = d_2 = 8, \sigma = 1$ and $d_3 \in \{20, 40, 60, 80, 100, 200\}$). (Top right): Clustering Error under different noise level σ (Model setting $\mu = 1, \alpha = 1, d_1 = d_2 = 8, d_3 = 40$ and $\sigma \in \{0, 0.01, 0.1, 0.25, 0.5, 0.75, 1, 1.5, 2\}$). (Bottom right): Clustering error under different sample size d_3 (Model setting $\mu = 1, \alpha = 1, d_1 = d_2 = 8, \sigma = 1$ and $d_3 \in \{20, 40, 60, 80, 100, 200\}$).

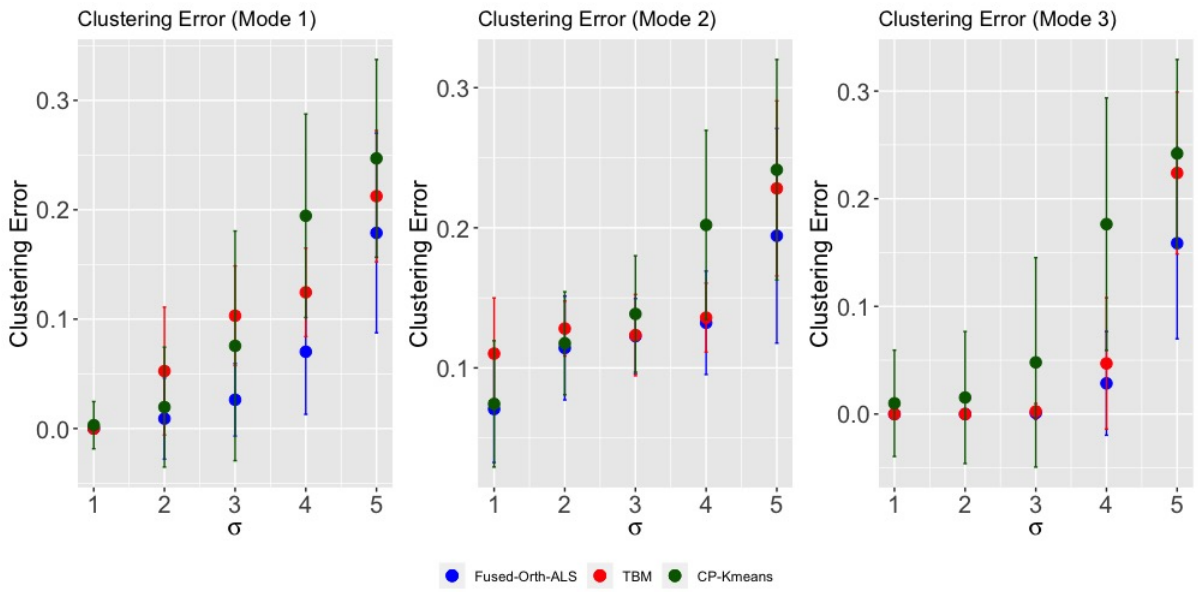


Figure 2.2: Multi-modes clustering performance comparison for Fused-Orth-ALS, CP-Kmeans and Tensor block model (TBM) (Model setting: $d_1 = d_2 = 20, d_3 = 40, \mu = 1, \sigma \in \{1, 2, 3, 4, 5\}$)

CHAPTER 3

LOW RANK TENSOR COMPLETION WITH FIBERS MISSING NOT AT RANDOM

3.1 Introduction

Studies on tensors, multidimensional array, have revealed their crucial and exigent impact on analyzing the interactions among different dimensions in many real applications such as recommendation system [Frolov and Oseledets, 2017, Bi et al., 2018, Zhang et al., 2021], neurogenomics study of brain [Liu et al., 2022, Zhang et al., 2014, Sun and Li, 2017] and biomedical imaging analysis [Tang et al., 2020, Zhou et al., 2013]. However, a large proportion of missing data in tensor observations poses huge challenges in achieving two major goals for different research and business projects. One goal is imputing the missing values in tensors, which has broad applications to recommendation systems related to movie ratings or online shopping. In particular, users' ratings or purchasing histories are collected through combining multiple information sources, including item profile, time, location and any promotion strategies. It is understandable that users cannot rate or purchase all the items, and imputing missing values provides beneficial insights in recommending potential items to users that meet their preferences. This can boost customer satisfaction, loyalty and increase corporate's profit from a long term perspective. Another goal is to recover underlying latent structure of tensor observations, which results in analyzing complicate interactions among different modes. In this case, the existence of large quantity of missing data may leave negative effects in implementing traditional statistical analysis methodologies. For example, developments in neuroscience are dependent on neuroimaging technologies such as anatomical magnetic resonance imaging (MRI), functional magnetic resonance imaging (fMRI), electroencephalography (EEG), diffusion tensor imaging, and positron emission tomography (PET). Missing data or poor image quality could lead to difficulty in establishing

association between brain images and clinical traits.

The majority of literature in tensor completion focuses on recovering tensors from partial observations which is a random subset of its entries. Yuan and Zhang [2016] filled the gap in generalizing the nuclear norm regularization technique to impose a low rank constraint from matrix completion to tensor completion. Xia and Yuan [2017] analyzed statistical optimality and proposed efficient algorithms in recovering noisy low rank tensor as well. Ghadermarzy et al. [2019] generalized the matrix max-norm definition to tensors and successfully achieved low rank tensor completion using a subset of tensor noisy measurements. Nevertheless, abundance of real application problems belong to another missing schema, where partial fibers along a specific dimension are completely unrevealed. Fibers of tensor are the higher-order analogue of matrix rows and columns. They represents slices of tensors obtained by fixing every but only one index in all dimensions of tensor. Thus, a matrix column is a mode-1 fiber and a row is a mode-2 fiber. In practice, a lot of applications involve tensor data with missing fibers. For example, Genotype-Tissue Expression (GTEx) program consists of transcriptome data in a variety of human tissue types as well as genome sequencing data from a large amount of donors. Wang et al. [2016a] developed a multi-tissue imputation method for gene expression in matrix case, which may ignore the important associations and interactions between genes expression and donor clinical traits. A reasonable way is to formulate the data into an order three tensor with donors, genes and tissues representing three modes, and furthermore implement tensor statistical analysis tools. But the consequent challenge is imputing gene expression missing in fibers. Due to the collection difficulty in inaccessible tissues, fibers (vectors of gene expression) for a certain donor over specific tissues will be unavailable. In summary, despite the powerful tools and algorithms developed for tensor completion problem, tensor completion with missing fibers is an open problem and worth of further investigation.

When considering the potential impact of missing data on the registry findings, another

critical aspect is the underlying reason for why the data are missing, which leads to three typical types of missingness: missing completely at random, missing at random, and missing not at random. Missing fibers is not an exception and our method focuses on the scenario when fibers are missing not at random. Missing not at random in matrix completion has been studied in Schnabel et al. [2016], Ma and Chen [2019], disclosing the significance of dealing with the selection bias using propensity score in the matrix case. However, fibers missing not at random appear frequently in tensor datasets and are not well understood. In the GTEx program, a major challenge was posed by the project's requirements for a diverse set of tissue types from which high-quality RNA could be isolated and characterized. Thus, the donors and their biospecimens should present with no evidence of disease (henceforth termed "normal tissues" or "normal biospecimens"). For example, kidneys from donors who died of renal disease are unqualified in terms of the requirements of the experiments and thus are more likely inaccessible from those donors. Donors having a strong habit in smoking will also affect eligibility of lungs to be included as "normal tissues". Multiple traits for donors (either observable or unobservable) can make a big impact on tissue collections and that's why the assumption that fibers are missing not at random plays an important role in analyzing the association between different genes and expression of particular diseases.

There are several aspects of our contributions to this problem. First, in order to obtain the reliable estimate for the propensity score, we generalize the 1bit matrix completion method [Ma and Chen, 2019] to tensors, which applies a link function on a latent continuous valued parameter tensor to model the propensity score. By imposing max norm with low rank constraint regularization on the parameter tensor, we successfully quantify the error bound for propensity score estimate, which achieves a faster convergence rate than the matrix case. Second, we analyze sample size complexity for deriving consistency estimate for the true underlying tensor given noisy observations with partially observed fibers. Similarly, max norm with low rank constraint regularization is employed on the true tensor since without

assuming further structure on it, there is no hope of recovering the missing entries as they are independent of the observed entries. We prove that with high probability, $O(r^{D-1}d\log d)$ total observations or equivalently, $O(r^{D-1}d)$ fibers are sufficient to get a consistent recovery of an order D tensor. Lastly, we explore highly efficient algorithms both for estimating propensity score and for imputing missing fibers by utilizing the Tucker decomposition of low rank tensors. The efficacy of both algorithms are illustrated via simulations and real datasets.

The organization of this chapter follows this framework: in section 3.2, we briefly demonstrated the validity of the assumption that fibers are missing not at random using a synthetic dataset. Section 3.3 proposes a propensity-scored tensor completion method with missing fibers in two scenarios: experimental settings when propensity score is given (a rare situation), and observational settings when propensity score has to be estimated (the common case). Theoretical properties on empirical risk minimization error bounds under these two scenarios are provided. Section 3.4 narrows down the focus on the observational setting and demonstrates in detail the performance of propensity-scored tensor completion method in recovering true tensor given noisy observations. The corresponding algorithms related to propensity score estimate and imputation for tensor missing fibers can be found in section 3.5. Moreover, section 3.6 involves multiple simulation experiments to validate the efficiency and consistency of proposed algorithm in terms of their finite sample performance. We also include comparison between our proposed method and alternatives on both synthetic datasets and real datasets. We put details on proof of theorems and more simulation experiments in Appendices.

3.2 Background for missing not at random

As mentioned above, tensors with missing fibers appear frequently in modern applications such as recommendation system and biomedical datasets. In particular, the fibers for the

tensor observation are not revealed uniformly at random. In general, the major task is to recover the tensor observation $\mathcal{X}$ through an estimatior $\hat{\mathcal{X}}$ and evaluate how well the predicted tensor reflects the true observations. A commonly used criterion to measure the distance between $\mathcal{X}$ and its estimatior $\hat{\mathcal{X}}$ is mean squared error (MSE), i.e.,

$$L(\hat{\mathcal{X}}) = \frac{1}{\prod_{k=1}^{D} d_k} \sum_{i_1=1}^{d_1} ... \sum_{i_D=1}^{d_D} (\mathcal{X}_{i_1,...,i_D} - \hat{\mathcal{X}}_{i_1,...,i_D})^2 \tag{3.1}$$

Suppose $\mathcal{X}$ is partially observed with missing fibers along kth mode, a conventional practice to estimate $\hat{\mathcal{X}}$ is to minimize (over a specified set of tensors)

$$L_{\text{Naive}}(\hat{\mathcal{X}}) = \frac{1}{|\mathbb{O}|d_k} \sum_{(i_1,..,i_{k-1},i_{k+1},...,i_D)\in\mathbb{O}} \sum_{i_k=1}^{d_k} (\mathcal{X}_{i_1,...,i_D} - \hat{\mathcal{X}}_{i_1,...,i_D})^2 \tag{3.2}$$

$\mathbb{O} := \{(i_1, .., i_{k-1}, i_{k+1}, ..., i_D) | \mathcal{X}_{i_1,..,i_{k-1},:,i_{k+1},...,i_D} \text{ is revealed}\}$ denotes the indices set of observed fibers along the kth mode and $|\mathbb{O}|$ represents the number of indices in $\mathbb{O}$. We call the function we minimize above L_{Naive} (the naive estimator) and it makes sense to use it when fibers are missing completely at random. However, $L_{\text{Naive}}(\hat{\mathcal{X}})$ may not be a good estimator for $L(\hat{\mathcal{X}})$ under the situation of fibers missing not at random due to the error introduced by selection bias. Thus, the Inverse-Propensity-Scoring (IPS) estimator is proposed to deal with this issue that fibers are missing not at random; it minimizes

$$L_{\text{IPS}}(\hat{\mathcal{X}}|\mathcal{P}) = \frac{1}{\prod_{k=1}^{D} d_k} \sum_{(i_1,..,i_{k-1},i_{k+1},...,i_D)\in\mathbb{O}} \sum_{i_k=1}^{d_k} \frac{(\mathcal{X}_{i_1,...,i_D} - \hat{\mathcal{X}}_{i_1,...,i_D})^2}{\mathcal{P}_{i_1,..,i_{k-1},i_{k+1},...,i_D}} \tag{3.3}$$

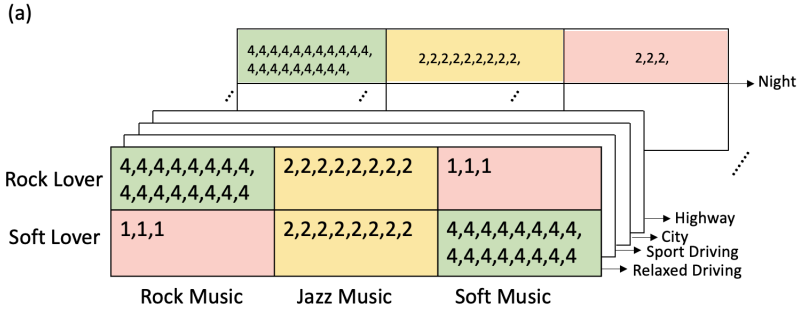
where $\mathcal{P}$ is an order $(D-1)$ tensor, with each component describing the probability of each corresponding fiber being revealed. It can be easily proved that $L_{\text{IPS}}(\hat{\mathcal{X}}|\mathcal{P})$ is an unbiased estimator of $L(\hat{\mathcal{X}})$.

Inspired by the toy example proposed in Schnabel et al. [2016], we will illustrate the

significant effect of selection bias in missing tensor fibers on prediction accuracy using the following synthetic tensor dataset. We use the experimental settings of InCarMusic [Baltrunas et al., 2011] and assume the user ratings on some music songs are collected given specific contextual factors. The data form an order three tensor and we denote the observation $\mathcal{X}_{ijk}$ as rating from user i towards music k under the contextual factors j. Due to the fact that not all the music songs can be listened to by a specific user, there will be missing fibers along the contextual factors mode, i.e., $\mathcal{X}_{i,:,k}$ representing the fibers of user i ratings on song k over the 10 different contextual factors. Figure 3.1 (a) shows the observational rating tensor under different contextual factors, where a subset of users are rock music lovers who rate music belonging to rock genre higher than soft music. Similarly, a subset of soft music lovers rate the opposite way. The number of ratings within each 'block' (block represents a small box in Figure 3.1 (a) and for example, the bottom left red block shows the ratings collected from soft music lovers on rock music songs under relaxed driving mode) can be found imbalanced and there is an obvious higher tendency to rate music songs consistent with users' taste preferences. Since the observation in this experiment is an order three tensor, the corresponding propensity score will be a matrix and we denote the propensity matrix as $\mathbf{P}$, describing the marginal probabilities such that $\mathbf{P}_{i,k} = \mathbb{P}(\mathcal{X}_{i,:,k} \text{ is observed})$.

In the simulation experiment, the number of users in rock lover and soft lover group are both set to be 50, and the number of songs belonging to rock music, jazz music and soft music are 30, 30, 30 respectively. Thus, under this setting, the number of fibers within each block will be 1500 in total. Each fiber within a specific block will be determined as observed based on independent Bernoulli distribution model with probability listed in the propensity matrix $\mathbf{P}$ in Figure 3.1 (c). For components in a fiber, the ratings under different contextual factors are specified to be a number as demonstrated in Figure 3.1 (b). For this experiment, $\mathcal{X}$ is completely known so that we can evaluate the performance via MSE in (3.1). We propose the following prediction schemes and evaluate their prediction accuracy

(a)



(b)

	Rock Music		Jazz Music		Soft Music	
	Rock lover	Soft lover	Rock lover	Soft lover	Rock lover	Soft lover
Relaxed Driving	4	0	2	2	2	6
Sports Driving	6	2	2	2	0	4
City	5	1	2	2	5	1
Highway	5	1	2	2	5	1
Awake	6	2	2	2	0	4
Sleepy	4	0	2	2	2	6
Rainy	6	2	2	2	2	6
Sunny	6	2	2	2	2	6
Daytime	4	0	2	2	0	4
Night	4	0	2	2	0	4

(c)

	Rock Music	Jazz Music	Soft Music
Rock lover	0.8	0.5	0.2
Soft lover	0.2	0.5	0.8

Figure 3.1: InCarMusic synthetic dataset: (a) tensor observation for user ratings towards songs under different contextual factors. (b) users ratings within each block under specific contextual factor. (c) Propensity matrix. $\mathbf{P}$

Estimating scheme	True MSE	Naive MSE	IPS MSE
REC_ONES	0.333	0.074(0.0047)	0.343(0.0220)
ROTATE	2.433	2.776(0.0065)	2.4374(0.0198)
SKEWED	1.401	1.2155(0.0063)	1.3901(0.0200)

Table 3.1: Comparison of performance evaluation under different estimating schemes. Numbers in parenthesis represents the standard error based on 20 replications of experiments.

through (3.2) and (3.3).

- REC_ONES: the estimator rating tensor $\hat{\mathcal{X}}$ is identical to the true rating tensor $\mathcal{X}$, except that half of the rating 1 are randomly selected and changed to 5.
- ROTATE: Each estimated rating $\hat{\mathcal{X}}_{i_1,\ldots,i_D}$ is changed to $\mathcal{X}_{i_1,\ldots,i_D}-1$ when $\mathcal{X}_{i_1,\ldots,i_D} > 0$ and $\hat{\mathcal{X}}_{i_1,\ldots,i_D} = 6$ for $\mathcal{X}_{i_1,\ldots,i_D} = 0$.
- SKEWED: Estimated rating $\hat{\mathcal{X}}_{i_1,\ldots,i_D}$ is sampled from a normal distribution with mean $\mathcal{X}_{i_1,\ldots,i_D}$ and standard deviation $(7 - \mathcal{X}_{i_1,\ldots,i_D})/2$.

Table 3.1 shows the corresponding true, naive and IPS MSE calculated for the above three estimators and, as is shown, the IPS MSE matches the true MSE under different estimating schemes while the naive MSE is severely biased from true MSE, indicating the selection bias introduces heavily discrepancy from true MSE when the fibers are missing not at random.

3.3 Propensity-Scored Tensor Completion with Missing Fibers

As proposed in (3.3), the IPS estimator is based on removing the selection bias via the propensity score $\mathcal{P}$, which is determined by the (missing data) mechanism generating the observation pattern in tensor fibers. Prior work in missing data analysis [Little and Rubin, 2019] states the following two mechanisms that decide whether propensity score $\mathcal{P}$ is known or not. First, the experimental setting is a scenario when $\mathcal{P}$ is known. This is widely used in ads related recommendation system, where experimenters are responsible for determining which ads are shown to the users. Second, we have the observational setting where $\mathcal{P}$ is

unknown. This is a common setting when some factors that are not under the control of experimenters are involved in generating the missing data. For example, in InCarMusic dataset [Baltrunas et al., 2011], users will self-select the songs they would like to listen and rate. Similarly, studies of gene expression in multiple tissues may face the challenge of collecting multi-tissue expression data on account of the fact that collection of inaccessible tissues from living subjects is impossible or imoral or certain samples only include limited available tissue biopsies. Thus, to evaluate the performance of completion and estimation, we choose to analyze experimental setting and observational setting separately.

3.3.1 Experimental setting

As stated above, when propensity score $\mathcal{P}$ is known, we define the IPS experimental estimator as

$$\hat{\mathcal{X}}_{\text{Exp}} = \arg\min_{\hat{\mathcal{X}} \in \mathcal{H}_{\text{Exp}}} L_{\text{IPS}}(\hat{\mathcal{X}}|\mathcal{P}), \quad \mathcal{H}_{\text{Exp}} = \{\hat{\mathcal{X}} \in \bigotimes_{i=1}^{D} \mathbb{R}^{d_i} : \text{rank}(\hat{\mathcal{X}}) \leq \mathbf{r}_1, \quad \|\hat{\mathcal{X}}\|_\infty \leq \psi_1\} \tag{3.4}$$

$\hat{\mathcal{X}}_{\text{Exp}}$ is restricted within $\mathcal{H}_{\text{Exp}}$ where a Tucker rank constraint (for convenience of notation, rank$(\cdot)$ represents the Tucker rank for tensor throughout the whole chapter) is imposed and an upper bound ψ_1 is employed on the elementwise maximum norm of tensor components. These two constraints are widely used in regularized tensor estimation literature. Tucker low rank assumes low rankness on tensor and without this assumption, recovering the missing entries will be impossible if the observed entries are independent of each other. Max norm constraint also makes sense in many real life applications. For instance, in most of the recommendation systems, the user ratings towards different items are ordinal values (star 1 to star 5) or even binary values (thumb up/thumb down or like/dislike), where ψ_1 can be set as any number larger than the maximum rating.

The following proposition establishes error bound of empirical risk for $\hat{\mathcal{X}}_{\text{Exp}}$.

Proposition 1. *Suppose* $\mathcal{X} \in \bigotimes_{i=1}^{D} \mathbb{R}^{d_i}$ *is an order* D *tensor observation with missing fibers along the* k*th mode and there exists a constant* $\phi \in (0, \psi_1)$ *such that* $\|\mathcal{X}\|_\infty \leq \phi$. $\mathcal{P} \in \bigotimes_{i \neq k} \mathbb{R}^{d_i}$ *is the corresponding given propensity score, which is an order* $(D-1)$ *tensor with component* $\mathcal{P}_{i_1,\ldots,i_{k-1},i_{k+1},..,i_D}$ *specifying the probability of fiber* $\mathcal{X}_{i_1,\ldots,i_{k-1},i_{k+1},..,i_D}$ *being revealed. For IPS estimator defined in* (3.4)*, with probability at least* $1-\delta$*, we have*

$$|L_{IPS}(\hat{\mathcal{X}}_{Exp}|\mathcal{P}) - L(\hat{\mathcal{X}}_{Exp})| \leq \frac{4\psi_1^2}{\|\mathcal{P}\|_{\min}} \sqrt{\frac{1}{2\prod_{i=1}^{D} d_i} \log \frac{2}{\delta}} \tag{3.5}$$

Remarkably, it's noted from proposition 1 that as long as $\|\mathcal{P}\|_{\min}$ is lower bounded away from 0, empirical risk error bound for $\hat{\mathcal{X}}_{\text{Exp}}$ will converge to 0.

3.3.2 Observational setting

We move on to discuss the observational setting where the propensity score need to be estimated. In general, the propensities can depend on the following: $\mathbf{X}_{\text{obs}}$, $\mathbf{X}_{\text{hid}}$, $\mathcal{X}$ indicating the observable, unobservable features for multiple modes of tenor and tensor observation respectively. One popular approach for propensity score estimation assumes that dependencies between covariates $\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{hid}}$ are negligible and propensity score will be fully estimated based on missingness mask $\mathcal{M} \in \bigotimes_{i=1}^{D} \mathbb{R}^{d_i}$ where components of $\mathcal{M}$ take binary values, i.e., when fiber $\mathcal{X}_{i_1,\ldots,i_{k-1},:,i_{k+1},\ldots,i_D}$ is observed, the corresponding entry $\mathcal{M}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D}$ will take the value 1 with probability $\mathcal{P}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D}$ and otherwise the value 0. Inspired by max norm constrained maximum likelihood estimation procedure for 1-bit matrix completion [Cai and Zhou, 2013, Davenport et al., 2014], we propose to use max-norm rank constrained maximum likelihood estimation procedure to estimate propensity score $\mathcal{P}$. Specifically, a user-specified strictly increasing link function $f : \mathbb{R} \to [0, 1]$ is employed on each entry of a parameter tensor $\mathcal{S} \in \bigotimes_{i=1}^{D} \mathbb{R}^{d_i}$ such that $\mathcal{P}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D} = f(\mathcal{S}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D})$. The observational model can be stated as the dependence of $\mathcal{M}$ on the underlying matrix $\mathcal{S}$

is in the following format:

$$\mathcal{M}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D} = \begin{cases} 1, & \text{if } \mathcal{S}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D} + \mathcal{E}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D} \geq 0 \\ 0, & \text{if } \mathcal{S}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D} + \mathcal{E}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D} < 0 \end{cases} \tag{3.6}$$

where $\mathcal{E}$ is an order $(D-1)$ tensor consisting of i.i.d errors. If we define the link function f connected with distribution function of $\mathcal{E}$ as $f(\theta) = \mathbb{P}(\mathcal{E}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D} \geq -\theta)$, model (3.6) reduces to the setting where $\mathcal{P}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D} = \mathbb{P}(\mathcal{M}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D} = 1) = f(\mathcal{S}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D})$. Thus, different choices of link function f have one-to-one correspondence with distribution assumption on noise tensor $\mathcal{E}$. The common choice for f can be logistic, probit or Laplacian, and their relationship with distribution of $\mathcal{E}$ has already been described in Cai and Zhou [2013], Davenport et al. [2014].

Given the tensor fiber observation $\mathcal{X}_{i_1,\ldots,i_{k-1},:,i_{k+1},\ldots,i_D}, \forall(i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D) \in \mathbb{O}$ and the above observational model, the log likelihood function can be formulated as

$$\begin{aligned} \ell(\mathcal{S}) = \sum_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D} & [\mathbf{1}_{\{\mathcal{M}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D}=1\}} \log f(\mathcal{S}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D}) + \\ & \mathbf{1}_{\{\mathcal{M}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D}\}=0} \log(1 - f(\mathcal{S}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D}))] \end{aligned} \tag{3.7}$$

The max norm rank constrained maximum likelihood estimator for $\mathcal{S}$ is defined to be

$$\hat{\mathcal{S}}_{\text{MLE}} = \arg\max_{\mathcal{S}\in\mathcal{J}} \ell(\mathcal{S}), \quad \mathcal{J} = \{\mathcal{S} : \text{rank}(\mathcal{S}) \leq \mathbf{r}_2, \|\mathcal{S}\|_\infty \leq \psi_2\} \tag{3.8}$$

Similarly, the optimization procedure requires $\mathcal{S}$ has Tucker low rank, which is an effective dimension reduction tool in tensor data analysis and max norm constrain ψ_2, which is expected to be reasonably effective for uniformly bounded data. It can be easily found that when $\mathcal{X}$ is an order three tensor observation, $\mathcal{S}$ reduces to an order two tensor, i.e., matrix, and $\hat{\mathcal{S}}_{\text{MLE}}$ will be 1-bit matrix completion estimator with full observation when changing

tucker rank constraint as nuclear norm constraint. Thus, our proposed model successfully generalize the matrix theory to higher order tensor.

To establish the error bound of estimator $\hat{\mathcal{S}}_{\text{MLE}}$ and $\hat{\mathcal{P}} = f(\hat{\mathcal{S}}_{\text{MLE}})$, we introduce the following assumption on f which controls the 'steepness' and 'convexity' of the link function f.

Assumption 5. *Suppose the link function f is monotonic increasing and differentiable. Let*

$$L_{\psi_2} = \sup_{|s|\leq\psi_2} \Big\{ \frac{f'(s)}{f(s)(1-f(s))} \Big\}, \quad U_{\psi_2} = \inf_{|s|\leq\psi_2} \Big\{ \frac{(f'(s))^2}{f^2(s)} - \frac{f''(s)}{f(s)} \Big\}$$

be finite, where f' and f'' represent the first order and second order derivative of f with respect to s respectively.

Now we are ready to state the result related to the recovery of $\mathcal{P}$ using max-norm rank constrained maximum likelihood estimator.

Theorem 3. *Assume Assumption 5 holds and let C such that $C = 8\log 12 + 9\log D$, with probability at least $1 - \frac{1}{\sum_{i\neq k} d_i} - 2\exp\left(-\log(D-1)\sum_{i\neq k} d_i\right)$, we have*

$$\frac{1}{\prod_{i\neq k} d_i} \|\hat{\mathcal{P}} - \mathcal{P}\|_F^2 \leq \min\Big\{4, 8eCL_{\psi_2}\psi_2 \sqrt{r_{2,\max}^{D-2} \frac{\sum_{i\neq k} d_i}{\prod_{i\neq k} d_i}}\Big\} \tag{3.9}$$

Theorem 3 provides a finite sample error bound on the propensity score estimate based on max norm rank regularized maximum likelihood estimation method. As shown in Ma and Chen [2019], the convergence rate for propensity score estimate via 1-bit matrix completion method is $O(\frac{1}{\sqrt{d_1}} + \frac{1}{\sqrt{d_2}})$. One the one hand, if we assume the maximum Tucker rank $r_{2,\max}$ to be a constant and will not increase as the number of modes D and dimension d_i change, the convergence rate of theorem 3 in order two tensor (i.e. matrix case) is $O(\sqrt{\frac{1}{d_1} + \frac{1}{d_2}})$, which improves the result in matrix case where low rank constraint is employed via regularization on nuclear norm of matrix. On the other hand, our theoretical bound also allows the Tucker

rank in each mode to diverge, which reflects a typical large scale scenario in big tensor data. A notable consequence of theorem 3 is, when all d_i are of the same order as d, a sufficient condition for consistency is that $d \to \infty$ and $r_{2,\max} = o(d)$. In particular, when $D = 3$, our proposed propensity score estimator is consistent as long as the maximum Tucker rank along each mode diverges slightly slower than d. Another interesting outcome is as we have more modes in the tensor data, i.e., D is getting larger and larger, the constraint on the rate of diverge of maximum element of Tucker rank, $r_{2,\max}$, does not become stricter, which allows our algorithm to perform stably even under extreme higher order case.

Similar to IPS estimator under experimental setting, we can define the IPS observational estimator as

$$\hat{\mathcal{X}}_{\text{Obs}} = \arg \min_{\hat{\mathcal{X}} \in \mathcal{H}_{\text{Obs}}} L_{\text{IPS}}(\hat{\mathcal{X}}|\hat{\mathcal{P}}), \quad \mathcal{H}_{\text{Obs}} = \{\hat{\mathcal{X}} \in \bigotimes_{i=1}^{D} \mathbb{R}^{d_i} : \text{rank}(\hat{\mathcal{X}}) \leq \mathbf{r}_1', \quad \|\hat{\mathcal{X}}\|_\infty \leq \psi_1'\} \tag{3.10}$$

where $\hat{\mathcal{P}}$ is the propensity score estimator via max norm rank constrained maximum likelihood approach. We can derive the error bound for empirical risk of $\hat{\mathcal{X}}_{\text{Obs}}$ in the following theorem.

Theorem 4. *Suppose $\mathcal{X} \in \bigotimes_{i=1}^{D} \mathbb{R}^{d_i}$ is an order D tensor with missing fibers along the kth mode and there exists a constant $\phi \in (0, \psi_1')$ such that $\|\mathcal{X}\|_\infty \leq \phi$, $\hat{\mathcal{P}} \in \bigotimes_{i \neq k} \mathbb{R}^{d_i}$ is the corresponding estimated propensity score in theorem 3, which is an order $(D-1)$ tensor with component $\hat{\mathcal{P}}_{i_1,\ldots,i_{k-1},i_{k+1},..,i_D}$ as the estimated probability of fiber $\mathcal{X}_{i_1,\ldots,i_{k-1},:,i_{k+1},..,i_D}$ being revealed. Then, under observational setting, for IPS estimator defined in* (3.10)*, with probability at least $1 - \delta - \frac{1}{\sum_{i \neq k} d_i} - 2\exp\left(-\log(D-1)\sum_{i \neq k} d_i\right)$, we have*

$$|L_{IPS}(\hat{\mathcal{X}}_{Obs}|\hat{\mathcal{P}}) - L(\hat{\mathcal{X}}_{Obs})| \leq \frac{4\psi_1'^2}{f(-\psi_2)} \sqrt{\frac{1}{2\prod_{i=1}^{D} d_i} \log \frac{2}{\delta}} + \min\left\{ \frac{16\psi_1'^2}{f(-\psi_2)^2}, \frac{32eCL_{\psi_2}\psi_1'^2\psi_2}{f(-\psi_2)^2} \sqrt{r_{2,\max}^{D-2} \frac{\sum_{i \neq k} d_i}{\prod_{i \neq k} d_i}} \right\}$$

where C is a constant satisfying $C = 8\log 12 + 9\log D$.

The above bound exhibits a bias-variance trade-off that does not arise in Proposition 1

for the experimental setting. The first term in error bound is similar to the bound derived in Proposition 1, quantifying how the accuracy of IPS estimator changes as the propensities become more 'non-uniform' while second term comes from theorem 3, i.e., the error bound for propensity score estimation.

3.4 Recovery Error for Tensor Completion with Missing Fibers

In many real life applications, the propensity score $\mathcal{P}$ is unknown and needs to be estimated delicately. We will focus on the observational setting in this section. To derive the statistical recovery error bound for $\hat{\mathcal{X}}_{\text{Obs}}$, we restricted our focus on the noise tensor completion setting. Given an order D tensor observation $\mathcal{X} \in \bigotimes_{i=1}^{D} \mathbb{R}^{d_i}$ with partially revealed fibers along kth mode, we assume it comes from true tensor $\mathcal{X}^*$ perturbed by some error tensor $\mathcal{E}$, i.e., $\mathcal{X} = \mathcal{X}^* + \mathcal{E}$. $\mathcal{X}^*$ enjoys low rankness with Tucker rank as $\mathbf{r}$ and entries in $\mathcal{E}$ are i.i.d Gaussian distributed with mean 0 and variance σ^2. Furthermore, the observed fibers along kth mode are assumed to follow an i.i.d Bernoulli model, that is $\mathcal{X}^*_{i_1,\ldots,i_{k-1},:,i_{k+1},\ldots,i_D}$ is observed independently with unknown probability $\mathcal{P}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D}$.

Compared with other popular methods in tensor completion, we can find that the one major aspect that differentiates our method lies in the assumed sampling scheme. For the matrix completion theory, Davenport et al. [2014] derived results on bounds of recovery error for both 1-bit measurements and unquantized measurements under uniform sampling distribution scheme, where each entry is discovered independently with equal probability. Cai and Zhou [2013] generalized uniform sampling distribution to a random sampling distribution Π with $\sum_{i,j} \pi_{i,j} = 1$ and the observation index set is drawn with replacement according to Π. Ghadermarzy et al. [2019] successfully extended the theory from matrix to tensor under general sampling scheme as in Cai and Zhou [2013] by proposing a max-qnorm constrained least squares tensor completion method. It is also worth noting that uniform sampling schemes are still popular in many papers related to low rank tensor completion from noisy

observation [Xia and Yuan, 2017, Xia et al., 2017]. Thus, we can conclude that excluding the obvious fact that our focus is random missing fibers instead of random missing entries, we further assume the reveal of each fiber follows an independent Bernoulli model with different probability determined by propensity score $\mathcal{P}$.

To investigate the asymptotic property of $\hat{\mathcal{X}}_{\text{Obs}}$, we restate the following assumptions on the missing structure,

Assumption 6. *The missingness mask tensor $\mathcal{M}$ has mutually independent entries and are also independent of error tensor $\mathcal{E}$. $\mathcal{M}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D}$ follows a Bernoulli distribution with probability of success $\mathcal{P}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D} = f(\mathcal{S}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D})$. Moreover, there exists a lower bound $\mathcal{P}_L \in (0,1)$ (which is allowed to depend on $d_1,\ldots,d_D$) such that $\min_{i_1,\ldots i_D} \mathcal{P}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D} \geq \mathcal{P}_L \geq \frac{1}{\mu \prod_{i=1}^D d_i}$ for some positive constant $\mu > 1$.*

Assumption 7. *Each entry in perturbed error $\mathcal{E}$ are independently distributed random variables such that $\mathbb{E}(\mathcal{E}_{i_1,\ldots,i_D}) = 0$ and $\mathbb{E}(\mathcal{E}^2_{i_1,\ldots,i_D}) = \sigma^2_{i_1,\ldots,i_D} \leq \infty$. Moreover, for some finite positive constant c_σ and η, $\max_{i_1,\ldots i_D} \mathbb{E}|\mathcal{E}_{i_1,\ldots,i_D}|^l \leq \frac{l!}{2} c_\sigma^2 \eta^{l-2}$ for any positive integer $l \geq 2$.*

Assumption 8. *There exists a constant $\phi \in (0, \psi_1')$ such that $\|\mathcal{X}^*\|_\infty \leq \phi$.*

Denote

$$\Gamma \asymp \max\left\{ \frac{d_{\max} \log(\sum_{i=1}^D d_i)}{\mathcal{P}_L \prod_{i=1}^D d_i}, \frac{(\log(\sum_{i=1}^D d_i))^2}{\mathcal{P}_L^2 \prod_{i=1}^D d_i} \right\} + r_{2,\max}^{D-1} \frac{\sum_{i=1}^D d_i (\log(\sum_{i=1}^D d_i))^{1/2}}{\prod_{i=1}^D d_i}$$

where $d_{\max} = \max\{d_1,\ldots,d_D\}$ and

$$\delta_{d_1,\ldots,d_D}(c_\sigma,\eta) = 2/\sum_{i=1}^D d_i + \exp\{-c\sum_{i=1}^D d_i\} + 12c_\sigma^2\eta^2/\log(\sum_{i=1}^D d_i)$$

Now we state the main result on the performance of proposed IPS estimator with estimated propensity score for recovering a bounded low rank tensor.

Theorem 5. *Assume Assumptions 5,6,7 and 8 hold. There exists constant C that only depends on D, and with probability at least $1-\delta_{d_1,...,d_D}(c_\sigma,\eta)$, we have*

$$\frac{1}{\prod_{i=1}^{D} d_i}\|\hat{\mathcal{X}}_{Obs}-\mathcal{X}^*\|_F^2 \le 4\mu^2(r'_{1,\max})^{D-1}\Big(\Gamma + C\frac{d_{\max}\log(\sum_{i=1}^{D} d_i)}{|\mathbb{O}|d_k}\Big) \tag{3.11}$$

The diminishing $\delta_{d_1,...,d_D}(c_\sigma,\eta)$ means that error bound for $\hat{\mathcal{X}}_{\text{Obs}}$ is bounded by the right hand side of (3.11) with probability approaching 1 for large enough $d_1,...,d_D$. Taking a look at the constitution of Γ, we can find part of Γ is related to the estimation performance of propensity score, which implies that accurate estimation of propensity score is a prerequisite for a successful recovery of underlying true tensor $\mathcal{X}^*$. We can tell easily from Theorem 5, as long as $d_{\max}\log(\sum_{i=1}^{D} d_i)/(|\mathbb{O}|d_k) \to 0$ as $d_i \to 0$, $\frac{1}{\prod_{i=1}^{D} d_i}\|\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*\|_F^2 \to 0$, or in other words, $\hat{\mathcal{X}}_{\text{Obs}}$ is a consistent estimator for $\mathcal{X}^*$. If we further assume $\{d_1,...,d_D\} = O(d)$, $\Gamma \asymp \mathcal{P}_L^{-1}d^{1-D}\log d + r_{2,\max}^{(D-1)/2}\mathcal{P}_L^{-1/2}d^{1-D}\sqrt{\log d} + r_{2,\max}^{D-1}d^{1-D}$. Besides, we know the expected observed sample size $\mathbb{E}(|\mathbb{O}|d_k)$ is at least $\mathcal{P}_L\prod_{i=1}^{D} d_i$, which leads to $d_{\max}\log(\sum_{i=1}^{D} d_i/(|\mathbb{O}|d_k) \asymp \mathcal{P}_L^{-1}d^{1-D}\log d$. In summary, $\|\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*\|_F^2/\prod_{i=1}^{D} d_i$ are $O(\mathcal{P}_L^{-1}d^{1-D}\text{polylog}(d))$. This rate attained is noted to coincide with that of the other tensor completion methods, for instance, Ghadermarzy et al. [2019] calibrating max qnorm constrained estimator, Xia et al. [2017]'s minimax optimal low rank tensor estimator in a general ℓ_p loss and Lee and Wang [2020]'s low rank constrained ordinal tensor completion, under either entries uniform sampled at random or drawn randomly with replacement from a predefined probability distribution Π.

3.5 Algorithms

Recall that under observational setting, we need to estimate the underlying parameter tensor $\mathcal{S}$ using the max norm rank constrained maximum likelihood approach proposed in (3.8) and then apply a link function f on $\hat{\mathcal{S}}_{\text{MLE}}$ to get the final propensity score estimate $\hat{\mathcal{P}}$. It is

not difficult to figure out that (3.8) is a non-convex problem due to the non-convexity in the feasible set $\mathcal{J}$. To solve this, we can use the Tucker decomposition representation and change the optimization problem into a block-wise convex problem. Let's try to write out the Tucker decomposition of $\mathcal{S}$ as $\mathcal{C} \times_1 \mathbf{S}_1 \times_2 \mathbf{S}_2 .... \times_{k-1} \mathbf{S}_{k-1} \times_{k+1} \mathbf{S}_{k+1} ... \times_D \mathbf{S}_D$, where $\mathcal{C} \in \mathbb{R}^{r_1 \times r_2 ... \times r_{k-1} \times r_{k+1} ... \times r_D}$ is the core tensor and $\mathbf{S}_i \in \mathbb{R}^{d_i \times r_i}, \forall i \in \{1, 2, ..., k-1, k+1, ..., D\}$ are the corresponding factor matrices for each mode. By introducing this low rank decomposition, log likelihood function defined in (3.7) can be treated as a function of core tensor and factor matrices, i.e., $\ell(\mathcal{S}) = \ell(\mathcal{C}, \mathbf{S}_1, ..., \mathbf{S}_D)$. Since we will implement a block-wise convex optimization algorithm to update $\mathcal{C}, \mathbf{S}_i, \forall i \in \{1, 2, ..., k-1, k+1, ..., D\}$ iteratively, for the purpose of notational convenience, when updating one block and keeping other blocks fixed, we will write $\ell(\mathcal{C}, \mathbf{S}_1, ..., \mathbf{S}_D)$ as only a function of the block unfixed. For example, when updating $\mathcal{C}$, $\ell(\mathcal{C}, \mathbf{S}_1, ..., \mathbf{S}_D)$ will be simplified as $\ell(\mathcal{C})$.

Then we can update the core tensor $\mathcal{C}$ and factor matrices $\mathbf{S}_i$ block-by-block. Specifically, the update for each factor matrix involves solving a number of separate Generalized Linear Models (GLM). We take updating the jth factor matrix at tth iteration as an example to illustrate how we determine the responses and predictors for GLM. Denote $\hat{\mathbf{S}}_j^{(t)}$ as the jth factor matrix estimate at tth iteration and for notational simplicity, we define

$$\hat{\mathbf{S}}_{(-j)} \in \mathbb{R}^{(\prod_{i \neq j,k} d_i) \times r_j} = \left[\hat{\mathcal{C}}^{(t)} \times_1 \hat{\mathbf{S}}_1^{(t+1)} ... \times_{j-1} \hat{\mathbf{S}}_{j-1}^{(t+1)} \times_{j+1} \hat{\mathbf{S}}_{j+1}^{(t)} ... \times_{k-1} \hat{\mathbf{S}}_{k-1}^{(t)} \times_{k+1} \hat{\mathbf{S}}_{k+1}^{(t)} ... \times_D \hat{\mathbf{S}}_D^{(t)} \right]_{(j)}^\top$$

Then, $\hat{\mathbf{S}}_j^{(t+1)}$ can be estimated row-by-row through solving the following d_j separate GLMs

$$\text{vec}(\mathcal{M}_{:,...,:,j_l,:,...,:}) \sim \text{GLM}(\hat{\mathbf{S}}_{(-j)}), \quad \forall l \in \{1, 2, ..., d_j\} \tag{3.12}$$

where the derived coefficient estimates constitute the lth row of $\hat{\mathbf{S}}_j^{(t+1)}$. Obviously, these low-dimensional GLMs facilitate the use of a fast GLM solver as well as parallel processing to speed up computation. Similarly, the core tensor $\mathcal{C}$ can also be updated via solving GLM

as follows

$$\text{vec}(\mathcal{M}) \sim \text{GLM}(\hat{\mathbf{S}}_1^{(t+1)} ... \odot \hat{\mathbf{S}}_{k-1}^{(t+1)} \odot \hat{\mathbf{S}}_{k+1}^{(t+1)} ... \odot \hat{\mathbf{S}}_D^{(t+1)}) \tag{3.13}$$

where the derived coefficient estimates composes the estimate for $\text{vec}(\hat{\mathcal{C}}^{(t+1)})$ and to get $\hat{\mathcal{C}}^{(t+1)}$, we only need to transfer it back to tensor format. The max norm constraint $\|\mathcal{S}\|_\infty \leq \psi_2$ can be imposed while performing the line search for the final estimate of $\hat{\mathcal{C}}^{(t+1)}$.

To summarize, the whole algorithm for estimating the propensity score $\mathcal{P}$ can be stated in Algorithm 3. Analogously, observational IPS estimator defined in (3.10) can be solved

Algorithm 3: Propensity score estimate via max norm rank constrained MLE

Input: Missingness mask $\mathcal{M} \in \bigotimes_{i \neq k} \mathbb{R}^{d_i}$, tensor tucker rank $\mathbf{r}_2 \in \mathbb{R}^{D-1}$, elementwise bound ψ_2 and link function f

Output: Estimated propensity score $\hat{\mathcal{P}}$

Initialize core tensor $\hat{\mathcal{C}}^{(0)}$ and factor matrices $\hat{\mathbf{S}}_i^{(0)}, \forall i \in \{1, 2, ..., k-1, k+1, ..., D\}$ and iteration index $t = 1$;

while *the relative increase in objective function* $\ell(\hat{\mathcal{S}} = \hat{\mathcal{C}}^{(0)} \times_1 \hat{\mathbf{S}}_1^{(0)} ... \times_D \hat{\mathbf{S}}_D^{(0)})$ *is larger than tolerance level* **do**

- Get $\hat{\mathbf{S}}_i^{(t)}$ via solving separate d_i GLMs defined in (3.12);
- Get $\hat{\mathcal{C}}^{(t)}$ via solving GLMs defined in (3.13);
- Perform line search to get $\rho^* = \arg\max_{\rho \in [0,1]} \ell(\rho \hat{\mathcal{C}}^{(t-1)} + (1-\rho)\hat{\mathcal{C}}^{(t)})$, subject to $\|\mathcal{S}\|_\infty \leq \psi_2$ and set $\hat{\mathcal{C}}^{(t)} = \rho^* \hat{\mathcal{C}}^{(t-1)} + (1-\rho^*)\hat{\mathcal{C}}^{(t)}$;
- Update iteration index $t \leftarrow t + 1$;

end

Get $\hat{\mathcal{S}}_{\text{MLE}} \leftarrow \hat{\mathcal{C}} \times_1 \hat{\mathbf{S}}_1 \times_2 \hat{\mathbf{S}}_2 .... \times_{k-1} \hat{\mathbf{S}}_{k-1} \times_{k+1} \hat{\mathbf{S}}_{k+1} ... \times_D \hat{\mathbf{S}}_D$;

Get $\hat{\mathcal{P}}_{i_1,...,i_{k-1},i_{k+1},...,i_D} \leftarrow f(\hat{\mathcal{S}}_{\text{MLE},i_1,...,i_{k-1},i_{k+1},...,i_D})$

using the same block-wise optimization framework. Denote the tucker decomposition of $\mathcal{X}^*$ as $\mathcal{D} \times_1 \mathbf{X}_1 ... \times_D \mathbf{X}_D$ where $\mathcal{D}$ and $\mathbf{X}_1, ..., \mathbf{X}_D$ are the corresponding order D core tensor and factor matrices along each mode. For each block update, our aim is to solve a constrained optimization function, which minimizes $L_{\text{IPS}}(\hat{\mathcal{X}}|\hat{\mathcal{P}})$ under the constraint $\|\hat{\mathcal{X}}\|_\infty \leq \psi_1'$. The algorithm has been illustrated in Algorithm 4.

Algorithm 4: Imputation for tensor missing fibers via estimated propensity score

Input: Tensor with partially observed fibers $\mathcal{X} \in \bigotimes_{i=1}^{D} \mathbb{R}^{d_i}$, tensor tucker rank $\mathbf{r}_1' \in \mathbb{R}^D$, elementwise bound ψ_1' and propensity score estimate $\hat{\mathcal{P}}$

Output: Estimated tensor with full observations $\hat{\mathcal{X}}_{\text{Obs}}$

Initialize core tensor $\hat{\mathcal{D}}^{(0)}$ and factor matrices $\hat{\mathbf{X}}_i^{(0)}, \forall i \in \{1, 2, ..., D\}$ and iteration index $t = 1$;

while *the relative increase in objective function* $L_{IPS}(\hat{\mathcal{X}}|\hat{\mathcal{P}})$ *is larger than tolerance level* **do**

$\hat{\mathbf{X}}_i^{(t)} \leftarrow \arg\min_{\mathbf{X}_i} L_{\text{IPS}}(\hat{\mathcal{D}}^{(t-1)} \times_1 \hat{\mathbf{X}}_1^{(t)} ... \times_{i-1} \hat{\mathbf{X}}_{i-1}^{(t)} \times_i \mathbf{X}_i \times_{i+1} \hat{\mathbf{X}}_{i+1}^{(t-1)} ... \times_D \hat{\mathbf{X}}_D^{(t-1)}), \text{s.t.} \|\hat{\mathcal{X}}\|_\infty \leq \psi_1', \forall i \in \{1, 2, ..., D\}$;

$\hat{\mathcal{D}}^{(t)} \leftarrow \arg\min_{\mathcal{D}} L_{\text{IPS}}(\hat{\mathcal{D}}^{(t-1)} \times_1 \hat{\mathbf{X}}_1^{(t)} ... \times_{i-1} \hat{\mathbf{X}}_{i-1}^{(t)} \times_{i+1} \hat{\mathbf{X}}_{i+1}^{(t-1)} ... \times_D \hat{\mathbf{X}}_D^{(t-1)}), \text{s.t.} \|\hat{\mathcal{X}}\|_\infty \leq \psi_1'$;

Update iteration index $t \leftarrow t + 1$;

end

Get $\hat{\mathcal{X}}_{\text{Obs}} \leftarrow \hat{\mathcal{D}} \times_1 \hat{\mathbf{X}}_1 ... \times_D \hat{\mathbf{X}}_D$;

In practice, the rank $\mathbf{r}_2, \mathbf{r}_1'$ are hardly known or given. One appropriate method is to manually inspect a sequence of underlying tensor estimation for a range of all possible combinations of hyperparameters and use the domain knowledge related to a specific application to achieve a better imputation performance. However, this approach is time-consuming and sometimes may require expert knowledge. Thus, we may prefer an automated and data-driven approach. Cross validation are popular techniques for tuning parameter selection but may be unreasonable in higher order tensor settings due to their computational burden. In this chapter, we utilize a commonly implemented criteria, extended Bayesian Information Criterion (eBIC) proposed in Chen and Chen [2008, 2012], which is defined in genral as

$$\text{eBIC}(\mathbf{r}) = -n \log \frac{\text{Likelihood}}{n} + 2\text{df}_{\mathbf{r}} \log(n)$$

where n is the number of observations, $\text{df}_{\mathbf{r}} = \sum_i (d_i - r_i) r_i + \prod_i r_i$ is the degrees of freedom for a particular value of $\mathbf{r}$. In particular, Likelihood in eBIC should take $\ell(\hat{\mathcal{S}})$ in Algorithm 3

and $L_{\text{IPS}}(\hat{\mathcal{X}}|\hat{\mathcal{P}})$ in Algorithm 4. This criterion successfully balances between goodness-of-fit for data and degrees of freedom for model. By choosing a grid of candidate values for $\mathbf{r}_1', \mathbf{r}_2$, the optimal ranks are selected which corresponds to the smallest value of eBIC over all the candidate values.

3.6 Experiments

3.6.1 Synthetic Data Analysis

In this section, we report the simulation results for recovery performance of our proposed propensity score estimator $\hat{\mathcal{P}}$ and observational IPS estimator $\hat{\mathcal{X}}_{\text{Obs}}$. Results come from conduction of Algorithms 3 and 4 and for simplicity, we consider an order four tensor observation $\mathcal{X} \in \mathbb{R}^{d\times d\times d\times d}$ whose low rank factors and core tensor are generated based on the standard Gaussian distribution. However, as mentioned in section 3.5, it is easy to generalize to higher order situations, i.e., $D > 4$. The fibers along the fourth mode are assumed to be missing not at random, where the probability of a specific fiber being observed is determined by order three propensity score $\mathcal{P} \in \mathbb{R}^{d\times d\times d}$. Moreover, mean squared error (MSE) is employed as criterion to measure the performance of estimation for propensity score and true tensor. The performance is measured by average value of MSE on a natural logarithm scale via 20 replications of each simulation setting. To illustrate the variation in simulation results, the error bars are provided which is calculated based on one standard deviation of logMSE.

First, we analyze the finite sample performance of Algorithm 3 and check its consistency with our derived theoretical property statement in Theorem 3. We consider a low rank tensor $\mathcal{S} = \mathcal{C} \times_1 \mathbf{S}_1 \times_2 \mathbf{S}_2 \times_3 \mathbf{S}_3$ with Tucker low rank $\mathbf{r} = (r, r, r)$, where $\mathcal{C} \in \mathbb{R}^{r\times r\times r}$ is the core tensor with entries sampling from i.i.d uniform distribution $[0, \psi]$. $\mathbf{S}_i \in \mathbb{R}^{d\times r}, \forall i \in \{1, 2, 3\}$ are the corresponding factor matrices with respect to each mode and we sample

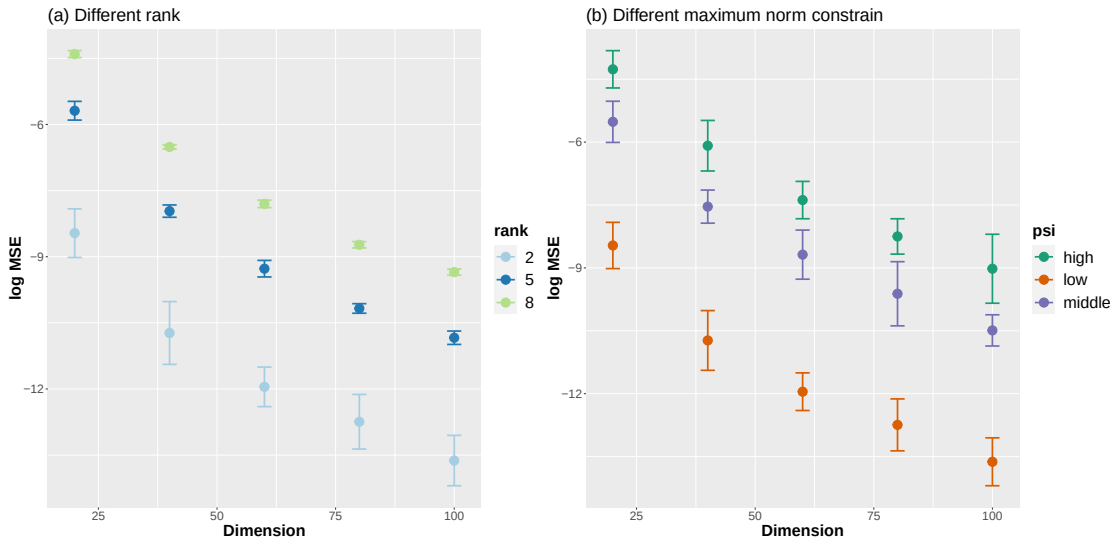


Figure 3.2: $\log \text{MSE} = \log(\|\hat{\mathcal{P}} - \mathcal{P}\|_F^2/d^3)$ of propensity score estimate via Algorithm 3 under different choices of dimension d, Tucker rank r and max norm ψ. (a) Performance of experiments with various $d \in \{20, 40, 60, 80, 100\}$ and $r \in \{2, 5, 8\}$ by setting $\psi = 1$. (b) Performance of experiments with various $d \in \{20, 40, 60, 80, 100\}$ and $\psi = \{1(\text{low}), 5(\text{middle}), 10(\text{high})\}$ by setting $r = 2$.

their components independently from Uniform $[0, 1]$ distribution. Logit link is implemented on $\mathcal{S}$ to obtain the true underlying propensity score $\mathcal{P}$. We vary $d \in \{20, 40, 60, 80, 100\}$, $r \in \{2, 5, 8, 10\}$ and $\psi \in \{1, 5, 10\}$. Table 3.2 shows that MSE decays exponentially as the dimension d increases. Theoretically, large sample sizes provide more information, leading to more accurate estimation of propensity score. However, another important finding is that a larger rank results in increasing MSE, which is due to more complicated underlying latent structure of parameter tensor $\mathcal{S}$ as rank increases. Moreover, large ψ choice means a large maximum element in parameter tensor $\mathcal{S}$, thus generating large estimation error. All these conclusions are consistent with the theoretical property derived in Theorem 3.

Next, we compare the performance of our proposed propensity score estimate based on max norm rank constrained maximum likelihood approach with logistic regression (LogisticReg). As introduced in Schnabel et al. [2016], there are two popular approaches which can be implemented to estimate propensity score: one uses naive Bayes and the other logistic

regression. The naive Bayes approach utilizes the Bayes formula to estimate the propensity score. However, it is commonly used in recommendation system where the tensor observation are finite integers and moreover, it requires a small sample of missing completely at random (MCAR) data to estimate the marginal probability. In general, naive Bayes is inappropriate for many applications that are not recommendation systems and MCAR samples are inaccessible. Thus, we only compare the performance of logistic regression with our proposed method. The main modelling assumption of logistic regression is that there exists parameters w, β, γ, τ such that $\mathcal{P}_{i_1,i_2,i_3} = \sigma(w^\top \mathbf{s}_{i_1,i_2,i_3} + \beta_{i_1} + \gamma_{i_2} + \tau_{i_3})$, where $\mathbf{s}_{i_1,i_2,i_3}$ represents a vector including related observable features of item i_1, i_2, i_3 and $\beta_{i_1}, \gamma_{i_2}, \tau_{i_3}$ are the corresponding offsets parameters along three modes. We compare those two estimation methods under two model settings, inductive model and logistic model.

For the first simulation setting, we formulate the parameter tensor $\mathcal{S}$ in an inductive way, $\mathcal{S} = \mathcal{C} \times_1 \mathbf{S}_1 \times_2 \mathbf{S}_2 \times_3 \mathbf{S}_3$. $\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3$ will be explained as the feature/side information matrices along each mode. This model is referred as inductive tensor completion [Nimishakavi et al., 2018], which is a natural generalization from inductive matrix completion [Jain and Dhillon, 2013, Zhang and Chen, 2019, Zhong et al., 2019, Natarajan and Dhillon, 2014, Chiang et al., 2015] by incorporating the features along each mode into latent structure of tensor. Thus, we treat $\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3$ as feature matrices and the corresponding entry in the mask tensor $\mathcal{M}$ as responses to fit linear regression model. The dimension will vary as $d \in \{20, 40, 60, 80, 100\}$ and we set $r = 5$. As is shown in Table 3.2, under inductive tensor completion model setting, our proposed propensity score estimation performance is much better than logistic regression even though logistic regression model inputs the feature matrices into model fitting. This is on the account of inductive tensor completion model sharing the same underlying Tucker decomposition structure as our model assumption on parameter tensor $\mathcal{S}$. Besides, both methods enjoy a decreasing MSE, indicating better estimation performance when dimension d gets large, which matches a theoretical property of our model and logistic regression.

Dimension	Our Method	LogisticReg
$d = 20$	0.00238(0.000688)	0.00212(0.001808)
$d = 40$	0.00046(0.0001727)	0.00032(0.000234)
$d = 60$	0.00019(6.827e-5)	0.00025(0.000238)
$d = 80$	8.2728e-5(2.8922e-5)	0.00016(0.000117)
$d = 100$	4.9808e-5(2.6371e-5)	0.00019(0.000152)

Table 3.2: Inductive model: comparison of MSE $= \|\hat{\mathcal{P}} - \mathcal{P}\|_F^2/d^3$ for propensity score estimate under different choice of dimension d (Tucker rank is set to be $(5, 5, 5)$).

The second setting follows a logistic regression framework. Three feature matrices $\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3 \in \mathbb{R}^{d\times r}$ are generated with each entry i.i.d from uniform $[0, 1]$ distribution. The weights $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3 \in \mathbb{R}^r$ are sampled i.i.d from Uniform $[0, 1]$ distribution as well. The propensity score $\mathcal{P} \in \mathbb{R}^{d_1\times d_2\times d_3}$ are formulated by setting each component $\mathcal{P}_{i_1,i_2,i_3}$ as $f(\mathbf{S}_1[i_1]^\top \mathbf{w}_1 + \mathbf{S}_2[i_2]^\top \mathbf{w}_2 + \mathbf{S}_3[i_3]^\top \mathbf{w}_3)$ where $\mathbf{S}_k[i_k], \forall k \in \{1, 2, 3\}$ represents the i_kth row of feature matrix $\mathbf{S}_k$. We choose link function f as standard logit link. Frequently, feature matrices might only exist for a subset of modes. Taking this challenge into account, we include the performance of logistic regression with only one or two feature matrices available, i.e., $\{\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3\}$ or $\{\mathbf{S}_1\&\mathbf{S}_2, \mathbf{S}_1\&\mathbf{S}_3, \mathbf{S}_2\&\mathbf{S}_3\}$ fitted into model to mimic the situation when some matrices are inaccessible. From data generation, we can tell there is no big difference in the role of $\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3$. Thus, we only include the logistic regression performance with only $\mathbf{S}_1$ and $\mathbf{S}_1\&\mathbf{S}_2$ for comparison. Analogously, we set $d \in \{20, 40, 60, 80, 100\}$ and $r = 5$, the estimation performance comparison can be found in Table 3.3. Not surprisingly, our proposed method does not perform as well as logistic regression model. However, we find that the performance of those two methods is comparable, especially when the dimension d is large. Thus, it is rational to use max norm rank constrained maximum likelihood propensity score estimate as a substitute for logistic regression estimates, particularly when the features along all modes or some specific modes are inaccessible.

After demonstrating that we can get a reliable estimate for propensity score, we move on to evaluate the performance of our proposed IPS estimator on imputing tensor missing

Dimension	Our Method	LogisticReg
$d = 20$	0.11386(0.039128)	0.01368(0.01027)
$d = 40$	0.10651(0.024352)	0.05868(0.018871)
$d = 60$	0.10449(0.037411)	0.08008(0.038011)
$d = 80$	0.12823(0.019296)	0.10881(0.021841)
$d = 100$	0.13752(0.013853)	0.11832(0.078435)
Dimension	LogisticReg ($\mathbf{S}_1$)	LogisticReg ($\mathbf{S}_1$, $\mathbf{S}_2$)
$d = 20$	0.06610(0.046743)	0.04322(0.024425)
$d = 40$	0.12469(0.039124)	0.08572(0.029572)
$d = 60$	0.14707(0.065591)	0.12082(0.047563)
$d = 80$	0.16076(0.015316)	0.12573(0.034654)
$d = 100$	0.21044(0.042412)	0.15757(0.035254)

Table 3.3: Logistic model: comparison of MSE $= \|\hat{\mathcal{P}} - \mathcal{P}\|_F^2/d^3$ for propensity score estimate under different choice of dimension d (Tucker rank is set to be $(5, 5, 5)$).

fibers. As introduced early, we set the observation as an order four tensor with fibers along the fourth mode missing not at random. The generation of propensity score $\mathcal{P}$ for fibers along the fourth mode is aligned with experiments performed in Figure 3.2. Similarly, true tensor is formulated via a Tucker rank decomposition where $\mathcal{X}^* \in \mathbb{R}^{d\times d\times d\times d} = \mathcal{D} \times_1 \mathbf{X}_1 \times_2 \mathbf{X}_2 \times_3 \mathbf{X}_3 \times_4 \mathbf{X}_4$, where $\mathcal{D} \in \mathbb{R}^{r\times r\times r\times r}$ is order four core tensor with entries i.i.d sampled from uniform $[0, \psi]$ distribution and $\mathbf{X}_i \in \mathbb{R}^{d_i\times r_i}$ are factor matrices with entries i.i.d sampled from uniform $[0, 1]$ distribution as well. A noise tensor $\mathcal{E}$ is added to $\mathcal{X}^*$ as a perturbation with each entry i.i.d sampled from Gaussian distribution with mean 0 and variance σ^2. We report the logMSE of observational IPS estimator via varying $d \in \{10, 20, 30, 40, 50\}$, $r \in \{2, 5, 8\}$, $\sigma \in \{0, 0.1, 0.5, 1\}$ and $\psi \in \{1, 5, 10\}$. The results are summarized in Figure 3.3. Clearly, MSE is getting lower in an exponential trend as the observed sample size $|\mathbb{O}|d$ increases. As expected, increasing rank r, perturbation level σ and the max norm ψ will lead to a little bit worse estimation, which is consistent to our theoretical analysis in theorem 5. To show the stability of our proposed method, we also consider rectangular tensor case when d_1, d_2, d_3, d_4 are not equivalent as well as imbalanced Tucker rank case when r_1, r_2, r_3, r_4 are not all equivalent and the results are summarized in Appendix.

Three alternative methods are selected for comparison with our proposed method for

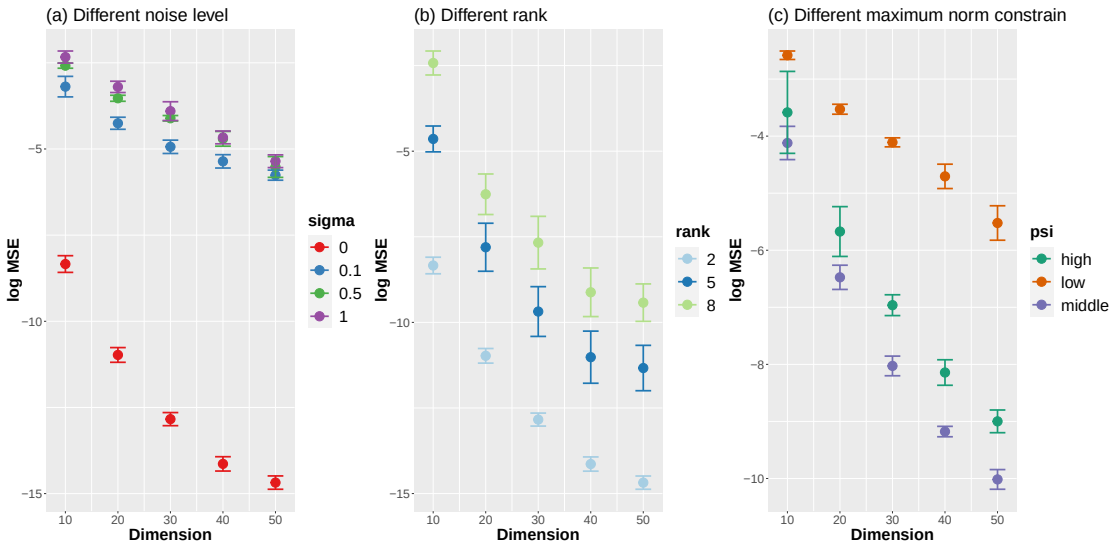


Figure 3.3: $\log \text{MSE} = \log(\|\hat{\mathcal{X}}_{Obs} - \mathcal{X}^*\|_F^2/d^4)$ of imputation for tensor with missing fibers with estimated propensity score via Algorithm 3 and 4 under different choice of dimension d, tucker rank r and max norm ψ. (a) Performance of experiments with various $d \in \{10, 20, 30, 40, 50\}$, $\sigma \in \{0, 0.5, 1, 5\}$ by setting $r = 2$ and $\psi = 1$. (b) Performance of experiments with various $d \in \{10, 20, 30, 40, 50\}$, $r = \{2, 5, 8\}$ by setting $\sigma = 0$ and $\psi = 1$. (c) Performance of experiments with various $d \in \{10, 20, 30, 40, 50\}$ and $\psi = \{1(\text{low}), 5(\text{middle}), 10(\text{high})\}$ by setting $\sigma = 0.5$, $r = 2$.

tensor completion with missing fibers.

- FUnif-TC: To show the benefit of considering fibers are missing not at random, we include tensor completion with fiber missing at uniform (FUnif-TC) into comparison, which follows the similar estimation framework as our proposed method with propensity score estimated uniformly as $|\mathbb{O}|/(d_1 d_2 d_3)$, i.e., the proportion of observed fibers along the fourth mode.

- TenALS: Tensor factorization with missing data via Alternating Least Square (TenALS) proposed in Jain and Oh [2014], which is an alternating minimization based method, iteratively refining estimates of the singular vectors. The available code published online for TenALS is restricted to order three tensors. We generalize it to order four tensor and adopt the similar Robust Tensor Power Method (RTPM) and clipping scheme for initialization.

- MNC-TC: Max-qnorm constrained tensor completion (MNC-TC) was first proposed in Ghadermarzy et al. [2019]. This method tackles the tensor completion problem via solving a constrained least squares estimation using nonconvex max-qnorm constraint.

For convenience of reference, our proposed method is denoted by FMNAR-TC. Through careful investigation, we can summarize the differences among four methods in the following two aspects. First, missingness assumptions are different. TenALS and MNC-TC assume each element in tensor is included in the final observations with a fixed probability p, i.e., the observations are selected uniformly at random. MNC-TC also considers a general version when the observation sampling scheme is non-uniform. FMNAR-TC and FUnif-TC take the fibers missingness into consideration while FMNAR-TC is the only one embracing the assumption of fibers missing not random. Second, the four methods employ the low rank constraint on tensor in a different manner. TenALS utilizes the CP low rank decomposition and iteratively updates each factor matrix in a column-by-column way. MNC-TC develops the definition of max-qnorm for tensor and proves its cruciality in recovering the CP low rank bounded tensor. In contrast, FMNAR-TC and FUnif-TC apply the Tucker low rank decomposition structure. These four methods are implemented with default parameters and the rank is selected using the recommended approach of each. The proposed eBIC is implemented to choose rank for TenALS and FMNAR-TC. Besides, MNC-TC uses the proposed five-point search for the optimal max-qnorm bound in [Ghadermarzy et al., 2019]. We follow the similar data generating framework as in Figure 3.3. To make the comparison of error bars easier, we implement dodging in Figure 3.4 to preserve the vertical position of each result while adjusting its horizontal position.

By varying dimension $d \in \{10, 20, 30, 40, 50\}$ and rank $r \in \{2, 5, 8, 10\}$, Figure 3.4 shows how FMNAR-TC outperforms in different scenarios. As expected, fixing either dimension d or Tucker rank r, FMNAR-TC shows the most accurate estimation. FUnif-TC is comparable to FMNAR-TC when Tucker rank is small, i.e., $r = 2$. This result come as no surprise as

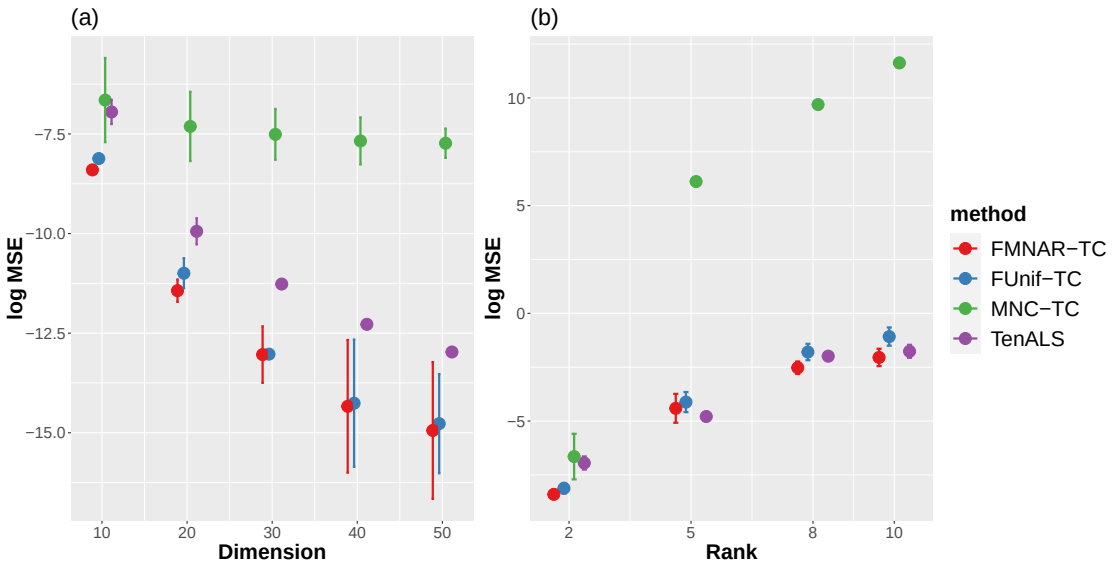


Figure 3.4: $\log \text{MSE} = \log(\|\hat{\mathcal{X}}_{Obs} - \mathcal{X}^*\|_F^2/d^4)$ compare among TenALS, MNC-TC, FMNAR-TC and FUnif-TC under different choice of dimension d, tucker rank r. (a) Performance of experiments with various $d \in \{10, 20, 30, 40, 50\}$ by setting $\sigma = 0$, $r = 2$ and $\psi = 1$. (b) Performance of experiments with various $r = \{2, 5, 8, 10\}$ by setting $d = 10$, $\sigma = 0$ and $\psi = 1$.

when true low rank structure of propensity score is simple, the advantage of fibers missing not at random is not obvious enough to differentiate from uniform missingness. TenALS performs at an intermediate level owing to the fact that CP low rank decomposition can be treated as a specialty of Tucker low rank decomposition by setting the Tucker rank along each modes all equal and core tensor as a super diagonal tensor. Besides, the results for TenALS seem more stable (i.e., shorter error bar) than other three methods and we give this credit to the robust initialization strategy implemented in TenALS. The worst performing method in these experiments is MNC-TC and we believe a more dedicated approach should be introduced to choose max qnorm bound for capturing the correct low rank structure.

3.6.2 Real Data Analysis

In this section, we analyze the *Last.fm Dataset-1K users* collected from Last.fm API to evaluate and compare the performance of popular tensor completion methods. This data was

collected by Òscar Celma [Celma, 2010] during Fall 2008 and is publicly available to download via `http://ocelma.net/MusicRecommendationDataset/lastfm-360K.html`. *Last.fm Dataset-1K users* is a popular music recommendation dataset, which represents the whole listening habits (until May, 5th 2009) for nearly 1,000 users. Tuples (user, timestamp, artist, song) are collected, including 19,150,868 listening records in total. Besides, users profiles are also included which contain gender, age, country and first signup time. The whole dataset contains 992 users, 107528 artists with MBID and 69420 without MBID. To facilitate the performance comparison among FMNAR-TC, TenALS, MNC-TC and FUnif-TC, we extract a small subset from the original dataset and formulate an order four tensor with modes presenting users, artists, songs and yearly time based on the collected tuples. The small subset is obtained following several steps: first, select the top engaged users. We define 'heavy' users who are more dedicated to the Last.fm API by having a large number of listening records. Second, artists and songs without ID are filtered out. Third, select the top 'popular' artists and songs via the number of corresponding listening records. Fourth, transfer the timestamp information into yearly time point. At the end, we conduct performance evaluation on a subset with 101 users, 47 artists, 83 songs and 5 years from 2005 to 2009, which is formulated into an order four tensor $\mathcal{X} \in \mathbb{R}^{101\times 47\times 83\times 5}$.

After further investigation, we can conclude that the fibers along the year mode are missing not random. To recognize this fact, it is important to differentiate the true missing fibers from observations with listening records 0. For example, we can treat the following type of observations as listening record 0, i.e., $\mathcal{X}_{i,j,k,l} = 0$ and $\mathcal{X}_{i,j,k,:} \neq \mathbf{0}$. It means user i had listened to song k from artist j sometime among Year 2005 to Year 2009. But due to unknown reasons, probably user i didn't like artist j anymore, or song k was not popular anymore, or the song was released after year l, or maybe the song was released before year l but the user started to follow artist j after year l...., the listening record for song k at year l is equivalent to 0. Another type of missing is what we care about, the whole fibers

along the year mode is unrevealed, i.e., $\mathcal{X}_{i,j,k,:} = \mathbf{0}$. The missingness mechanism for those fibers are not random. For example, users who used to listen to nostalgic songs may have a lower tendency or probability to be recommended with new electronic music songs, which explains why no listening records related to electronic music for them. But this does not mean electronic music should never be recommended to this type of user. Thus, our major task is to estimate the possible listening records for songs that users didn't play during year 2005 to 2009, furthermore gaining some insights about improvement on recommendation mechanism.

Considering that listening records are non-negative integers, we would like to remove the means along each mode before performing any tensor completion algorithms. Likewise we may also wish to standardize the fibers to have unit variance. This is widely utilized in literatures dealing with recommendation system challenges. For example, Olshen and Rajaratnam [2010] implements a centering and scaling algorithm for complete data and analyze its convergence property. Hastie et al. [2015] proposes a similar centering and scaling scheme for incomplete data. Since selected subset has imbalanced data, we will adopt the similar idea of in Hastie et al. [2015] and generalize to tensor case by learning the centering and scaling parameters for the fibers along four modes. We assume tensor $\mathcal{X}$ with each component $\{\mathcal{X}_{i,j,k,l}\}$ follows the model where the fibers along four modes have mean zero and standard deviation 1 simultaneously, in other words, $\mathcal{X}_{i,j,k,l}$ follows a distribution with mean $\mu_{i,j,k,l}$ and standard deviation $\sigma_{i,j,k,l}$ where

$$\mu_{i,j,k,l} = \alpha_i + \beta_j + \gamma_k + \tau_l$$
$$\sigma_{i,j,k,l} = a_i b_j c_k d_l$$

$\alpha_i, \beta_j, \gamma_k, \tau_l$, a_i, b_j, c_k, d_l represent the mean and standard deviation of fibers along the user, artist, song and year respectively. Given those parameters, each observation is centered and

scaled via

$$\tilde{\mathcal{X}}_{i,j,k,l} = \frac{\mathcal{X}_{i,j,k,l} - \alpha_i - \beta_j - \gamma_k - \tau_l}{a_i b_j c_k d_l}$$

Consider the estimation equation for the mode-1 fiber mean,

$$\frac{1}{|\Omega_i|} \sum_{(j,k,l)\in\Omega_i} \tilde{\mathcal{X}}_{i,j,k,l} = \frac{1}{|\Omega_i|} \sum_{(j,k,l)\in\Omega_i} \frac{\mathcal{X}_{i,j,k,l} - \alpha_i - \beta_j - \gamma_k - \tau_l}{a_i b_j c_k d_l}$$

$\Omega_i = \{(j,k,l)|$ component (i,j,k,l) is revealed$\}$. By setting the above expression equals zero, we can obtain

$$\alpha_i = \frac{\sum_{(j,k,l)\in\Omega_i} \frac{1}{b_j c_k d_l}(\mathcal{X}_{i,j,k,l} - \beta_j - \gamma_k - \tau_l)}{\sum_{(j,k,l)\in\Omega_i} \frac{1}{b_j c_k d_l}}$$

Analogously, the mean for mode-2, mode-3 and mode-4 fibers can be estimated through

$$\beta_j = \frac{\sum_{(i,k,l)\in\Omega_j} \frac{1}{a_i c_k d_l}(\mathcal{X}_{i,j,k,l} - \alpha_i - \gamma_k - \tau_l)}{\sum_{(i,k,l)\in\Omega_j} \frac{1}{a_i c_k d_l}}$$

$$\gamma_k = \frac{\sum_{(i,j,l)\in\Omega_k} \frac{1}{a_i b_j d_l}(\mathcal{X}_{i,j,k,l} - \alpha_i - \beta_j - \tau_l)}{\sum_{(i,j,l)\in\Omega_k} \frac{1}{a_i b_j d_l}}$$

$$\tau_l = \frac{\sum_{(i,j,k)\in\Omega_l} \frac{1}{a_i b_j c_k}(\mathcal{X}_{i,j,k,l} - \alpha_i - \beta_j - \gamma_k)}{\sum_{(i,j,k)\in\Omega_l} \frac{1}{a_i b_j c_k}}$$

The variance condition for mode-1 fiber is

$$\frac{1}{|\Omega_i|} \sum_{(j,k,l)\in\Omega_i} \tilde{\mathcal{X}}^2_{i,j,k,l} = \frac{1}{|\Omega_i|} \sum_{(j,k,l)\in\Omega_i} \frac{(\mathcal{X}_{i,j,k,l} - \alpha_i - \beta_j - \gamma_k - \tau_l)^2}{a_i^2 b_j^2 c_k^2 d_l^2} = 1, \quad \forall i \in [d_1]$$

And we can derive

$$a_i^2 = \frac{1}{|\Omega_i|} \sum_{(j,k,l)\in\Omega_i} \frac{(\mathcal{X}_{i,j,k,l} - \alpha_i - \beta_j - \gamma_k - \tau_l)^2}{b_j^2 c_k^2 d_l^2}$$

Analogously, we can derive the variance for fibers along other modes in the same manner,

$$b_j^2 = \frac{1}{|\Omega_j|} \sum_{(i,k,l)\in\Omega_j} \frac{(\mathcal{X}_{i,j,k,l} - \alpha_i - \beta_j - \gamma_k - \tau_l)^2}{a_i^2 c_k^2 d_l^2}$$

$$c_k^2 = \frac{1}{|\Omega_k|} \sum_{(i,j,l)\in\Omega_k} \frac{(\mathcal{X}_{i,j,k,l} - \alpha_i - \beta_j - \gamma_k - \tau_l)^2}{a_i^2 b_j^2 d_l^2}$$

$$d_l^2 = \frac{1}{|\Omega_l|} \sum_{(i,j,k)\in\Omega_l} \frac{(\mathcal{X}_{i,j,k,l} - \alpha_i - \beta_j - \gamma_k - \tau_l)^2}{a_i^2 b_j^2 c_k^2}$$

Empirically, $\alpha_i, \beta_j, \gamma_k, \tau_l$ are estimated iteratively by the above equations until the following 'residual' converges to zero,

$$\begin{aligned}
\text{residual} = &\sum_{i=1}^{d_1} (\frac{1}{|\Omega_i|} \sum_{(j,k,l)\in\Omega_i} \tilde{\mathcal{X}}_{i,j,k,l})^2 + \sum_{j=1}^{d_2} (\frac{1}{|\Omega_j|} \sum_{(i,k,l)\in\Omega_j} \tilde{\mathcal{X}}_{i,j,k,l})^2 \\
&+ \sum_{k=1}^{d_3} (\frac{1}{|\Omega_k|} \sum_{(i,j,l)\in\Omega_k} \tilde{\mathcal{X}}_{i,j,k,l})^2 + \sum_{l=1}^{d_4} (\frac{1}{|\Omega_l|} \sum_{(i,j,k)\in\Omega_l} \tilde{\mathcal{X}}_{i,j,k,l})^2 \\
&+ \sum_{i=1}^{d_1} \log^2 (\frac{1}{|\Omega_i|} \sum_{(j,k,l)\in\Omega_i} \tilde{\mathcal{X}}_{i,j,k,l}^2) + \sum_{j=1}^{d_2} \log^2 (\frac{1}{|\Omega_j|} \sum_{(i,k,l)\in\Omega_j} \tilde{\mathcal{X}}_{i,j,k,l}^2) \\
&+ \sum_{k=1}^{d_3} \log^2 (\frac{1}{|\Omega_k|} \sum_{(i,j,l)\in\Omega_k} \tilde{\mathcal{X}}_{i,j,k,l}^2) + \sum_{l=1}^{d_4} \log^2 (\frac{1}{|\Omega_l|} \sum_{(i,j,k)\in\Omega_l} \tilde{\mathcal{X}}_{i,j,k,l}^2)
\end{aligned}$$

We perform the above centering steps, and it is obvious from Figure 3.5 that the residual terms converges to 0 pretty fast. In order to evaluate the effect of ratio of observations, we vary the ratio of observations $p \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. For a specific p, we randomly select p of observed fibers in selected subset as training set and the remaining

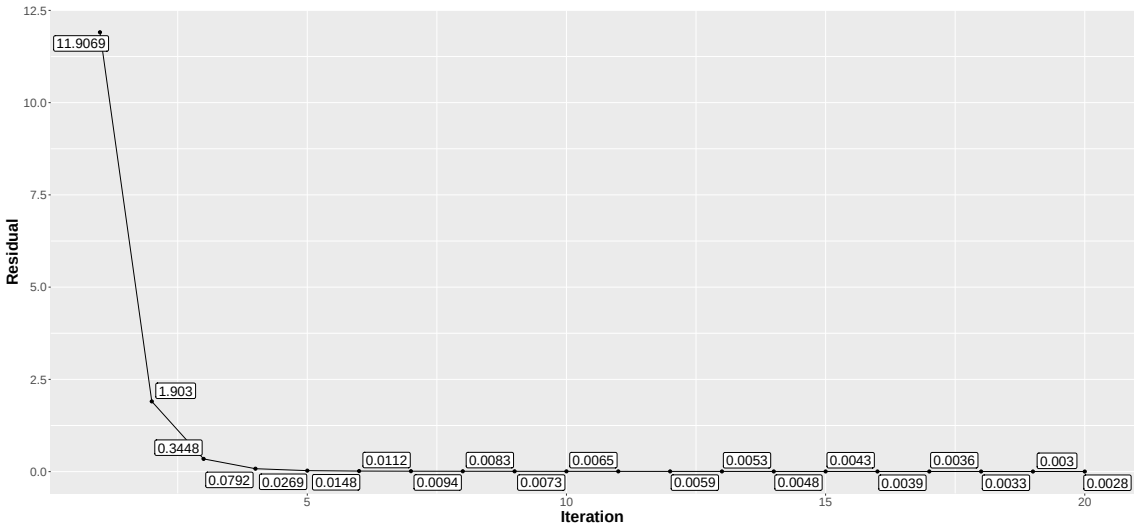


Figure 3.5: Residual path of centering and scaling procedure

entries as testing set. The error bars show one standard error which is calculated based on 10 replications. Entries in testing set are treated as missing and we report the performance of four methods on testing set in terms of mean squared error (MSE) under different p in Figure 3.6. Similarly, as the simulation results reported in the previous section, FMNAR-TC and FUnif-TC produced highly comparable results but FMNAR-TC is always better by achieving lower MSE. In particular, the difference between FMNAR-TC and FUnif-TC are larger when p is small, indicating the applicability of FMNAR-TC to larger dataset whenever the observations are scarce. Although FMNAR-TC is slightly inferior to MNC-TC when the observational proportion p is less than 0.4, we observe this dominance of MNC-TC diminishes gradually as more and more observations are included. This might be due to difficulty in choosing the 'optimal' hyperparameters in FMNAR-TC when only a few observations are revealed. Moreover, FMNAR-TC and FUnif-TC offere more consistent results when p is higher than 0.6 while MNC-TC exhibits much larger variations. As a consequence of simple structure of CP low rank, we can conclude that TenALS does not perform as well as FMNAR-TC and FUnif-TC to capture the structure of missing fibers, which is different from the simulation results. This can be partially explained by the incapability of CP low

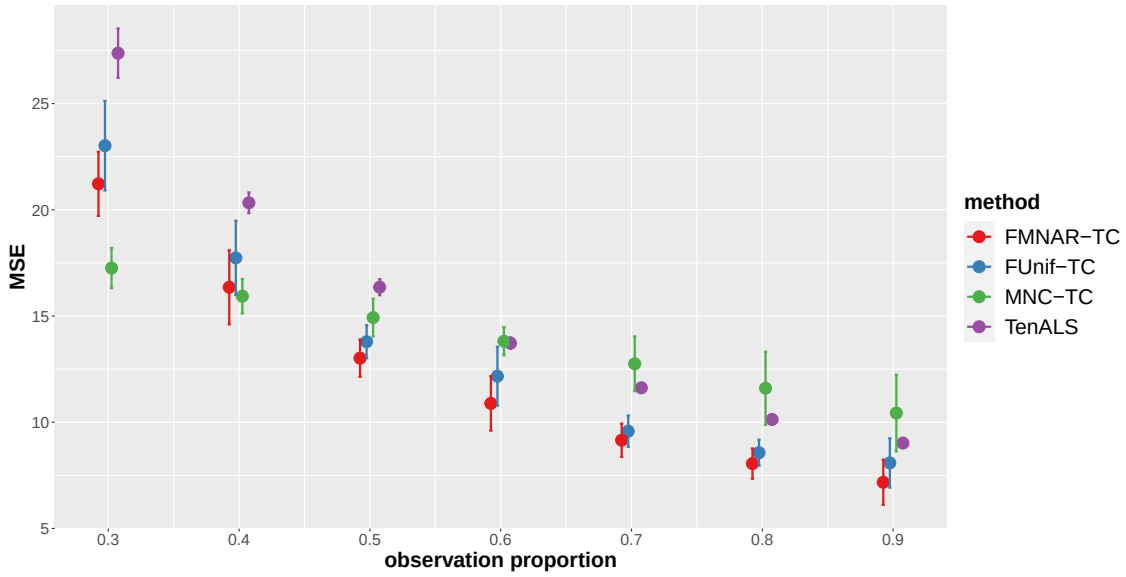


Figure 3.6: Comparison of MSE on testing set among TenALS, MNC-TC, FUnif-TC and FMNAR-TC under different choices of observation proportion.

rank structure of TenALS to deal with imbalanced dimension along each mode for selected subset, where the dimension for year mode is significantly smaller than dimensions for user, music or song modes. Overall, FMNAR-TC and FUnif-TC are among the top performers of the analysis reported in Figure 3.6 and we recommend to use FMNAR-TC when the number of revealed entries are scarce.

CHAPTER 4

LEARNING INCOMPLETE SPARSE TENSOR WITH AUXILIARY INFORMATION

Tensors, or multi-dimensional arrays, arise naturally in many applications, including recommender systems [Bi et al., 2018, Frolov and Oseledets, 2017, Nasiri et al., 2014], neuroimaging analysis [Zhou et al., 2013, Sun and Li, 2017, Li et al., 2018], statistical genetics analysis [Hore et al., 2016, Wang et al., 2019, Feizi et al., 2017] and computer vision [Cohen et al., 2017, Mordohai and Medioni, 2006]. Low-rank tensor completion, whose major goal is to recover low rank tensors by imputing missing entries based on limited numbers of measurable entries, has revived by massive interesting media and business datasets. For example, drug bank dataset which aims to characterize the drug-protein interaction signatures by analyzing drug effects on different diseases, can be formulated into an order three tensor with dimension 593(drugs) $\times$ 501(target protein) $\times$ 313(diseases). MovieLens 1M dataset, which is a standard dataset of 1-million movie-ratings (1-5) star and a popular dataset extensively utilized in recommender systems, can be organized as an order three tensor with dimension 6040(movies) $\times$ 3883(users) $\times$ 150(timestamp).

There are abundant well-developed methods for order two tensor (matrix) completion in the fields of high-dimensional statistics and machine learning, providing many useful insights on higher order tensor completion. Candès and Recht [2009], Candès and Tao [2010], Recht [2011], Gross [2011] focus on matrix recovery using convex approximation such as nuclear norm minimization via uniformly randomly selected observed entries. As shown in Gross [2011], for a $n_1 \times n_2$ matrix of rank r, one can achieve excellent recovery property with high probability as long as roughly $O(r(n_1 + n_2)\log^2(n_1 + n_2))$ entries are randomly and uniformly observed. Another streamline work of matrix completion assumes the observed entries are sampled by some deterministically sampling patterns [Chen et al., 2014] or column subset selection [Wang and Singh, 2015, Cai et al., 2016]. Strong theoretical

properties can be obtained under these settings. Chen et al. [2014] has successfully proved that by $O(\max\{n_1, n_2\} r \log^2(n_1 + n_2))$, nuclear norm minimization can recover an arbitrary matrix provided that revealed entries are drawn proportionally to the local row and column coherences. Thus, perfect recovery of a matrix with low rank structure is plausible when using a small fraction of observed entries.

After the great success of matrix completion, tensor completion has been a popular generalization in recent years. Similarly, most of the literatures focus on the situation when the observed entries in order three tensor $\mathcal{R} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ are uniformly and randomly selected. For example, inspired by approximating low rank constrain on the matrix by minimizing nuclear norm, Gandy et al. [2011], Liu et al. [2012], Tomioka et al. [2011] propose sum of nuclear norm via matricization of order three tensor and derive that the tensor can be reconstructed with high probability provided the number of observed entries is larger than $O\Big((n_1 n_2 r + n_1 n_3 r + n_2 n_3 r) \log^2(n_1 + n_2 + n_3)\Big)$. Jain and Oh [2014] takes the alternative minimization method for imputing low rank tensor with CP decomposition structure and orthogonal factor matrices, which requires $O(r^5 n^{3/2} \log^4 n)$ observed samples to achieve perfect recovery by assuming n_1, n_2, n_3 take the same value n. Xia et al. [2021] establishes minimax optimal rates of convergence for low rank noisy tensor completion with conditions on observed sample size of order $O(r\sqrt{n_1 n_2 n_3} \log^5 \max\{n_1, n_2, n_3\} + r^2 \max\{n_1, n_2, n_3\} \log^{10} \max\{n_1, n_2, n_3\})$.

Even though there have been a lot of progress on tensor completion, one critical challenge dealing with higher order tensor completion problem is large proportional missing entries, which may violate the conditions on sample size in the above proposed methods. For instance, the drug bank dataset we mentioned contains 20778 observed entries, consisting of only 0.0223% of the whole entries in a $593 \times 501 \times 313$ tensor. Similarly, for MovieLens 1M there are only 1000209 observed values, or in other words, 0.0284% observed entries among the whole $6040 \times 3883 \times 150$ tensor. To achieve better recovery performance with fewer observed samples, auxiliary information for each mode are employed. One commonly used form of the

auxiliary information is the feature matrix, which captures the statistical properties of tensor modes. For example, in the MovieLens 1M dataset, auxiliary information are accessible for 'movies' and 'users' modes. Different information for occupation type, gender, age etc have been collected for all the users. Moreover, there are 18 genres for movies in total, which facilitates to construct dummy-variables feature matrix for movies. Another form of auxiliary information is through constructing similarity matrix. The drug bank dataset constructs similarity matrix for the drug modes by analyzing the chemical structure and side effects for different drugs. Similarly, protein sequence and GO annotation for protein will be insightful to establish the similarity matrix for the target mode. Moreover, similarity matrix for disease can utilize the information from disease phenotype and human phenotype ontology. Many trials have been conducted to introduce the auxiliary information into tensor completion. Zhou et al. [2017] proposes a Riemannian conjugate gradient descent algorithm to impute the missing entries in tensor with auxiliary information. Nimishakavi et al. [2018] focuses on multi-aspect streaming settings for inductive tensor completion, meaning that the tensor size and the size of corresponding auxiliary matrices will monotonically increase as time t increases. Bertsimas and Pawlowski [2020] considers the situation when the side information are not available for all the modes, i.e., we can only access the auxiliary information for one or two modes of the tensor observations. Unfortunately, all these methods fail to provide a rigorous theoretical sample complexity analysis for demonstrating good recovery.

The goal of this chapter is to answer the following questions: is it possible to generalize convex relaxation methods of matrix completion to tensor completion with auxiliary information? If so, what is the sample complexity, i.e., the order of minimum number of entries to be observed in $\mathcal{R}$ so that the perfect recovery property can be achieved under the setting when auxiliary information can be used to fully describe the latent structure of $\mathcal{R}$? What if the auxiliary information are not perfect, i.e., some noisy terms may exist that perturb the auxiliary information, reducing the ability of auxiliary information revealing the latent struc-

ture of $\mathcal{R}$? Can we still obtain reasonable recovery performance with some mild assumptions on those noisy terms provided limited number of observed entries? Considering the previous discussion, we propose **L**earn **I**ncomplete **S**parse **T**ensor with **A**uxiliary **I**nformation (LISTAI) for tensor completion with auxiliary information by borrowing the idea of using nuclear norm as regularization term to impose low rank constraint and inductive settings for matrix/tensor completion. To illustrate the properties of the proposed procedure, both theoretical analysis and implementation on datasets are provided. We successfully show that with perfect auxiliary information, exact recovery can be attained with high probability if requirement for the minimum number of observed entries is satisfied. Moreover, with noisy auxiliary information, a theoretical foundation is still obtained to show the effectiveness of this model. Formally, we quantify the quality of the auxiliary information and derive the corresponding the sample complexity of the proposed model to achieve arbitrarily small expected risk from a statistical learning perspective. In addition, on account of tensor nuclear norm can be computational NP hard and there is no practical applicable algorithm to implement, we propose a nested double ADMM algorithm to solve the approximation of the original optimization problem and derive the convergence property in finite iterations. Empirically, we show our proposed algorithm outperforms other popular tensor completion methods with auxiliary information on synthetic data as well as real world datasets.

This chapter is organized as follows. Section 4.1 introduces our proposed LISTAI model which aims to recover the low rank tensors with partially observed entries as well as general auxiliary information matrices. We theoretically analyze the recovery property of our proposed model under full rank auxiliary information matrices in section 4.2 and corrupted auxiliary information in section 4.3. A new nested double ADMM algorithm is proposed to solve the approximated LISTAI and its convergence property is analyzed in section 4.4. We show experimental results on simulation experiments and real datasets in section 4.5.

For the purpose of notational convenience, we use capital letter P to represent an operator

but different operators have different subscripts on P. Since linear operators are widely used in this chapter, we define the induced operator norm of operator $P : \mathbb{R}^{n_1 \times n_1 \times n_3} \to \mathbb{R}$ as

$$\|P\| = \max\{\|P(\mathcal{R})\|_F : \mathcal{R} \in \mathbb{R}^{n_1 \times n_2 \times n_3}, \|\mathcal{R}\|_F \leq 1\}$$

4.1 Methods

Let $\mathcal{R} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ be a partially observed order three tensor and Ω be a subset of $[n_1] \times [n_2] \times [n_3]$ where $[n_i] = \{1, 2, ..., n_i\}, i \in \{1, 2, 3\}$. The goal of tensor completion with auxiliary information is to recover $\mathcal{R}$ when observing only entries $\mathcal{R}_{i,j,k}$ for $(i, j, k) \in \Omega$ given auxiliary information matrices $\mathbf{X'} \in \mathbb{R}^{n_1 \times d_1'}, \mathbf{Y'} \in \mathbb{R}^{n_2 \times d_2'}, \mathbf{Z'} \in \mathbb{R}^{n_3 \times d_3'}$ for the three modes respectively.

Inspired by the simple linear predictive model, i.e., $\mathcal{R}_{i,j,k} = \mathbf{x}_i'^\top \boldsymbol{\beta}_x + \mathbf{y}_j'^\top \boldsymbol{\beta}_y + \mathbf{z}_k'^\top \boldsymbol{\beta}_z + \beta_0$, where $\mathbf{x}_i', \mathbf{y}_j', \mathbf{z}_k'$ are the feature vectors respectively for the first, second and third mode and $\boldsymbol{\beta}_x, \boldsymbol{\beta}_y, \boldsymbol{\beta}_z, \beta_0$ are the corresponding model parameters, we consider adding interaction terms on the basis of simple linear predictive model by introducing an interaction tensor $\mathcal{G}' \in \mathbb{R}^{d_1' \times d_2' \times d_3'}$ and three interaction matrices $\mathbf{B}_{xy} \in \mathbb{R}^{d_1' \times d_2'}$,$\mathbf{B}_{xz} \in \mathbb{R}^{d_1' \times d_3'}$, $\mathbf{B}_{yz} \in \mathbb{R}^{d_2' \times d_3'}$, leading to the following interactive-predictive model

$$\mathcal{R}_{i,j,k} = [\mathcal{G}'; \mathbf{x}_i', \mathbf{y}_j', \mathbf{z}_k'] + \mathbf{x}_i'^\top \mathbf{B}_{xy} \mathbf{y}_j' + \mathbf{x}_i'^\top \mathbf{B}_{xz} \mathbf{z}_k' + + \mathbf{y}_j'^\top \mathbf{B}_{yz} \mathbf{z}_k' + \mathbf{x}_i'^\top \boldsymbol{\beta}_x + \mathbf{y}_j'^\top \boldsymbol{\beta}_y + \mathbf{z}_k'^\top \boldsymbol{\beta}_z + \beta_0 \quad (4.1)$$

By setting $\mathbf{x}_i = [\mathbf{x}_i'^\top \ \ 1]^\top$, $\mathbf{y}_j = [\mathbf{y}_j'^\top \ \ 1]^\top$, $\mathbf{z}_k = [\mathbf{z}_k'^\top \ \ 1]^\top$ and reformulating the interaction tensor as $\mathcal{G} \in \mathbb{R}^{(d_1 = d_1' + 1) \times (d_2 = d_2' + 1) \times (d_3 = d_3' + 1)}$ by augmenting $\mathcal{G}', \mathbf{B}_{xy}, \mathbf{B}_{xz}, \mathbf{B}_{yz}, \boldsymbol{\beta}_x, \boldsymbol{\beta}_y, \boldsymbol{\beta}_z$ as in Figure 4.1, the above interactive-predictive model can be simplified as

$$\mathcal{R} = [\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}] \quad (4.2)$$

where $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ are the corresponding augmented auxiliary information matrices formulated

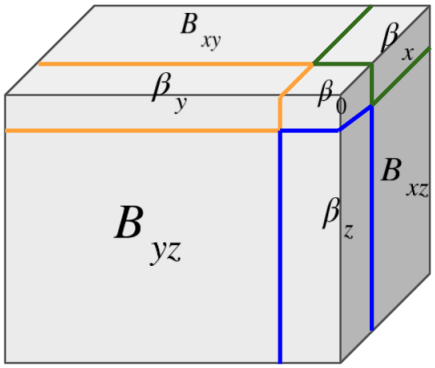


Figure 4.1: Formulation of interaction tensor $\mathbf{G}$

by rows $\mathbf{x}_i, i \in [n_1]$, $\mathbf{y}_j, j \in [n_2]$, and $\mathbf{z}_k, k \in [n_3]$ respectively.

Inspired by the interactive-predictive model, we propose a new model called Learning Incomplete Sparse Tensor with Auxiliary Information (LISTAI) which incorporates the low rank constrains on latent tensor structure and sparsity regularization on interaction tensor $\mathcal{G}$.

Classically, low rank tensor completion problem aims to solve

$$\min_{\mathcal{Q}} \ \text{rank}(\mathcal{Q}), \quad \text{s.t.} P_\Omega(\mathcal{Q}) = P_\Omega(\mathcal{R}) \tag{4.3}$$

where rank$(\cdot)$ constrains $\mathcal{Q}$ to be identified with a certain low-rank structure (i.e., CP low rank or Tucker low rank) and $P_\Omega(\cdot)$ is the mapping giving a tensor whose (i, j, k) entry as $\mathcal{R}_{i,j,k}$ if $(i, j, k) \in \Omega$ (or 0 otherwise).

Following low rank tensor completion framework, we will impose a low-rank structure tensor $\mathcal{Q} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ to approximate the partially observed tensor $\mathcal{R}$; $\mathcal{Q}$ can be further constructed by inductive approximation using interactive-predictive model, i.e., $[\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}]$. Different from most existing methods which impose low rank constrain on $\mathcal{G}$, our model takes the following fact into consideration: the number of features for auxiliary information matrices $\mathbf{X}', \mathbf{Y}', \mathbf{Z}'$ are much smaller than the number of observations along each mode of $\mathcal{R}$,

i.e., $d_1 \ll n_1, d_2 \ll n_2, d_3 \ll n_3$. For example, a well known baseline dataset in recommender system is MovieLens1M (`https://grouplens.org/datasets/movielens/1m/`), which can be organized as an order three tensor of users $\times$movies $\times$weeks. The number of users, movies and weeks are 6040, 3952 and 149 respectively. Besides, different features are provided for users and movies. For users, each person's occupation type is collected and there are 21 occupations in total. For movies, each movie is classified into one group among 18 genres. Obviously, the number of feature vectors $21 \ll 6040, 18 \ll 3952$. Thus, adding low rank constraint on $\mathcal{G}$ can be over-restrictive since small number of feature vectors will lead to limited number of parameters to estimate in $\mathcal{G}$.

To add a low rank constraint on $\mathcal{Q}$, we adopt the strategy from matrix completion by adding nuclear norm regularization, i.e., $\|\mathcal{Q}\|_*$. We will analyze the exact recovery and ϵ recovery property of LISTAI method in section 4.2 and 4.3. However, considering the fact that there are no practical applicable algorithms to calculate the tensor nuclear norm, a common strategy to overcome the challenges with high order tensor is to unfold them into matrices and then resort to usual nuclear norm minimization heuristics for matrices. In section 4.4, $\mathcal{Q}$ will be unfolded into matrices along first, second and third mode respectively and nuclear norm regularizer will be imposed on those three unfolded matrices. We use $(\|\mathcal{Q}_{(1)}\|_* + \|\mathcal{Q}_{(2)}\|_* + \|\mathcal{Q}_{(3)}\|_*)/3$ as an approximation to $\|\mathcal{Q}\|_*$. This method is known as sum of nuclear norms (SNN) and has been widely implemented in Liu et al. [2012], Signoretto et al. [2010], Tomioka et al. [2011].

Instead of adding a low rank constraint, a sparse regularizer is employed on $\mathcal{G}$ so that not all the interaction terms among three modes, i.e. (entries in interaction tensor $\mathcal{G}'$) or two modes, i.e., (entries in interaction matrices, $\mathbf{B}_{xy}, \mathbf{B}_{xz}, \mathbf{B}_{yz}$) are useful for predicting entries in $\mathcal{R}$. Naturally, our model impose sparsity on $\mathcal{G}$ by introducing a regularizer on $\|\mathcal{G}\|_1$.

More specifically, LISTAI follows the below estimation logic

$$[\underbrace{\mathcal{G}}_{\text{sparsity}};\mathbf{X},\mathbf{Y},\mathbf{Z}] \approx \underbrace{\mathcal{Q}}_{\text{low rank constraint}} \approx \mathcal{R}$$

and it aims to solve the following optimization

$$\begin{aligned} &\min_{\mathcal{Q},\mathcal{G}} \ \lambda_G\|\mathcal{G}\|_1 + \lambda_Q\|\mathcal{Q}\|_* \\ &\text{s.t. } P_\Omega(\mathcal{Q}) = P_\Omega(\mathcal{R}), \ \ [\mathcal{G};\mathbf{X},\mathbf{Y},\mathbf{Z}] = \mathcal{Q} \end{aligned} \tag{4.4}$$

The solution for $\mathcal{Q}$ provides recovery $\hat{\mathcal{R}}$ for the partially observed tensor. λ_G and λ_Q are two crucial tuning parameters that need to be chosen carefully since they control the penalization imposed on sparsity of $\mathcal{G}$ and low rank constraint on $\mathcal{Q}$. When $\lambda_G = \infty$, $\mathcal{G}$ will be enforced as 0 and the optimization problem 4.4 will simplify to the standard tensor completion problem as optimization problem 4.3 without auxiliary information. Thus, our proposed LISTAI model is still applicable even when there is no access to auxiliary information.

4.2 Sample complexity for exact recovery

In this section, we study the theoretical performance of LISTAI when the latent structure of tensor $\mathcal{R}$ can be fully described by the auxiliary information matrices provided. Let $\mathcal{G}_0, \mathcal{Q}_0$ satisfy $P_\Omega(\mathcal{Q}_0) = P_\Omega(\mathcal{R})$ and $[\mathcal{G}_0;\mathbf{X},\mathbf{Y},\mathbf{Z}] = \mathcal{Q}_0$. Our goal is to determine what requirement should be imposed on the sample size $|\Omega|$ to guarantee that, with high probability, $(\mathcal{G}_0, \mathcal{Q}_0)$ is the unique optimizer of problem 4.4, i.e., $\hat{\mathcal{R}}$ is an exact recovery of $\mathcal{R}$.

4.2.1 Preliminaries

Assume $\mathcal{R}$ enjoys a CP low rank structure, i.e., it can be decomposed into multiplication among weights $\mathbf{a} = [a_1, ..., a_r]^\top \in \mathbb{R}^r$ and three factor matrices $\mathbf{U} = [\mathbf{U}_{:1}, ..., \mathbf{U}_{:r}] \in$

$\mathbb{R}^{n_1 \times r}, \mathbf{V} = [\mathbf{V}_{:1}, ..., \mathbf{V}_{:r}] \in \mathbb{R}^{n_2 \times r}, \mathbf{W} = [\mathbf{W}_{:1}, ..., \mathbf{W}_{:r}] \in \mathbb{R}^{n_3 \times r}$, i.e., $\mathcal{R} = \sum_{i=1}^{r} a_i \mathbf{U}_{:i} \circ \mathbf{V}_{:i} \circ \mathbf{W}_{:i}$ where $\mathbf{U}, \mathbf{V}, \mathbf{W}$ have columnwise unit 1 norm, $\|\mathbf{U}_{:i}\| = \|\mathbf{V}_{:i}\| = \|\mathbf{W}_{:i}\| = 1$. Because CP low rank structure can be considered as a special situation of tensor multiplication, we formulate $\mathcal{R}$ as $[\mathcal{A}; \mathbf{U}, \mathbf{V}, \mathbf{W}]$ where $\mathcal{A}$ is a diagonal tensor with $\mathbf{a}$ on its super diagonal and all the other entries as 0. We choose CP low rank decomposition due to its relatively simple formulation and competitive empirical performance. It has been extensively implemented in many tensor related research topics, like tensor clustering [Sun and Li, 2019, Wang et al., 2019] and recommendation systems [Bi et al., 2018]. Specifically, mode-1 fiber of $\mathcal{R}$ is $\mathcal{R}_{:,j,k} \in \mathbb{R}^{n_1} = (\mathcal{R}_{1,j,k}, ..., \mathcal{R}_{n_1,j,k})^\top, \forall j \in [n_2], k \in [n_3]$. Mode-2 and mode-3 fiber of $\mathcal{R}$ can be defined in a similar manner. Let $\mathcal{L}_1(\mathcal{R}), \mathcal{L}_2(\mathcal{R}), \mathcal{L}_3(\mathcal{R})$ be the linear space spanned by a collection of mode-1, mode-2, mode-3 fibers respectively and the dimension of $\mathcal{L}_i(\mathcal{R})$ is $r_i, \forall i \in \{1, 2, 3\}$ which are often referred to the tucker rank of $\mathcal{R}$.

Let P_j be an arbitrary projection from $\mathbb{R}^{n_j}$ to a linear space of $\mathbb{R}^{n_j}$ and we define the tensor projection $P_1 \otimes P_2 \otimes P_3$ on $\mathcal{R}$ as

$$(P_1 \otimes P_2 \otimes P_3)\mathcal{R} = [\mathcal{A}; P_1\mathbf{U}, P_2\mathbf{V}, P_3\mathbf{W}]$$

Next, we consider the singular value decomposition of $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ as $\mathbf{X} = \mathbf{U}_X \mathbf{\Sigma}_X \mathbf{V}_X^\top, \mathbf{Y} = \mathbf{U}_Y \mathbf{\Sigma}_Y \mathbf{V}_Y^\top, \mathbf{Z} = \mathbf{U}_Z \mathbf{\Sigma}_Z \mathbf{V}_Z^\top$ where all $\mathbf{\Sigma}_X, \mathbf{\Sigma}_Y, \mathbf{\Sigma}_Z$ are full rank and define the following projections,

$$P_U = \mathbf{U}\mathbf{U}^\top, P_V = \mathbf{V}\mathbf{V}^\top, P_W = \mathbf{W}\mathbf{W}^\top$$
$$P_X = \mathbf{U}_X\mathbf{U}_X^\top, P_Y = \mathbf{U}_Y\mathbf{U}_Y^\top, P_Z = \mathbf{U}_Z\mathbf{U}_Z^\top$$

where they project a vector onto the subspaces spanned by the columns in $\mathbf{U}, \mathbf{V}, \mathbf{W}, \mathbf{X}, \mathbf{Y}, \mathbf{Z}$ respectively. Assume we are under the perfect feature settings where feature matrices $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$

fully describe the true latent structure space of $\mathcal{R}$, in other words,

$$\mathcal{L}_1(\mathcal{R}) \subseteq \mathcal{L}(\mathbf{X}), \quad \mathcal{L}_2(\mathcal{R}) \subseteq \mathcal{L}(\mathbf{Y}), \quad \mathcal{L}_3(\mathcal{R}) \subseteq \mathcal{L}(\mathbf{Z})$$

where $\mathcal{L}(\mathbf{X}), \mathcal{L}(\mathbf{Y}), \mathcal{L}(\mathbf{Z})$ represent the linear space spanned by the columns in $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$. Under this assumption, we define the $P_{U^\perp}$ satisfying $P_{U^\perp} + P_U = P_X$. Similarly, $P_{V^\perp} + P_V = P_Y$ and $P_{W^\perp} + P_W = P_Z$. We also define the following tensor projections that will be used extensively later:

$$\begin{aligned}
P^0_{\mathcal{R}} &= P_U \otimes P_V \otimes P_W & P^0_{\mathcal{R}^\perp} &= P_{U^\perp} \otimes P_{V^\perp} \otimes P_{W^\perp} \\
P^1_{\mathcal{R}} &= P_{U^\perp} \otimes P_V \otimes P_W & P^1_{\mathcal{R}^\perp} &= P_U \otimes P_{V^\perp} \otimes P_{W^\perp} \\
P^2_{\mathcal{R}} &= P_U \otimes P_{V^\perp} \otimes P_W & P^2_{\mathcal{R}^\perp} &= P_{U^\perp} \otimes P_V \otimes P_{W^\perp} \\
P^3_{\mathcal{R}} &= P_U \otimes P_V \otimes P_{W^\perp} & P^3_{\mathcal{R}^\perp} &= P_{U^\perp} \otimes P_{V^\perp} \otimes P_W \\
P_{\mathcal{R}} &= P^0_{\mathcal{R}} + P^1_{\mathcal{R}} + P^2_{\mathcal{R}} + P^3_{\mathcal{R}} & P_{\mathcal{R}^\perp} &= P^0_{\mathcal{R}^\perp} + P^1_{\mathcal{R}^\perp} + P^2_{\mathcal{R}^\perp} + P^3_{\mathcal{R}^\perp}
\end{aligned}$$

A crucial concept in matrix completion is coherence and we will generalize it to the tensor case. Recall that the coherence of an r dimensional linear subspace $\mathbf{U}$ of $\mathbb{R}^{n_1}$ is defined to be

$$\mu(\mathbf{U}) = \frac{n_1}{r} \max_{1 \le i \le n_1} \|P_U \mathbf{e}_i\|^2$$

where P_U is the orthogonal projection onto $\mathbf{U}$ and $\mathbf{e}_i$ is the canonical basis for $\mathbb{R}^{n_1}$. Thus, we define the coherence measure for tensor $\mathcal{R}$ as

$$\begin{aligned}
\mu_0 &= \max\Big\{\mu(\mathcal{L}_1(\mathcal{R})), \mu(\mathcal{L}_2(\mathcal{R})), \mu(\mathcal{L}_3(\mathcal{R}))\Big\} \\
&= \max\Big\{\frac{n_1}{r} \max_{1 \le i \le n_1} \|P_U \mathbf{e}_i\|^2, \frac{n_2}{r} \max_{1 \le j \le n_2} \|P_V \mathbf{e}_j\|^2, \frac{n_3}{r} \max_{1 \le k \le n_3} \|P_W \mathbf{e}_k\|^2\Big\}
\end{aligned}$$

Obviously, $\mu_0 \geq 1$ since $\mu(\mathbf{U})$ is the ratio of the ℓ_∞ and the length normalized ℓ_2 norms of a vector. Also, we define the coherence measurement among auxiliary information matrices as

$$\mu_{xyz} = \max\Big\{ \max_{1\leq i\leq n_1} \frac{n_1}{d_1}\|\mathbf{x}_i\|_2^2, \max_{1\leq j\leq n_2} \frac{n_2}{d_2}\|\mathbf{y}_j\|_2^2, \max_{1\leq k\leq n_3} \frac{n_3}{d_3}\|\mathbf{z}_k\|_2^2 \Big\}$$

4.2.2 Sample size requirement for exact tensor recovery

With the above preliminaries, we can obtain the following theorem by denoting $d_s = d_1+d_2+d_3, d_p = d_1d_2d_3, d_{\max} = \max\{d_1, d_2, d_3\}$. If the auxiliary information matrices $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ are all full column rank, $\hat{\mathcal{R}}$, which is the solution to 4.4, will be an exact recovery for $\mathcal{R}$ with high probability provided there are $O(\max\{r^2 d_s \log(r^2 d_s), (d_p + r^3 - r^2 d_s)\log(d_p + r^3 - r^2 d_s)\})$ observed entries in $\mathcal{R}$.

Theorem 6. *Let Ω be a uniformly sampled subset of $[n_1] \times [n_2] \times [n_3]$ and $\hat{\mathcal{R}}$ be the solution to 4.4. Define q_1, q_2 as constants which separate $|\Omega|$ into q_2 subsequences with each length as q_1 and $q_2 \geq (1/\log 2)\log(\sqrt{32}n_1n_2n_3|\Omega|^{-1/2})$ and we further denote $\gamma = \max\{\|\mathbf{\Sigma}_X^{-1}\|, \|\mathbf{\Sigma}_Y^{-1}\|, \|\mathbf{\Sigma}_Z^{-1}\|\}$. If we assume there exists a constant C such that $\|\mathcal{G}_0\|_1 \leq \lambda_G(\frac{1}{2} - \frac{\xi}{4}\sqrt{\frac{|\Omega|}{2n_1n_2n_3}})/(C\lambda_Q)$ with $\xi \leq \sqrt{5/2}$, then with probability at least $1 - \psi_1 - \psi_2 - 2q_2(d_s-2r)r^2 \exp\Big\{-\frac{3}{32}\frac{q_1}{\gamma\mu_{xyz}\mu_0^2 r^2 d_s}\Big\} - 2q_2n_1n_2n_3 \exp\Big\{-\frac{3q_1}{32\gamma\mu_{xyz}\mu_0^2 r^2 d_s}\Big\} - q_2(n_1n_2n_3)^{-\beta-1}$ for $\beta > 0$, we have $\hat{\mathcal{R}} = \mathcal{R}$ as long as*

$$\begin{aligned}|\Omega| \geq \max\Big\{ &\frac{32}{3}\log\frac{2(d_s-2r)r^2}{\psi_1}\gamma\mu_{xyz}\mu_0^2 r^2 d_s, \\ &\frac{128}{3}\log\frac{2(d_p - d_s r^2 + 2r^3)}{\psi_2}(\gamma^2\mu_{xyz}^3 d_p - \gamma\mu_{xyz} r^2 d_s + 2r^3)\Big\}\end{aligned}$$

From the above theorem, we can easily see that for $d_1 \asymp d_2 \asymp d_3 \asymp d$ and fixed $\{\psi_1, \psi_2, \gamma, \mu_{xyz}, \mu_0\}$, the sample complexity for LISTAI reduces to $O(\max\{r^2 d\log(r^2 d), (d^3 - r^2 d + r^3)\log(d^3 - r^2 d + r^3)\})$. Furthermore, if we assume $n_1 \asymp n_2 \asymp n_3 \asymp n$ and $d^3\log d$ has smaller order than $n^{3/2}\text{polylog}(n)$, LISTAI requires a smaller number of observed entries with

the benefit of incorporating auxiliary information, compared to tensor completion via convex relaxation with sum of nuclear norms [Tomioka et al., 2011] requiring $O(rn^2)$ observed entries, tensor factorization with alternating least squares [Jain and Oh, 2014] requiring $O(r^5 n^{3/2}(\log n)^4)$ observed entries, and tensor completion via nuclear norm minimization [Yuan and Zhang, 2016] with $O(\sqrt{r}n^{3/2}\text{polylog}(n))$ observed entries. Another interesting finding is that, by definition, γ is determined by the smallest singular value of the auxiliary information matrices. Thus, it is clearly observed from Theorem 6 as γ increases, the sample complexity also increases, indicating that more observed entries are required to achieve exact recovery if the smallest singular values of auxiliary information matrices are small. Intuitively speaking, this result makes sense since auxiliary information matrices with small singular values means limited information provided to uncover the latent structure of tensor observation. Under the setting that auxiliary information can fully describe the latent structure of $\mathcal{R}$, auxiliary information and observed entries in $\mathcal{R}$ work as two reciprocal forces. When the ability of auxiliary information matrices to recover the missing entries gets weak, more observed entries can guarantee a reasonable recovery performance.

4.3 Sample complexity for recovery with corrupted auxiliary information

The sample complexity analysis has been investigated in detail under the scenario when the auxiliary information matrices are full column rank. In many situations, this assumption may not be satisfied, which means the auxiliary information matrices $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ are not full column rank or $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ do not fully describe the latent structure of tensor. In this section, we analyze the sample complexity of the model to achieve reasonable recovery when $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ are not full column rank or their rank is difficult to obtain.

Assume the true $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ are contaminated by some noise terms $\Delta\mathbf{X}, \Delta\mathbf{Y}, \Delta\mathbf{Z}$ respectively, or in other words, we could only obtain corrupted version of auxiliary information

matrices, $\mathbf{X}+\Delta\mathbf{X}, \mathbf{Y}+\Delta\mathbf{Y}, \mathbf{Z}+\Delta\mathbf{Z}$ in practice. Transferring the sparsity on $\mathcal{G}$ and low rank constrain on $\mathcal{Q}$ into hard constraints, the optimization problem will be

$$\begin{aligned} \min_{\mathcal{Q},\mathcal{G}} \ & \|P_\Omega([\mathcal{G}; \mathbf{X}+\Delta\mathbf{X}, \mathbf{Y}+\Delta\mathbf{Y}, \mathbf{Z}+\Delta\mathbf{Z}] - \mathcal{R})\|_F^2 \\ \text{s.t.} \ & \mathcal{Q} - [\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}] \in B_\phi(0) \\ & \|\mathcal{G}\|_1 \le \alpha_1, \|\mathcal{Q}\|_* \le \alpha_2 \end{aligned} \tag{4.5}$$

For convenience, let $\Theta = \{(\mathcal{G}, \mathcal{Q}) | \mathcal{Q} - [\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}] \in B_\phi(0), \|\mathcal{G}\|_1 \le \alpha_1, \|\mathcal{Q}\|_* \le \alpha_2\}$ be the feasible solution set and $\theta \in \Theta$ be any feasible solution, where $B_\phi(0)$ is a small ball with radius ϕ at center 0. Also, let $f_\theta(i,j,k) = [\mathcal{G}; \mathbf{x}_i, \mathbf{y}_j, \mathbf{z}_k]$ be the estimation function for $\mathcal{R}_{i,j,k}$ parametrized by θ and $F_\Theta = \{f_\theta | \theta \in \Theta\}$ be the set of feasible solutions. Based on statistical learning theory, we are interested in the following two ℓ-risk quantities:

- Expected ℓ-risk: $\text{Risk}(f) = \mathbb{E}_{i,j,k}(f(i,j,k) - \mathcal{R}_{i,j,k})^2$
- Empirical ℓ-risk: $\widehat{\text{Risk}}(f) = \frac{1}{|\Omega|}\sum_{(i,j,k)\in\Omega}(f(i,j,k) - \mathcal{R}_{i,j,k})^2$

Thus, (4.5) is set to solve for θ^* that parametrizes $f^* = \arg\min_{f\in F_\Theta} \widehat{\text{Risk}}(f)$ and the next theorem shows that with bounded perturbation, the recovery can be attained if $\text{Risk}(f)$ approaches zero for large enough dimension.

Theorem 7. *Suppose* $\|\mathcal{G}\|_1 \le \alpha_1, \|\mathcal{Q}\|_* \le \alpha_2$, $\|\mathcal{Q} - [\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}]\|_F \le \phi$ *and the auxiliary information matrices are perturbed by noise terms* $\Delta\mathbf{X}, \Delta\mathbf{Y}, \Delta\mathbf{Z}$ *which are uniformly bounded by* $\Delta_{\max} = \max\{\|\Delta\mathbf{X}\|_F, \|\Delta\mathbf{Y}\|_F, \|\Delta\mathbf{Z}\|_F\}$, *we have the expected* ℓ*-risk* $\text{Risk}(f)$ *can be bounded by any arbitrary small* $\varepsilon > 0$ *with* $O\Big(\Big[\max\Big\{(\alpha_2+\phi)^2, \alpha_1^2\mu_{xyz}^3\Big\}\log n_{\max} + \alpha_1 d_1 d_2 d_3 \log d_{\max}\Delta_{\max}^3\Big]/\varepsilon^2\Big)$ *samples.*

Obviously, α_1 and α_2 affect the sample complexity when the auxiliary information matrices are corrupted. Intuitively, large α_2 indicates the low rank structure of $\mathcal{Q}$ is relatively complicated, which leads to more observations needed for obtaining better recovery of the

whole tensor. Similarly, lower ℓ_1 norm of $\mathcal{G}$, i.e., more sparsity in $\mathcal{G}$, representing fewer interaction terms among the modes of $\mathcal{R}$, implies more samples are required to reveal the latent structure of $\mathcal{R}$. Besides, theorem 7 suggests that the sample complexity of our model depends on the bound of auxiliary information perturbation $\Delta_{\max}$. Assume $d_1 \asymp d_2 \asymp d_3 \asymp d$, if auxiliary information are perfect, i.e., $\Delta_{\max} = O(1)$, our result suggests only $O(d^3 \log d)$ samples are required for recovery, which matches the result in the previous section provided $\log(n) = O(d^3 \log d)$. We also compare our sample complexity result with alternatives. If $\Delta_{\max} = O(\sqrt{n})$, or in other words, the perturbation is comparable to true 'signal' level, our result shows that $O(n^{3/2}\text{polylog}(n))$ samples are required, which is consistent with sample size requirement in terms of tensor completion method without auxiliary information [Jain and Oh, 2014, Yuan and Zhang, 2016, Ibriga and Sun, 2021, Xia et al., 2021].

4.4 Nested double ADMM algorithm

To allow for the existence of Gaussian noise when using $[\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}]$ to approximate $\mathcal{Q}$, the constraint $[\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}] = \mathcal{Q}$ is relaxed as minimizing their difference squared residual error and the optimization 4.4 can be formulated as a convex optimization problem

$$\begin{aligned} \min_{\mathcal{Q},\mathcal{G}} \ & \frac{1}{2}\|[\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}] - \mathcal{Q}\|_F^2 + \lambda_G \|\mathcal{G}\|_1 + \lambda_Q \|\mathcal{Q}\|_*, \\ \text{s.t.} \ & P_\Omega(\mathcal{Q}) = P_\Omega(\mathcal{R}) \end{aligned} \tag{4.6}$$

As stated before, we will use sum of nuclear norm as an approximation since there has been no practical algorithm to calculate tensor nuclear norm. Another type of approximation for tensor nuclear norm is also considered, which is the maximum value of nuclear norm for matricization of tensor along three mode and is polynomial-time computable [Friedland and Lim, 2018]. We include the comparison of those two approximations in Appendix.

With sum of nuclear norm, we aim to solve

$$\begin{aligned}
&\min_{\mathcal{Q},\mathcal{G}} \ \frac{1}{2}\|[\mathcal{G};\mathbf{X},\mathbf{Y},\mathbf{Z}]-\mathcal{Q}\|_F^2+\lambda_G\|\mathcal{G}\|_1+\frac{\lambda_Q}{3}(\|\mathcal{Q}_{(1)}\|_*+\|\mathcal{Q}_{(2)}\|_*+\|\mathcal{Q}_{(3)}\|_*),\\
&\text{s.t. } P_\Omega(\mathcal{Q})=P_\Omega(\mathcal{R})
\end{aligned}$$

which is in the format of convex optimization and inspires us to use the alternating direction method of multipliers (ADMM) algorithm to solve it. Recall that the usual ADMM is applied to solve problems with separable blocks of variables, we first introduce an auxiliary variable $\mathcal{F}$ which is equivalent to $[\mathcal{G};\mathbf{X},\mathbf{Y},\mathbf{Z}]-\mathcal{Q}$ and write down the optimization problem as

$$\begin{aligned}
&\min_{\mathcal{Q},\mathcal{G}} \ \frac{1}{2}\|\mathcal{F}\|_F^2+\lambda_G\|\mathcal{G}\|_1+\frac{\lambda_Q}{3}(\|\mathcal{Q}_{(1)}\|_*+\|\mathcal{Q}_{(2)}\|_*+\|\mathcal{Q}_{(3)}\|_*),\\
&\text{s.t. } P_\Omega(\mathcal{Q})=P_\Omega(\mathcal{R}) \text{ and } \mathcal{F}=[\mathcal{G};\mathbf{X},\mathbf{Y},\mathbf{Z}]-\mathcal{Q}
\end{aligned} \tag{4.7}$$

Thus the augmented Lagrangian function is

$$\begin{aligned}
\mathcal{L}_\beta(\mathcal{G},\mathcal{Q},\mathcal{F},\mathcal{M}_1,\mathcal{M}_2)=&\frac{1}{2}\|\mathcal{F}\|_F^2+\lambda_G\|\mathcal{G}\|_1+\frac{\lambda_Q}{3}(\|\mathcal{Q}_{(1)}\|_*+\|\mathcal{Q}_{(2)}\|_*+\|\mathcal{Q}_{(3)}\|_*)\\
&+\Big\langle\mathcal{M}_1,\mathcal{F}-([\mathcal{G};\mathbf{X},\mathbf{Y},\mathbf{Z}]-\mathcal{Q})\Big\rangle+\Big\langle\mathcal{M}_2,P_\Omega(\mathcal{Q}-\mathcal{R})\Big\rangle\\
&+\frac{\beta}{2}\|\mathcal{F}-([\mathcal{G};\mathbf{X},\mathbf{Y},\mathbf{Z}]-\mathcal{Q})\|_F^2+\frac{\beta}{2}\|P_\Omega(\mathcal{Q}-\mathcal{R})\|_F^2
\end{aligned} \tag{4.8}$$

For notational convenience we will drop notation for dependence on $\mathcal{L}_\beta(\mathcal{G},\mathcal{Q},\mathcal{F},\mathcal{M}_1,\mathcal{M}_2)$ and use $\mathcal{L}$ as a substitution. Following the framework of ADMM algorithm, we can construct

the following algorithm to solve 4.7,

$$\begin{aligned}
\mathcal{F}^{t+1} &= \arg\min_{\mathcal{F}} \mathcal{L}_\beta(\mathcal{G}^t, \mathcal{Q}^t, \mathcal{F}, \mathcal{M}_1^t, \mathcal{M}_2^t) \\
\mathcal{G}^{t+1} &= \arg\min_{\mathcal{G}} \mathcal{L}_\beta(\mathcal{G}, \mathcal{Q}^t, \mathcal{F}^{t+1}, \mathcal{M}_1^t, \mathcal{M}_2^t) \\
\mathcal{Q}^{t+1} &= \arg\min_{\mathcal{Q}} \mathcal{L}_\beta(\mathcal{G}^{t+1}, \mathcal{Q}, \mathcal{F}^{t+1}, \mathcal{M}_1^t, \mathcal{M}_2^t) \\
\mathcal{M}_1^{t+1} &= \mathcal{M}_1^t + \beta(\mathcal{F}^{t+1} - ([\mathcal{G}^{t+1}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}] - \mathcal{Q}^{t+1})) \\
\mathcal{M}_2^{t+1} &= \mathcal{M}_2^t + \beta P_\Omega(\mathcal{Q}^{t+1} - \mathcal{R}^{t+1})
\end{aligned}$$

Obviously, the iterative step for updating $\mathcal{Q}$ goes back to the original tensor completion problem with sum of nuclear norm as regularizer without auxiliary information matrices. Thus, we implement another inner loop of ADMM to update $\mathcal{Q}$ and that is the reason why we call our algorithm nested double ADMM. We now derive the closed form solution for updating $\mathcal{F}, \mathcal{Q}, \mathcal{G}$.

4.4.1 Algorithm updating steps

Updating $\mathcal{F}^{t+1}$ We take the derivative of $\mathcal{L}$ with respect to $\mathcal{F}$ and obtain

$$\frac{\nabla \mathcal{L}}{\nabla \mathcal{F}} = \frac{\nabla\left(\frac{1}{2}\|\mathcal{F}\|_F^2 + \left\langle \mathcal{M}_1, \mathcal{F} \right\rangle + \frac{\beta}{2}\|\mathcal{F} - ([\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}] - \mathcal{Q})\|_F^2\right)}{\nabla \mathcal{F}}$$

By setting it to be 0, we have

$$\mathcal{F}^{t+1} = \frac{\beta}{\beta + 1}([\mathcal{G}^t; \mathbf{X}, \mathbf{Y}, \mathbf{Z}] - \mathcal{Q}^t - \frac{\mathcal{M}_1}{\beta})$$

Updating $\mathcal{G}^{t+1}$ Let's write down the augmented Lagrangian that are related to $\mathcal{G}$,

$$\begin{aligned}\mathcal{G}^{t+1} =& \arg\min_{\mathcal{G}} \lambda_G \|\mathcal{G}\|_1 + \left\langle \mathcal{M}_1^t, \mathcal{F}^{t+1} - ([\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}] - \mathcal{Q}^t) \right\rangle \\ &+ \frac{\beta}{2} \|\mathcal{F}^{t+1} - ([\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}] - \mathcal{Q}^t)\|_F^2 \\ =& \arg\min_{\mathcal{G}} \lambda_G \|\mathcal{G}\|_1 + \frac{\beta}{2} \|\mathcal{F}^{t+1} + \frac{\mathcal{M}_1^t}{\beta} - [\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}] + \mathcal{Q}^t\|_F^2\end{aligned} \tag{4.9}$$

which is equivalent to updating $\mathcal{G}$ in a vectorized form,

$$\text{vec}(\mathcal{G}) = \arg\min_{\text{vec}(\mathcal{G})} \lambda_G \|\text{vec}(\mathcal{G})\|_1 + \frac{\beta}{2} \|\text{vec}(\mathcal{F}^{t+1} + \frac{\mathcal{M}_1^t}{\beta} + \mathcal{Q}^t) - (\mathbf{Z} \otimes \mathbf{Y} \otimes \mathbf{X})\text{vec}(\mathcal{G})\|_2^2$$

The above problem is similar to 'LASSO' and we can implement any well-developed algorithm [Efron et al., 2004] to solve it. Considering that LASSO does not have a closed form solution and has to be solved iteratively in practice, to facilitate our convergence analysis, we do linearization on steps to update $\mathcal{G}$. Since $1/2\|(\mathbf{Z}\otimes\mathbf{Y}\otimes\mathbf{X})\text{vec}(\mathcal{G}) - \text{vec}(\mathcal{F}^{t+1} + \frac{\mathcal{M}1^t}{\beta} + \mathcal{Q}^t)\|_2^2 \approx 1/2\|(\mathbf{Z}\otimes\mathbf{Y}\otimes\mathbf{X})\text{vec}(\mathcal{G}^t) - \text{vec}(\mathcal{F}^{t+1} + \frac{\mathcal{M}_1^t}{\beta} + \mathcal{Q}^t)\|_2^2 + \langle f_3^t, \text{vec}(\mathcal{G}) - \text{vec}(\mathcal{G}^t)\rangle + \rho_1/2\|\text{vec}(\mathcal{G}) - \text{vec}(\mathcal{G}^t)\|_2^2$, where f_3^t is the gradient of $1/2\|(\mathbf{Z}\otimes\mathbf{Y}\otimes\mathbf{X})\text{vec}(\mathcal{G}) - \text{vec}(\mathcal{F}^{t+1} + \frac{\mathcal{M}_1^t}{\beta} + \mathcal{Q}^t)\|_2^2$ at $\text{vec}(\mathcal{G}^t)$, i.e., $f_3^t = (\mathbf{Z}\otimes\mathbf{Y}\otimes\mathbf{X})^\top[(\mathbf{Z}\otimes\mathbf{Y}\otimes\mathbf{X})\text{vec}(\mathcal{G}) - \text{vec}(\mathcal{F}^{t+1} + \frac{\mathcal{M}_1^t}{\beta} + \mathcal{Q}^t)]$, we have the approximation for the step to update $\mathcal{G}$

$$\text{vec}(\mathcal{G})^{t+1} = \arg\min_{\text{vec}(\mathcal{G})} \lambda_G \|\text{vec}(\mathcal{G})\|_1 + \frac{\beta\rho_1}{2} \|\text{vec}(\mathcal{G}) - \text{vec}(\mathcal{G}^t) + \frac{f_3^t}{\rho_1}\|_2^2$$

The closed form solution is

$$\text{vec}(\mathcal{G})^{t+1} = \max(0, |\text{vec}(\mathcal{G})^t - f_3^t/\rho_1| - \lambda_G/(\rho_1\beta))\text{sign}(\text{vec}(\mathcal{G})^t - f_3^t/\rho_1)$$

Updating $\mathcal{Q}^{t+1}$ Solving the augmented Lagrangian that are related to $\mathcal{Q}$, we have,

$$\begin{aligned}
\mathcal{Q}^{t+1} =& \arg\min_{\mathcal{Q}} \frac{\lambda_Q}{3}(\|\mathcal{Q}_{(1)}\|_* + \|\mathcal{Q}_{(2)}\|_* + \|\mathcal{Q}_{(3)}\|_*) \\
&+ \left\langle \mathcal{M}_1^t, \mathcal{F}^{t+1} - ([\mathcal{G}^{t+1}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}] - \mathcal{Q}) \right\rangle + \left\langle \mathcal{M}_2^t, P_\Omega(\mathcal{Q} - \mathcal{R}) \right\rangle \\
&+ \frac{\beta}{2}\|\mathcal{F}^{t+1} - ([\mathcal{G}^{t+1}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}] - \mathcal{Q})\|_F^2 + \frac{\beta}{2}\|P_\Omega(\mathcal{Q} - \mathcal{R})\|_F^2 \\
=& \arg\min_{\mathcal{Q}} \frac{\lambda_Q}{3}\sum_{i=1}^{3}\|\mathcal{Q}_{(i)}\|_* + \frac{\beta}{2}\|\mathcal{F}^{t+1} + \frac{\mathcal{M}_1^t}{\beta} - ([\mathcal{G}^{t+1}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}] - \mathcal{Q})\|_F^2 \\
&+ \frac{\beta}{2}\|P_\Omega(\mathcal{Q} + \frac{\mathcal{M}_2^t}{\beta} - \mathcal{R})\|_F^2
\end{aligned}$$

Different from the original tensor completion without auxiliary information, we are going to use linearization to combine the last two terms, where f_1, f_2 are the gradients of $1/2\|\mathcal{F}^{t+1} + \frac{\mathcal{M}_1^t}{\beta} - ([\mathcal{G}^{t+1}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}] - \mathcal{Q})\|_F^2$ and $1/2\|P_\Omega(\mathcal{Q} + \frac{\mathcal{M}_2^t}{\beta} - \mathcal{R})\|_F^2$ at $\mathcal{Q}^t$, i.e., $f_1 = \mathcal{Q}^t - ([\mathcal{G}^{t+1}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}] - \mathcal{F}^{t+1} - \frac{\mathcal{M}_1^t}{\beta})$ and $f_2 = P_\Omega(\mathcal{Q}^t - \mathcal{R} + \frac{\mathcal{M}_2^t}{\beta})$,

$$\begin{aligned}
\mathcal{Q}^{t+1} =& \arg\min_{\mathcal{Q}} \frac{\lambda_Q}{3}\sum_{i=1}^{3}\|\mathcal{Q}_{(i)}\|_* + \frac{\beta\rho_2}{2}\|\mathcal{Q} - (\mathcal{Q}^t - f_1/\rho_2)\|_F^2 + \frac{\beta\rho_2}{2}\|\mathcal{Q} - (\mathcal{Q}^t - f_2/\rho_2)\|_F^2 \\
=& \arg\min_{\mathcal{Q}} \frac{\lambda_Q}{3}\sum_{i=1}^{3}\|\mathcal{Q}_{(i)}\|_* + \beta\rho_2\|\mathcal{Q} - (\mathcal{Q}^t - (f_1 + f_2)/2\rho_2)\|_F^2 \qquad (4.10)
\end{aligned}$$

To solve problem 4.10, we implement the inner ADMM iteration for nested double ADMM algorithm. Problem 4.10 can be transferred as

$$\begin{aligned}
\min_{\mathcal{Q}, \{\mathbf{Q}_i\}_{i=1}^3} \quad & \frac{\lambda_Q}{3}\sum_{i=1}^{3}\|\mathbf{Q}_i\|_* + \beta\rho_2\|\mathcal{Q} - (\mathcal{Q}^t - (f_1 + f_2)/2\rho_2)\|_F^2, \\
\text{s.t.} \quad & \mathcal{Q}_{(i)} = \mathbf{Q}_i \qquad (4.11)
\end{aligned}$$

which is equivalent to

$$\begin{aligned}\min_{\mathcal{Q},\{\text{vec}(\mathbf{Q}_i)\}_{i=1}^3} \quad & \frac{\lambda_Q}{3}\sum_{i=1}^3 \|\mathbf{Q}_i\|_* + \beta\rho_2\|\text{vec}(\mathcal{Q} - (\mathcal{Q}^t - (f_1+f_2)/2\rho_2))\|_2^2, \\ \text{s.t.} \quad & \tilde{P}_i\text{vec}(\mathcal{Q}) = \text{vec}(\mathbf{Q}_i)\end{aligned}$$

where $\mathbf{Q}_i \in \mathbb{R}^{n_i \times \prod_{j\neq i} n_j}$ is an auxiliary matrix of the same size as the mode i unfolding of $\mathcal{Q}$ and $\tilde{P}_i$ is the matrix representation of mode i unfolding. It's easy to show that $\tilde{P}_i$ is a permutation matrix which satisfies $\tilde{P}_i^\top \tilde{P}_i = \mathbf{I}$. Therefore, we can derive its augmented Lagrangian function as

$$\begin{aligned}\mathcal{L}_\eta(\mathcal{Q},\{\mathbf{Q}_i\}_{i=1}^3) =& \frac{\lambda_Q}{3}\sum_{i=1}^3 \|\mathbf{Q}_i\|_* + \beta\rho_2\|\text{vec}(\mathcal{Q} - (\mathcal{Q}^t - (f_1+f_2)/2\rho_2))\|_2^2, \\ &+ \sum_{i=1}^3 \left\{\eta\boldsymbol{\alpha}_i^\top(\tilde{P}_i\text{vec}(\mathcal{Q}) - \text{vec}(\mathbf{Q}_i)) + \frac{\eta}{2}\|\tilde{P}_i\text{vec}(\mathcal{Q}) - \text{vec}(\mathbf{Q}_i)\|_2^2\right\}\end{aligned}$$

Note that we rescaled the Lagrangian multiplier vector $\boldsymbol{\alpha}$ by the factor η for the sake of notational simplicity. Thus, we can iteratively update $\mathcal{Q}, \{\mathbf{Q}_i\}_{i=1}^3, \{\boldsymbol{\alpha}_i\}_{i=1}^3$ for solving problem 4.10.

To update $\mathcal{Q}$, we take the derivative of $\mathcal{L}_\eta(\mathcal{Q},\{Q_i\}_{i=1}^3)$ with respect to $\text{vec}(\mathcal{Q})$ and set it to be 0,

$$\begin{aligned}\frac{\nabla\mathcal{L}_\eta(\mathcal{Q},\{\mathbf{Q}_i\}_{i=1}^3)}{\nabla\text{vec}(\mathcal{Q})} =& 2\beta\rho_2(\text{vec}(\mathcal{Q} - (\mathcal{Q}^t - (f_1+f_2)/2\rho_2))) \\ &+ \sum_{i=1}^3 \left\{\eta\tilde{P}_i^\top\boldsymbol{\alpha}_i + \eta\tilde{P}_i^\top\tilde{P}_i\text{vec}(\mathcal{Q}) - \eta\tilde{P}_i^\top\text{vec}(\mathbf{Q}_i)\right\} \\ \text{vec}(\mathcal{Q}^{s+1}) =& \frac{2\beta\rho_2\text{vec}(\mathcal{Q}^t - (f_1+f_2)/2\rho_2)) + \sum_{i=1}^3\left\{\eta\tilde{P}_i^\top(\text{vec}(\mathbf{Q}_i^s)) - \boldsymbol{\alpha}_i^s\right\}}{2\beta\rho_2 + 3\eta}\end{aligned}$$

The remaining step to get estimate update for $\mathcal{Q}$ will be folding $\text{vec}(\mathcal{Q})$ into $\mathbb{R}^{n_1\times n_2\times n_3}$ tensor.

Updating $\{\mathbf{Q}_i\}_{i=1}^3$ are similar to matrix completion problem with nuclear norm regularization. We aim to solve

$$\begin{aligned}\mathbf{Q}_i^{s+1} &= \arg\min_{\mathbf{Q}_i} \frac{\lambda_Q}{3}\|\mathbf{Q}_i\|_* + \frac{\eta}{2}\|\tilde{P}_i \mathrm{vec}(\mathcal{Q})^{s+1} - \mathrm{vec}(\mathbf{Q}_i) + \boldsymbol{\alpha}_i^s\|_2^2 \\ &= \arg\min_{\mathbf{Q}_i} \frac{\lambda_Q}{3}\|\mathbf{Q}_i\|_* + \frac{\eta}{2}\|\mathcal{Q}_{(i)}^{s+1} - \mathbf{Q}_i + \mathbf{A}_i^s\|_F^2\end{aligned}$$

where $\mathbf{A}_i \in \mathbb{R}^{n_i \times \prod_{j\neq i} n_j}$ represents the matrix folded by $\boldsymbol{\alpha}_i$. Then by spectral soft thresholding algorithm, we can obtain

$$\mathbf{Q}_i^{s+1} = \mathrm{prox}_{\lambda_Q/(3\eta)}(\mathcal{Q}_{(i)}^{s+1} + \mathbf{A}_i^s)$$

where $\mathrm{prox}_\lambda(\mathbf{Y}) = \mathbf{U}\max(\mathbf{S} - \lambda, 0)\mathbf{V}^\top$ ($\mathbf{Y} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$ is the SVD of input matrix $\mathbf{Y}$ and the max operation is taken elementwise) is the spectral soft thresholding operation which can be considered as a shrinkage operation on the singular values. There are lots of existing work in matrix completion ([Mazumder et al., 2010, Hastie et al., 2015]) discussing how to solve it in a more efficient way.

Updating $\boldsymbol{\alpha}_i$ is conducted by performing $\boldsymbol{\alpha}_i^{s+1} = \boldsymbol{\alpha}_i^s + \mathcal{Q}_{(i)}^{s+1} - \mathbf{Q}_i^{s+1}$.

We perform inner ADMM loop for $s = 1, 2, ...$ until certain convergence criterion is satisfied and later we will be able to get the final iterative update for $\mathcal{Q}^{t+1}$.

4.4.2 Convergence analysis

The convergence of nested double ADMM algorithm mildly depends on the choice of the step-size β, η and the scale parameters ρ_1, ρ_2. Stated in the following Theorem 8, we will set the step size β to be updated at each iteration to be $\beta_{t+1} = \min\{\beta_{\max}, \beta_0\beta_t\}, \forall t \in \{1, 2, ...\}$, where $\beta_0 \geq 1$ is a scale parameter to guarantee $\{\beta_t\}$ is a non-decreasing sequence and furthermore, $\beta_{\max}$ controls the upper bound for $\{\beta_t\}$. For η, recall that we require the

inner ADMM iteration is invariant to scalar multiplication of the objective function 4.11. More precisely, when the input $\mathcal{Q}^t - (f_1 + f_2)/2\rho_2$ is multiplied by a constant c and the regularization ρ_2 is multiplied by a constant $1/c$, the solution of the minimization (4.11) should remain essentially the same as the original problem, except that the solution $\mathcal{Q}$ is also multiplied by the constant c. In order to make the algorithm follow the same path (instead of $\mathcal{Q}^s$,$\{\mathbf{Q}_i^s\}_{i=1}^3$, $\{\boldsymbol{\alpha}_i\}_{i=1}^3$ are all multiplied by c), we need to scale η inversely proportional to c and as a result, the last two terms in the augmented Lagrangian scale linearly if η scale inversely to c. Therefore, we choose $\eta = \eta_0/\text{std}(\mathcal{Q}^t - (f_1 + f_2)/2\rho_2)$ where η_0 is a constant and $\text{std}(\mathcal{Q}^t-(f_1+f_2)/2\rho_2)$ is the standard deviation of $\mathcal{Q}^t-(f_1+f_2)/2\rho_2$. Then by choosing appropriate ρ_1, ρ_2, we have the following convergence property.

Theorem 8. *Let* $P_{\mathcal{G}}(\mathcal{G}) = \begin{bmatrix} -[\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}] \\ 0 \end{bmatrix}$, $P_{\mathcal{Q}}(\mathcal{Q}) = \begin{bmatrix} \mathcal{Q} \\ P_\Omega(\mathcal{Q}) \end{bmatrix}$ *be the operators which satisfies* $\rho_1 \geq \|P_{\mathcal{G}}\|^2, \rho_2 \geq \|P_{\mathcal{Q}}\|^2$. *If the step size parameter* β_t *is non-decreasing and upper bounded, the sequence derived from nested double ADMM algorithm* $(\mathcal{G}^t, \mathcal{Q}^t, \mathcal{F}^t, \mathcal{M}_1^t, \mathcal{M}_2^t)$ *will converge to a global minimizer of 4.7.*

Algorithm 8 guarantees that the nested double ADMM algorithm will converge to the global optimum of convex approximation of LISTAI with appropriate step size choice.

4.5 Data Analysis

In this section, we will use both synthetic (simulated) datasets and a real world dataset to examine the efficacy of the proposed nested double ADMM algorithm. Specifically, we perform a simulation study to show that nested double ADMM algorithm recovers the low rank tensors under different quality of auxiliary information, either perfect or corrupted. We also compare nested double ADMM algorithm to an alternating minimization method for tensor completion proposed in Jain and Oh [2014] (Tensor-ALS) that is similar in spirit to various generalizations of tensor decomposition, clustering and completion algorithms in

tensor related literatures. Different from our model, Tensor-ALS is originally designed to impute missing values without incorporating auxiliary information. We believe this comparison is useful for understanding how the auxiliary information will facilitate recovering the latent structure of partially observed tensor dataset. In addition, we consider implementing nested double ADMM algorithm on different types of tensor as well, cubic or rectangular tensors. Its good performance also validates the stability of our algorithm under different dimensional tensor settings.

To evaluate the quality of completion performance, we choose metrics which are commonly used in matrix/tensor completion literatures: the average root mean squared error (RMSE) calculated on the observed entries (training error) and missing entries (testing error). The first one measures the distance between observed entries and estimated entries by implementing nested double ADMM algorithm. The second one is critical to assess how close the imputation and the true values will be.

- Training error:$\|P_{\Omega}([\mathcal{G};\mathbf{X},\mathbf{Y},\mathbf{Z}]-\mathcal{R})\|_F/\sqrt{|\Omega|}$
- Testing error: $\|P_{\Omega^c}([\mathcal{G};\mathbf{X},\mathbf{Y},\mathbf{Z}]-\mathcal{R})\|_F/\sqrt{|\Omega^c|}$

Ω^c represents the compliment of Ω, i.e., the index set containing all missing entries in $\mathcal{R}$. We run each experiment 10 times and report the average value with corresponding standard error in parenthesis.

4.5.1 Synthetic Data

To determine the efficacy of our model and method, we create synthetic datasets as follows. The sparse core tensor $\mathcal{G}$ is generated under the following framework: all the entries are generated from standard Gaussian distribution and the zero entries in $\mathcal{G}$ are randomly selected with probability $p_{\text{sparse}} = 0.6$. The entries in $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ are randomly generated from standard Gaussian distribution as well. The tensor observation $\mathcal{R}$ is created by $[\mathcal{G};\mathbf{X},\mathbf{Y},\mathbf{Z}]+\mathcal{N}$ where

Parameters				Perfect auxiliary information	
n	d	p_{miss}	σ^2	Train	Test
20	5	0.8	0.01	0.0038(0.00081)	0.0933(0.01531)
40	5	0.8	0.01	0.0038(0.00081)	0.0283(0.00361)
20	5	0.9	0.01	0.0053(0.00016)	0.0989(0.01552)
20	5	0.8	10	0.0037(0.00007)	0.0960(0.0132)
20	10	0.8	0.01	0.0038(0.00081)	0.2576(0.01905)

Table 4.1: Recovery performance for nested double ADMM algorithm when auxiliary information are perfect

$\mathcal{N} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is a noise tensor with entries iid Gaussian with mean 0 and variance σ^2. We randomly sample p_{miss} of entries from $\mathcal{R}$ as missing entries. We present the results in this section by reporting the average performance across 10 repetitions under each simulation settings.

To get an initial evaluation for how the nested double ADMM algorithm performs at recovering true underlying latent structure with auxiliary information incorporated, we first consider a simple situation where the tensor is cubic and all the auxiliary information matrices have the same number of columns. In other words, we set the dimensions for tensor $\mathcal{R}$ are equal in three modes, $n_1 = n_2 = n_3 = n$ and all the three auxiliary information matrices have the same dimensions as well, $d_1 = d_2 = d_3 = d$. In the perfect setting, $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ are set to be matrices with dimension $n \times d$ and all the features are generated to guarantee they have full column rank. Instead, in the corrupted setting, $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ are set to be matrices with dimension $n \times (d+1)$ where the first d columns in $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ are the same as perfect settings and the last column is created by arbitrarily linear combinations of first d columns.

As we will show, we vary the choice of $n, d, p_{\text{miss}}, \sigma$ to validate the robustness of our method in achieving good recovery performance under different scenarios. The tuning parameters λ_G, λ_Q are chosen via cross validation process with the candidate ratio λ_G/λ_Q within $\{10^{-3}, ..., 10^4\}$. The hyperparameters that give the best average testing error are used.

Parameters				Corrupted auxiliary information	
n	d	p_{miss}	σ^2	Train	Test
20	6	0.8	0.01	0.0038(0.00012)	0.1059(0.01708)
40	6	0.8	0.01	0.0013(0.00021)	0.0393(0.00305)
20	6	0.9	0.01	0.0053(0.00014)	0.1073(0.01604)
20	6	0.8	10	0.0038(0.00081)	0.1345(0.03606)
20	11	0.8	0.01	0.0038(0.00591)	0.2826(0.03923)

Table 4.2: Recovery performance for nested double ADMM algorithm when auxiliary information are corrupted

As we can observe from Table 4.1, if we treat the first scenario ($n = 20, d = 5, p_{\text{miss}}, \sigma^2 = 0.1$) as a benchmark and compare it with other following scenarios, it is clear that our proposed algorithm achieves a good recovery performance. When the sample size increases, in other words, when n increases or p_{miss} decreases, our algorithm achieves a lower training error and testing error, which is consistent with our theoretical analysis. Moreover, larger d increases the difficulty in recovering the latent structure of $\mathcal{R}$, leading to a relatively larger training error and testing error. Our algorithm still obtains a good recovery performance even when there is a large noisy perturbation. As is shown, increasing the variation of noise tensor σ^2 from 0.1 to 10 does not result in a dramatic change in training/testing error. When the auxiliary information becomes corrupted, we can see that our algorithm still obtains reasonable good performance by comparing results provided in Table 4.1 and 4.2.

Next, we compare the performance of our proposed algorithm with an alternative method, Ten-ALS. As discussed, another interesting point is to figure out what benefit we can obtain by incorporating the auxiliary information. Table 4.3 displays the performance comparison between nested double ADMM algorithm and Ten-ALS, where the former employs the auxiliary information and the latter just conduct tensor completion without using any auxiliary information.

Comparing the result in Table 4.1 and 4.3, our algorithm outperforms Ten-ALS with a

Parameters				Our approach		Ten-ALS	
n	d	p_{miss}	σ^2	Train	Test	Train	Test
20	5	0.8	0.01	0.0038(0.00081)	0.0933(0.01531)	0.0548(0.00780)	0.6453(1.13046)
40	5	0.8	0.01	0.0038(0.00081)	0.0283(0.00361)	0.0212(0.00469)	0.0114(0.00237)
20	5	0.9	0.01	0.0053(0.00016)	0.0989(0.01552)	0.0699(0.01560)	0.5990(0.80196)
20	5	0.8	10	0.0037(0.00007)	0.0960(0.0132)	0.2184(0.00701)	0.2578(0.17239)
20	10	0.8	0.01	0.0038(0.00081)	0.2576(0.01905)	0.2874(0.02310)	0.2467(0.05153)

Table 4.3: Recovery performance comparison between nested double ADMM and Ten-ALS [Jain and Oh, 2014] with different dimensions, missing percentage, noise level and latent low rank structure when auxiliary information are perfect. Reported values are average based on 10 replications and standard error of recovery error are provided in parenthesis.

large gap in either training error or testing error. Whenever tensor size gets increased (n increases from 20 to 40) or the underlying low rank structure gets complicated (d increases from 5 to 10), our algorithm always achieves a better recovery performance. Surprisingly, when a large proportion of entries are missing, i.e., $p_{\text{miss}} = 0.8$, our algorithm still improves the recovery rate though the testing error for Ten-ALS is relatively high. Finally, an interesting phenomenon is that our algorithm is less sensitive to the noise perturbation. Keeping other factors fixed, if σ^2 increases to 10, both training error and testing error for Ten-ALS experience a significant jump, while the training error for our algorithm increases slightly, but is still pretty close to 0. Even there is a relatively large increase in testing error, its magnitude is still at an acceptable level.

Up to this point, we restricted our implementation to cubic tensors. However, in applications, it is more common to deal with rectangular tensors. Moreover, we simplify the latent structure for tensors in previous sections to having the same rank for each mode, implying that the number of features for different mode auxiliary information matrices are the same. Clearly, this is unrealistic since in most cases, the number of features we could obtain can vary for different modes. Taking those concerns into consideration, we perform two different trials for rectangular tensors. One is with balanced low rank structure along 3 modes, i.e., $d_1 = d_2 = d_3$ and another one generalizes that ranks along 3 modes are different. This

	Nested double ADMM		Ten-ALS	
(n_1, n_2, n_3)	(10,20,30)	(10,20,30)	(10,20,30)	(10,20,30)
(d_1, d_2, d_3)	(5,5,5)	(4,8,12)	(5,5,5)	(4,8,12)
p_{miss}	0.8	0.8	0.8	0.8
σ^2	0.1	0.1	0.1	0.1
Train	0.0043(0.00011)	0.0044(0.00008)	0.0510(0.001311)	0.1478(0.01077)
Test	0.0933(0.01957)	0.1822(0.03214)	0.1627(0.05917)	0.3112(0.27962)

Table 4.4: Rectangular tensor recovery performance comparison for proposed nested double ADMM algorithm and Ten-ALS Jain and Oh [2014] when auxiliary information are perfect.

allows us to compare how the performance will be affected by having imbalanced low rank structure.

Table 4.4 shows that nested double ADMM algorithm performs well and much better than Ten-ALS in terms of testing error at both balanced and imbalanced low rank structure for rectangular tensors. Not surprisingly, there is a slightly increase in both training and testing errors for imbalanced low rank structures. Another issue worth noting is that Ten-ALS employs the CP rank decomposition and that partially explains its worse performance when implemented to rectangular tensor with imbalanced low rank structure.

4.5.2 Real Data

UCLAF dataset is a real dataset introduced in Zheng et al. [2010] which collects 164 users GPS trajectories over 168 locations from April 2007 to October 2009 in the city of Beijing, China. As indicated in Liao et al. [2005], this is a challenging but interesting question: by using the user's trajectories and annotations, our goal is to make useful, targeted and personalized recommendation based on different characteristics of users and locations. In Zheng et al. [2010], the user activities are classified into 5 different types, 'Food & Drink', 'Shopping', 'Movies & Shows', 'Sports & Exercise' and 'Tourism & Amusement'. Thus, UCLAF has the dataset formulated as an order three tensor with dimension 164(users) $\times$ 168(locations) $\times$ 5(activities). Each entry (i, j, k) in the tensor will be recorded

as the frequency of user i performing activity k at location j. We separate imputing the missing values in UCLAF into 2 different scenarios and each of them will facilitate us to answer the following questions:

- For a specific user, if she/he visited some places in Beijing, what type of the activity recommendations can be provided?
- For a specific user, if she/he would like to hang out, or do tourist sightseeing or have lunch/dinner with friends, which locations/places can be recommended given previous GPS trajectories?

To differentiate the two scenarios, a user-location matrix is provided to model the case when we only know a user has visited a certain place but have no idea what the user was doing there. In other words, for entry (i, j) in user location matrix with nonzero value, if there are no activities recorded in the corresponding tensor fiber, $\mathcal{R}_{i,j,:} \in \mathbb{R}^5$, we will treat the whole 5 entries in $\mathcal{R}_{i,j,:}$ as missing values. That's the first scenario and on the other hand, if the whole fiber $\mathcal{R}_{i,j,:}$ are all zero together with a zero entry (i, j) in user location matrix, we are under the second scenario.

In addition to the travel GPS trajectories, we have auxiliary information which can enhance our imputation performance. For users, we have a 164×164 similarity matrix obtained from social network analysis. This helps to give a more user-specified location/activity recommendation based on similar users GPS history. For locations, we have a 168×18 feature matrix with each column referring to the (normalized) number of point of interests such as museums, shopping centers or theatres. Moreover, a 5×5 activity similarity matrix is also provided to represent the activity-activity correlations, indicating how likely one activity is to happen if another activity happens.

To achieve the goal of dimension reduction, we implement the Principal Component Analysis (PCA) to pick useful 'feature components' for users and activities. Consequently, we choose the top 4 principle components of user-user similarity matrix and the top 2 principal

components of activity-activity similarity matrix. (More details about PCA analysis can be found in Appendix C.5.)

Considering the fact the UCLAF dataset has non-negative entries, we implement the centering strategy, removing means for each mode. This is widely utilized in papers dealing with recommendation system challenges. For example, Olshen and Rajaratnam [2010] implements a centering and scaling algorithm for complete data and analyze its convergence property. Hastie et al. [2015] proposes a similar centering and scaling scheme for incomplete data. Since UCLAF has imbalanced data, we will only adopt the idea of centering in Hastie et al. [2015] and generalize to tensor case by learning the centering parameters for the fibers along three modes. We assume order three tensor $\mathcal{R} = \{\mathcal{R}_{i,j,k}\} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ follows the centering model where the fibers along three modes have mean zero simultaneously, or in other words, $\mathcal{R}_{i,j,k}$ follows a distribution with mean $\mu_{i,j,k}$ where

$$\mu_{i,j,k} = \alpha_i + \beta_j + \gamma_k$$

$\alpha_i, \beta_j, \gamma_k$ represents the fibers along the first mode, second mode and third mode respectively. Given those parameters, each observation is centered via

$$\tilde{\mathcal{R}}_{i,j,k} = \mathcal{R}_{i,j,k} - \alpha_i - \beta_j - \gamma_k$$

The idea for estimation procedure is quite simple. The method of moments will be adopted by writing down the estimating equations that demand the transformed observed data have mean zero and these equations will be solved iteratively until a certain termination condition is satisfied.

Consider the estimation equation for the mode-1 fiber mean,

$$\frac{1}{|\Omega_i|}\sum_{(j,k)\in\Omega_i}\tilde{\mathcal{R}}_{i,j,k}=\frac{1}{|\Omega_i|}\sum_{(j,k)\in\Omega_i}(\mathcal{R}_{i,j,k}-\alpha_i-\beta_j-\gamma_k),\quad \forall i\in[n_1]$$

where $\Omega_i=\{(j,k)|(i,j,k)\in\Omega\}$. By setting the above expression equals zero, we can obtain

$$\alpha_i=\frac{1}{|\Omega_i|}\sum_{(j,k)\in\Omega_i}(\mathcal{R}_{i,j,k}-\beta_j-\gamma_k),\quad \forall i\in[n_1]$$

Analogously, the mean for mode-2, mode-3 fiber can be estimated through

$$\beta_j=\frac{1}{|\Omega_i|}\sum_{(j,k)\in\Omega_i}(\mathcal{R}_{i,j,k}-\alpha_i-\gamma_k),\quad \forall j\in[n_2]$$
$$\gamma_k=\frac{1}{|\Omega_i|}\sum_{(j,k)\in\Omega_i}(\mathcal{R}_{i,j,k}-\alpha_i-\beta_j),\quad \forall k\in[n_3]$$

Empirically, $\alpha_i,\beta_j,\gamma_k$ are estimated iteratively by the above three equations until the following 'residual' converges to zero,

$$\text{residual}=\sum_{i=1}^{n_1}(\frac{1}{|\Omega_i|}\sum_{(j,k)\in\Omega_i}\tilde{\mathcal{R}}_{i,j,k})^2+\sum_{j=1}^{n_2}(\frac{1}{|\Omega_j|}\sum_{(i,k)\in\Omega_j}\tilde{\mathcal{R}}_{i,j,k})^2+\sum_{k=1}^{n_3}(\frac{1}{|\Omega_k|}\sum_{(i,j)\in\Omega_k}\tilde{\mathcal{R}}_{i,j,k})^2$$

Generally speaking, there is no guarantee that this algorithm will converge except in certain noted cases. But when we implement this method to UCLAF, we found that a rapid convergence can be attained. (More details in Appendix C.5).

To compare the performance of different methods, we randomly split 50% of the observed data as training and hold out the remaining as testing. Two comparative approaches are chosen to compare with nested double ADMM. The first one is simple guess - it takes 0 as imputation without splitting the training and testing dataset. Another is Ten-ALS which is the method we discussed in Section 4.5.1. As is shown in Table 4.5, our algorithm outper-

	Our approach	Simple guess	Ten-ALS
Testing error	0.4487(0.00935)	1.5738(-)	0.7808(0.08250)

Table 4.5: Recovery performance comparison for proposed nested double ADMM algorithm, simple guess and Ten-ALS [Jain and Oh, 2014] on UCLAF dataset.

	Our approach	PTD	Bayesian CP
AUC	0.9770(0.01352)	0.9407 (0.01563)	0.9113(0.01134)

.

Table 4.6: AUC score comparison for proposed nested double ADMM algorithm, PTD [Rai et al., 2015] and Bayesian CP [Rai et al., 2014] on UCLAF dataset **after binarization**.

forms the other two alternatives taking advantages of auxiliary information incorporation and flexible nuclear norm constrained low rank assumption.

Recommendation system datasets often face the challenge of sparsity and the same problem appears in UCLAF dataset. Among the $164 \times 168 \times 5 = 137760$ entries, there are a large proportion, 98.957% of entries taking the value 0. Besides, the frequencies that particular users visit some locations are imbalanced, leading to a highly skewed distribution of the entries in tensor $\mathcal{R}$. A simple way to deal with this is to binarize the data which has been implemented in Rai et al. [2015]. We truncate all the entries with value larger than 1 to 1, indicating that the specific user i conducted some activities k at location j. Due to the fact that after binarization, the UCLAF problem can be treated as a binary classification, we compute the AUC score on testing dataset for nested double ADMM algorithm and compare with another two approaches, probabilistic tensor decomposition (denoted as PDT), which is the method proposed in [Rai et al., 2015] on a basis of probabilistic tensor decomposition model and Bayesian CP which is a Bayesian low rank decomposition method proposed in Rai et al. [2014]. The comparison result is reported in Table 4.6 and still, our algorithm beats other methods significantly.

To validate the advantage of the tensor completion method, we compare the nested double ADMM algorithm with another popular matrix completion approach, inductive tensor

completion (IMC) [Jain and Dhillon, 2013]. IMC is implemented on slices along each mode of $\mathcal{R} \in \mathbb{R}^{164\times168\times5}$. For example, in terms of activity mode, $\mathcal{R}_{:,:,i} \in \mathbb{R}^{164\times168}, i \in \{1,2,3,4,5\}$ will be treated as matrices with partial observations and the missing values will be imputed by taking into account the auxiliary information of user similarity and location covariates.

	Our approach	Activity	User	Location
Training error	0.2038 (0.00189)	0.3526(0.00841)	0.3198(0.00847)	0.4490(0.013928)
Testing error	0.4487(0.00935)	0.4625(0.00797)	0.4712(0.01282)	0.4784(0.01378)

Table 4.7: Recovery performance comparison for nested double ADMM algorithm with IMC [Jain and Dhillon, 2013]

The results are listed in Table 4.7. LISTAI achieves the best imputation performance on the testing dataset. This improvement in performance is attributed to the advantages of tensor-based LISTAI method which utilizes auxiliary information of all three modes for imputation. Moreover, LISTAI achieves a significantly lower training error, which results in a high goodness-of-fit for data without having an overfitting problem. By comparing the testing error difference between LISTAI and matrix completion along each mode, the contribution of auxiliary information to missing value imputation along each mode is straightforward to recognize. Incorporating the location covariates encourages the most evident improvement in imputation, where testing error reduces 2.97% (derived based on 0.4784 - 0.4487) compared to 1.38% and 2.25% for incorporating activity similarity and user similarity matrix respectively. This is expected, since location covariates contain 18 variables, indicating number of points of interests near the location, that provide rich information; this is in contrast to the top four and two principal components selected for user and activity similarity matrices.

APPENDIX A

TECHNICAL DETAILS OF CHAPTER 2

A.1 Proof of Theorem 1 and Corollary 1

A.1.1 Proof of Theorem 1

Using $\hat{\mathbf{C}}_{:i}$ as an example, Algorithm 2 applies the following steps to do the updates:

1. Orthogonal projection:

   Suppose $\hat{\mathbf{A}}_{:i}$ and $\hat{\mathbf{B}}_{:i}$ are estimates from previous iteration. With a slight abuse of notation, first calculate projection $\bar{\mathbf{A}}_{:i}$ and $\bar{\mathbf{B}}_{:i}$ to the previous $(i-1)$ orthogonal basis, $\{\bar{\mathbf{A}}_{:j}, j < i\}$ and $\{\bar{\mathbf{B}}_{:j}, j < i\}$, i.e.,

$$\bar{\mathbf{A}}_{:i} = \hat{\mathbf{A}}_{:i} - \sum_{j<i} \bar{\mathbf{A}}_{:j}^{\top} \hat{\mathbf{A}}_{:i} \bar{\mathbf{A}}_{:j}$$
$$\bar{\mathbf{B}}_{:i} = \hat{\mathbf{B}}_{:i} - \sum_{j<i} \bar{\mathbf{B}}_{:j}^{\top} \hat{\mathbf{B}}_{:i} \bar{\mathbf{B}}_{:j}$$

2. ALS-update:

   In the Algorithm 2, updates for unnormalized $\mathbf{Z}$ are calculated by

$$\mathbf{Z} = \mathcal{Y}_{(3)}(\bar{\mathbf{B}} \odot \bar{\mathbf{A}})$$

   Next we do normalization for each column in $\mathbf{Z}$ and denote the normalized column as $\mathbf{Z}_{:i}$, we can written this update in tensor power update form as

$$\mathbf{Z}_{:i} = \frac{\mathcal{Y}(\bar{\mathbf{A}}_{:i}, \bar{\mathbf{B}}_{:i}, \mathbf{I})}{\|\mathcal{Y}(\bar{\mathbf{A}}_{:i}, \bar{\mathbf{B}}_{:i}, \mathbf{I})\|_2}$$

3. Fuse operator:

$$\tilde{\mathbf{Z}}_{:i} = \arg\min_{\mathbf{C}_{:i}} \frac{1}{2}\|\mathbf{Z}_{:i} - \mathbf{C}_{:i}\|_2^2 + \lambda\| \ {}^3\mathbf{\Delta}\mathbf{C}_{:i}\|_1$$

4. Normalization:

$$\hat{\mathbf{C}}_{:i} = \frac{\tilde{\mathbf{Z}}_{:i}}{\|\tilde{\mathbf{Z}}_{:i}\|_2}$$

Step 1: convergence error for $\hat{\mathbf{C}}_{:1}$

Consider the first column in the third factor matrix $\mathbf{C}_{:1}$, the update for this column is not affected by 'orthogonalization' step and our first step is to prove the convergence bound for $\hat{\mathbf{C}}_{:1}$ and then use induction to prove the result holding for the following $K-1$ columns. Suppose we have initialization assumption $\max\{\|\hat{\mathbf{A}}_{:1} - \mathbf{A}\colon 1\|_2, \|\hat{\mathbf{B}}_{:1} - \mathbf{B}\colon 1\|_2\} \leq \epsilon_0$, then update for $\mathbf{Z}_{:1}$ can be written as

$$\mathbf{Z}_{:1} = \frac{\mathcal{Y}(\hat{\mathbf{A}}_{:1}, \hat{\mathbf{B}}_{:1}, \mathbf{I})}{\|\mathcal{Y}(\hat{\mathbf{A}}_{:1}, \hat{\mathbf{B}}_{:1}, \mathbf{I})\|_2}$$

and further we can write $\|\mathbf{Z}_{:1} - \mathbf{C}_{:1}\|_2$ as

$$\begin{aligned}\|\mathbf{Z}_{:1} - \mathbf{C}_{:1}\|_2 &= \|\frac{\mathcal{Y}(\hat{\mathbf{A}}_{:1}, \hat{\mathbf{B}}_{:1}, \mathbf{I})}{\|\mathcal{Y}(\hat{\mathbf{A}}_{:1}, \hat{\mathbf{B}}_{:1}, \mathbf{I})\|_2} - \mathbf{C}_{:1}\|_2 \\ &\leq \|\frac{\mathcal{Y}^*(\hat{\mathbf{A}}_{:1}, \hat{\mathbf{B}}_{:1}, \mathbf{I})}{\|\mathcal{Y}(\hat{\mathbf{A}}_{:1}, \hat{\mathbf{B}}_{:1}, \mathbf{I})\|_2} - \mathbf{C}_{:1}\|_2 + \|\frac{\mathcal{E}(\hat{\mathbf{A}}_{:1}, \hat{\mathbf{B}}_{:1}, \mathbf{I})}{\|\mathcal{Y}(\hat{\mathbf{A}}_{:1}, \hat{\mathbf{B}}_{:1}, \mathbf{I})\|_2}\|_2\end{aligned}$$

Follow the proof for Theorem 3 in Sun and Li [2019] and denote $f(\epsilon_0, \rho, K) := \alpha\epsilon_0^2 + \rho^2(K-1) + 2\epsilon_0\rho(K-1)$, we can show the following convergence error bound

$$\|\mathbf{Z}_{:1} - \mathbf{C}_{:1}\|_2 \leq \frac{2w_{\max}f(\epsilon_0, \rho, K) + 2\psi}{w_1(1-\epsilon_0^2) - w_{\max}f(\epsilon_0, \rho, K) - \psi}$$

By Lemma 1 and choosing appropriate tuning parameter λ, we have

$$\|\hat{\mathbf{C}}_{:1} - \mathbf{C}_{:1}\|_2^2 \leq \Big[\frac{2w_{\max}f(\epsilon_0,\rho,K) + 2\psi}{w_1(1-\epsilon_0^2) - w_{\max}f(\epsilon_0,\rho,K) - \psi}\Big]^2 + \frac{8M\| \ {}^3\mathbf{\Delta}\mathbf{C}_{:1}\|_1(w_{\max}f(\epsilon_0,\rho,K)+\psi)}{w_1(1-\epsilon_0^2) - w_{\max}f(\epsilon_0,\rho,K) - \psi}$$

Moreover, under bounded fusion assumption,

$$\|\hat{\mathbf{C}}_{:1} - \mathbf{C}_{:1}\|_2 \leq \frac{2\sqrt{2}(w_{\max}f(\epsilon_0,\rho,K)+\psi)}{w_1(1-\epsilon_0^2) - w_{\max}f(\epsilon_0,\rho,K) - \psi}$$

Denote $f(\epsilon_0,\rho,K) = \rho^2(K-1) + \tilde{q}\epsilon_0$ where $\tilde{q} = \alpha\epsilon_0 + 2\rho(K-1)$. Under assumption 2, we have $\epsilon_0 \leq \min\{\frac{w_{\min}}{6w_{\max}} - \rho^2(K-1), \frac{w_{\min}}{12\sqrt{2}w_{\max}\alpha} - \frac{2\rho(K-1)}{\alpha}\}$ and this leads to $\epsilon_0^2 \leq \frac{w_{\min}}{6w_{\max}}$ and $\tilde{q} \leq 1$. Furthermore, we can derive $f(\epsilon_0,\rho,K) \leq \frac{w_{\min}}{6w_{\max}}$.

$$\begin{aligned}
&w_1(1-\epsilon_0^2) - w_{\max}f(\epsilon_0,\rho,K) - \psi \\
\geq& w_{\min}(1 - \frac{w_{\max}}{w_{\min}}\epsilon_0^2 - \frac{w_{\max}}{w_{\min}}f(\epsilon_0,\rho,K) - \frac{\psi}{w_{\min}}) \\
\geq& w_{\min}(1 - \frac{1}{6} - \frac{1}{6} - \frac{1}{6}) \geq \frac{w_{\min}}{2}
\end{aligned}$$

Thus,

$$\|\hat{\mathbf{C}}_{:1} - \mathbf{C}_{:1}\|_2^2 \leq \frac{4\sqrt{2}w_{\max}}{w_{\min}}\rho^2 K + \frac{4\sqrt{2}}{w_{\min}}\psi + \frac{4\sqrt{2}w_{\max}}{w_{\min}}\tilde{q}\epsilon_0$$

where $\frac{4\sqrt{2}w_{\max}}{w_{\min}}\tilde{q} \leq \frac{1}{3}$. Then, by iteratively applying the above result, we can show that

$$\|\hat{\mathbf{C}}_{:1} - \mathbf{C}_{:1}\|_2^2 \lesssim \gamma\rho^2 K + \frac{\psi}{w_{\min}}$$

Step 2: convergence error for $\hat{\mathbf{C}}_{:i}, \forall i \in \{1,2,...,K\}$

We now prove that Fused-Orth-ALS algorithm also recovers the remaining columns. We have already shown that the algorithm recovers the first column and we would like to use induction to prove the result for the remaining columns: if the first $(i-1)$ columns have

converged, the ith column in the factor matrix $\mathbf{C}$ also converges. The main idea is that, since the correlation between columns in factor matrices are small, the orthogonalization step will not affect the factors which have not been recovered but ensures the ith estimate never has high correlation with the first $i-1$ columns which have already been recovered. Lemma 3 proved this claim. Next, we will bound $\|\hat{\mathbf{C}}_{:i} - \mathbf{C}_{:i}\|_2$, for $i > 1$. Still, we will start by bounding the difference between ALS update $\mathbf{Z}_{:i}$ and $\mathbf{C}_{:i}$ and then apply Lemma 1 to consider the effect of fuse operator.

$$\begin{aligned}\|\mathbf{Z}_{:i} - \mathbf{C}_{:i}\|_2 =& \|\frac{\mathcal{Y}(\bar{\mathbf{A}}_{:i}, \bar{\mathbf{B}}_{:i}, \mathbf{I})}{\|\mathcal{Y}(\bar{\mathbf{A}}_{:i}, \bar{\mathbf{B}}_{:i}, \mathbf{I})\|_2} - \mathbf{C}_{:i}\|_2 \\ =& \|\underbrace{\frac{\mathcal{Y}^*(\bar{\mathbf{A}}_{:i}, \bar{\mathbf{B}}_{:i}, \mathbf{I})}{\|\mathcal{Y}(\bar{\mathbf{A}}_{:i}, \bar{\mathbf{B}}_{:i}, \mathbf{I})\|_2} - \mathbf{C}_{:i}}_{II_1}\|_2 + \|\underbrace{\frac{\mathcal{E}(\bar{\mathbf{A}}_{:i}, \bar{\mathbf{B}}_{:i}, \mathbf{I})}{\|\mathcal{Y}(\bar{\mathbf{A}}_{:i}, \bar{\mathbf{B}}_{:i}, \mathbf{I})\|_2}}_{II_2}\|_2\end{aligned}$$

We will follow similar procedures we used in proving the convergence for the first column. Note that $\bar{\mathbf{A}}_{:i} = a(\hat{\mathbf{A}}_{:i} - \sum_{j<i} \bar{\mathbf{A}}_{:j}^\top \hat{\mathbf{A}}_{:i} \bar{\mathbf{A}}_{:j})$ and $\bar{\mathbf{B}}_{:i} = b(\hat{\mathbf{B}}_{:i} - \sum_{j<i} \bar{\mathbf{B}}_{:j}^\top \hat{\mathbf{B}}_{:i} \bar{\mathbf{B}}_{:j})$ where a, b are two normalization parameters to keep $\|\bar{\mathbf{A}}_{:i}\|_2 = \|\bar{\mathbf{B}}_{:i}\|_2 = 1$ holds. We will ignore the normalization parameters a, b when we analyze II_1, II_2 since they will both appear in the numerator and denominator and could be cancelled finally. We first take a look at the numerator of II_1.

$$\begin{aligned}\mathcal{Y}^*(\bar{\mathbf{A}}_{:i}, \bar{\mathbf{B}}_{:i}, \mathbf{I}) &= \mathcal{Y}^*(\hat{\mathbf{A}}_{:i} - \sum_{j<i} \bar{\mathbf{A}}_{:j}^\top \hat{\mathbf{A}}_{:i} \bar{\mathbf{A}}_{:j}, \hat{\mathbf{B}}_{:i} - \sum_{j<i} \bar{\mathbf{B}}_{:j}^\top \hat{\mathbf{B}}_{:i} \bar{\mathbf{B}}_{:j}, \mathbf{I}) \\ &= \underbrace{\mathcal{Y}^*(\hat{\mathbf{A}}_{:i}, \hat{\mathbf{B}}_{:i}, \mathbf{I})}_{II_{11}} - \underbrace{\sum_{j<i} \bar{\mathbf{A}}_{:j}^\top \hat{\mathbf{A}}_{:i} \mathcal{Y}^*(\bar{\mathbf{A}}_{:j}, \hat{\mathbf{B}}_{:i}, \mathbf{I})}_{II_{12}} - \underbrace{\sum_{j<i} \bar{\mathbf{B}}_{:j}^\top \hat{\mathbf{B}}_{:i} \mathcal{Y}^*(\hat{\mathbf{A}}_{:i}, \bar{\mathbf{B}}_{:j}, \mathbf{I})}_{II_{13}} + \\ &\underbrace{\sum_{j<i} \bar{\mathbf{A}}_{:j}^\top \hat{\mathbf{A}}_{:i} \sum_{j<i} \bar{\mathbf{B}}_{:j}^\top \hat{\mathbf{B}}_{:i} \mathcal{Y}^*(\bar{\mathbf{A}}_{:j}, \bar{\mathbf{B}}_{:j}, \mathbf{I})}_{II_{14}}\end{aligned}$$

Obviously, we have to bound $\bar{\mathbf{A}}_{:j}^{\top}\mathbf{A}_{:i}, \forall j < i$. Under Lemma 3, we have

$$\begin{aligned}\Delta :&= \max_{j<i} \bar{\mathbf{A}}_{:j}^{\top}\mathbf{A}_{:i} \\ &\leq (\mathbf{A}_{:j} + \xi_j)^{\top}(\mathbf{A}_{:i} + \hat{\xi}_i) \\ &\leq \alpha/\sqrt{d} + \epsilon_0 + 16\sqrt{2}\gamma\alpha K/\sqrt{d} + 16\sqrt{2}\gamma\alpha K\epsilon_0/\sqrt{d}\end{aligned}$$

Next, we will bound $II_{11}, II_{12}, II_{13}, II_{14}$ respectively.

Bound $\|II_{11}\|_2$ After re-randomization, we can use the conclusion from convergence result from step 1, i.e.,

$$\begin{aligned}II_{11} &= II'_{11} + w_i\mathbf{C}_{:i} \\ \|II'_{11}\|_2 &\leq w_{\max} f(\epsilon_0, \rho, K) + 2w_i\epsilon_0\end{aligned}$$

Bound $\|II_{12}\|_2$ Similarly, II_{12} can be written as

$$\sum_{j<i} \bar{\mathbf{A}}_{:j}^{\top}\hat{\mathbf{A}}_{:i}\mathcal{Y}^*(\bar{\mathbf{A}}_{:j}, \hat{\mathbf{B}}_{:i}, \mathbf{I}) \leq (K-1)\Delta\mathcal{Y}(\bar{\mathbf{A}}_{:j}, \hat{\mathbf{B}}_{:i}, \mathbf{I})$$

$$\begin{aligned}\mathcal{Y}^*(\bar{\mathbf{A}}_{:j}, \hat{\mathbf{B}}_{:i}, \mathbf{I}) =& \mathcal{Y}^*(\bar{\mathbf{A}}_{:j} - \mathbf{A}_{:j} + \mathbf{A}_{:j}, \hat{\mathbf{B}}_{:i} - \mathbf{B}_{:i} + \mathbf{B}_{:i}, \mathbf{I}) = \underbrace{\mathcal{Y}^*(\bar{\mathbf{A}}_{:j} - \mathbf{A}_{:j}, \hat{\mathbf{B}}_{:i} - \mathbf{B}_{:i}, \mathbf{I})}_{i_1} \\ &+ \underbrace{\mathcal{Y}^*(\mathbf{A}_{:j}, \hat{\mathbf{B}}_{:i} - \mathbf{B}_{:i}, \mathbf{I})}_{i_2} + \underbrace{\mathcal{Y}^*(\bar{\mathbf{A}}_{:j} - \mathbf{A}_{:j}, \mathbf{B}_{:i}, \mathbf{I})}_{i_3} + \underbrace{\mathcal{Y}^*(\mathbf{A}_{:j}, \mathbf{B}_{:i}, \mathbf{I})}_{i_4}\end{aligned}$$

Using CP low rank decomposition structure of $\mathcal{Y}^*$, we have

$$\begin{aligned}\|i_1\|_2 &= \| \sum_{l\in[K]} \mathcal{Y}^* \langle \bar{\mathbf{A}}_{:j} - \mathbf{A}_{:j}, \mathbf{A}_{:l}\rangle \langle \hat{\mathbf{B}}_{:i} - \mathbf{B}_{:i}, \mathbf{B}_{:l}\rangle \mathbf{C}_{:l} \|_2 \\ &\leq \epsilon_0 \xi_j \sum_{l\in[K]} w_l \leq \epsilon_0 \xi_j w_{\max} \alpha\end{aligned}$$

where the last inequality can be derived from Lemma 3 and assumption about spectral norm of $\mathcal{Y}^*$, i.e., $\|\mathcal{Y}^*\| \leq w_{\max}\alpha$. Similarly, by imposing the incoherence assumption 1, we can bound $\|i_2\|_2$

$$\begin{aligned}\|i_2\|_2 &= \| \sum_{l\neq j} \langle \mathbf{A}_{:j}, \mathbf{A}_{:l}\rangle \langle \hat{\mathbf{B}}_{:i} - \mathbf{B}_{:i}, \mathbf{B}_{:l}\rangle \mathbf{C}_{:l} + w_j \langle \hat{\mathbf{B}}_{:i} - \mathbf{B}_{:i}, \mathbf{B}_{:j}\rangle \mathbf{C}_{:j} \|_2 \\ &\leq \epsilon_0 \rho (K-1) w_{\max} + w_j \epsilon_0\end{aligned}$$

To bound $\|i_3\|_2, \|i_4\|_2$, we should first notice to split i_3, i_4 into 2 parts with the second part is related to $w_i \mathbf{C}_{:i}$,

$$\begin{aligned}i_3 &= i_3' + \xi_j w_i \mathbf{C}_{:i} \\ \|i_3'\|_2 &= \sum_{l\neq j} w_l \langle \bar{\mathbf{A}}_{:j} - \mathbf{A}_{:j}, \mathbf{A}_{:l}\rangle \langle \mathbf{B}_{:i}, \mathbf{B}_{:l}\rangle \mathbf{C}_{:l} \\ &\leq \xi_j \rho (K-1) w_{\max}\end{aligned}$$

$$\begin{aligned}i_4 &= i_4' + w_i \rho \mathbf{C}_{:i} \\ \|i_4'\|_2 &= \sum_{l\neq j} w_l \langle \mathbf{A}_{:j}, \mathbf{A}_{:l}\rangle \langle \mathbf{B}_{:i}, \mathbf{B}_{:l}\rangle \mathbf{C}_{:l} \\ &\leq \rho^2 (K-2) w_{\max} + w_j \rho\end{aligned}$$

Thus, combine the above results,

$$II_{12} \leq (K-1)\Delta II'_{12} + (K-1)\Delta(\xi_j + \rho)w_i \mathbf{C}_{:i}$$

with

$$\begin{aligned}\|II'_{12}\|_2 &\leq \|i_1\|_2 + \|i_2\|_2 + \|i'_3\|_2 + \|i'_4\|_2 \\ &\leq \xi_j \epsilon_0 w_{\max}\alpha + (\epsilon_0 + \xi_j)\rho(K-1)w_{\max} + \rho^2(K-2)w_{\max} + (\epsilon_0 + \rho)w_{\max}\end{aligned}$$

Bound $\|II_{13}\|_2$ Similar as $\|II_{12}\|_2$.

Bound $\|II_{14}\|_2$

$$\begin{aligned}II_{14} &= \sum_{j_1 \leq i} \bar{\mathbf{A}}_{:j_1}^\top \mathbf{A}_{:i} \sum_{j_2 \leq i} \bar{\mathbf{B}}_{:j_1}^\top \mathbf{B}_{:i} \mathcal{Y}^*(\bar{\mathbf{A}}_{:j_1} - \mathbf{A}_{:j_1} + \mathbf{A}_{:j_1}, \bar{\mathbf{B}}_{:j_1} - \mathbf{B}_{:j_1} + \mathbf{B}_{:j_1}, \mathbf{I}) \\ &\leq ((K-1)\Delta)^2 (\underbrace{\mathcal{Y}^*(\bar{\mathbf{A}}_{:j_1} - \mathbf{A}_{:j_1}, \bar{\mathbf{B}}_{:j_1} - \mathbf{B}_{:j_1}, \mathbf{I})}_{ii_1} + \underbrace{\mathcal{Y}^*(\bar{\mathbf{A}}_{:j_1} - \mathbf{A}_{:j_1}, \mathbf{B}_{:j_1}, \mathbf{I})}_{ii_2} \\ &+ \underbrace{\mathcal{Y}^*(\mathbf{A}_{:j_1}, \bar{\mathbf{B}}_{:j_1} - \mathbf{B}_{:j_1}, \mathbf{I})}_{ii_3} + \underbrace{\mathcal{Y}^*(\mathbf{A}_{:j_1}, \mathbf{B}_{:j_1}, \mathbf{I})}_{ii_4})\end{aligned}$$

Still, under the CP decomposition structure of $\mathcal{Y}^*$, we have

$$\begin{aligned}\|ii_1\|_2 &= \Big\| \sum_{l \in [K]} w_l \langle \bar{\mathbf{A}}_{:j_1} - \mathbf{A}_{:j_1}, \mathbf{A}_{:l} \rangle \langle \bar{\mathbf{B}}_{:j_2} - \mathbf{B}_{:j_2}, \mathbf{B}_{:l} \rangle \mathbf{C}_{:l} \Big\|_2 \\ &\leq \xi_{j_1} \xi_{j_2} w_{\max} \alpha\end{aligned}$$

$$\begin{aligned}\|ii_2\|_2 &= \| \sum_{l \in [K]} w_l \langle \mathbf{A}_{:j_1}, \mathbf{A}_{:l} \rangle \langle \bar{\mathbf{B}}_{:j_2} - \mathbf{B}_{:j_2}, \mathbf{B}_{:l} \rangle \mathbf{C}_{:l} \|_2 \\ &\leq w_{\max} \xi_{j_2} + \rho(K-1)\xi_{j_2} w_{\max}\end{aligned}$$

$\|ii_3\|_2$ can be bounded similar by $w_{\max}\xi_{j_1} + \rho(K-1)\xi_{j_1}w_{\max}$. For ii_4,

$$\begin{aligned}\|ii_4\|_2 &= \sum_{l\in[K]} w_l\langle \mathbf{A}_{:j_1}, \mathbf{A}_{:l}\rangle\langle \mathbf{B}_{:j_2}, \mathbf{B}_{:l}\rangle \mathbf{C}_{:l} \\ &\leq (K-2)\rho^2 w_{\max} + 2w_{\max}\rho\end{aligned}$$

Combine the above results, we have

$$\|II_{14}\|_2 \leq \xi_{j_1}\xi_{j_2}w_{\max}\alpha + w_{\max}(\xi_{j_1} + \xi_{j_2} + 2\rho) + \rho(K-1)(\xi_{j_1} + \xi_{j_2})w_{\max} + (K-2)^2\rho^2 w_{\max}$$

We denote $\xi := \max_j \xi_j \leq 16\sqrt{2}\gamma\alpha K/\sqrt{d}, \forall j \in [K]$. Then, we can summarize that

$$\mathcal{Y}^*(\bar{\mathbf{A}}_{:i}, \bar{\mathbf{B}}_{:i}, \mathbf{I}) \leq \Lambda + \underbrace{(2(K-1)\Delta(\xi+\rho)+1)}_{\eta_0} w_i \mathbf{C}_{:i}$$

where

$$\begin{aligned}\|\Lambda\|_2 &= \eta_1 w_{\max}\alpha + \eta_2\rho(K-1)w_{\max} + \eta_3\rho^2(K-2)w_{\max} \\ \eta_1 &= 2(K-1)\Delta\xi\epsilon_0 + (K-1)^2\xi^2\delta^2 + \epsilon_0^2 \\ \eta_2 &= 2(K-1)\Delta(\epsilon_0+\xi) + 2(K-1)^2\Delta^2\xi + \epsilon_0 \\ \eta_3 &= 2(K-1)\Delta + (K-1)^2\Delta^2 + 1\end{aligned}$$

Similarly, the denominator of II_1 can be lower bounded by

$$\|\mathcal{Y}^*(\bar{\mathbf{A}}_{:i}, \bar{\mathbf{B}}_{:i}, \mathbf{I})\|_2 \geq w_i(1-\epsilon_0^2) - w_{\max}\|\Lambda\|_2 - \psi$$

Thus,

$$II_1 \leq \frac{\Lambda + \eta_0 w_i \mathbf{C}_{:i} - (w_i(1-\epsilon_0^2) - w_{\max}\|\Lambda\|_2 - \psi)\mathbf{C}_{:i}}{w_i(1-\epsilon_0^2) - w_{\max}\|\Lambda\|_2 - \psi}$$

In Assumption 1, we have $K\rho^2 = o(1)$ combined with initialization condition

$$\epsilon_0 \leq \frac{(\sqrt{2}-1)\sqrt{d}/(K-1) - \alpha(1+16\sqrt{2}\gamma K)}{\sqrt{d}+16\sqrt{2}\alpha\gamma K}$$

it's easy to show that $\eta_0 \leq 2, \eta_1 \leq 2\epsilon_0^2, \eta_2 \leq 2\epsilon_0, \eta_3 \leq 2$

$$\|II_1\|_2 \leq \frac{4w_{\max}f(\epsilon_0,\rho,K)+\psi}{w_i(1-\epsilon_0^2)-2w_{\max}f(\epsilon_0,\rho,K)-\psi}$$

Following the similar argument in step 1, we can show

$$\|II_2\|_2 \leq \frac{\psi}{w_i(1-\epsilon_0^2)-2w_{\max}f(\epsilon_0,\rho,K)-\psi}$$

Therefore, combine $\|II_1\|_2$ and $\|II_2\|_2$,

$$\|\mathbf{Z}_{:i}-\mathbf{C}_{:i}\|_2 \leq \frac{4w_{\max}f(\epsilon_0,\rho,K)+2\psi}{w_i(1-\epsilon_0^2)-2w_{\max}f(\epsilon_0,\rho,K)-\psi}$$

Next we consider the effect of fuse operator. Similar as step 1, under Assumption 4 and by choosing appropriate tuning parameter λ, we have

$$\|\hat{\mathbf{C}}_{:i}-\mathbf{C}_{:i}\|_2 \leq \frac{4\sqrt{2}w_{\max}f(\epsilon_0,\rho,K)+2\sqrt{2}\psi}{w_i(1-\epsilon_0^2)-2w_{\max}f(\epsilon_0,\rho,K)-\psi}$$

Under Assumption 2, $w_i(1-\epsilon_0^2)-2w_{\max}f(\epsilon_0,\rho,K)-\psi \geq w_{\min}(1-\frac{1}{6}-\frac{1}{3}-\frac{1}{6}) = \frac{w_{\min}}{3}$. Then,

$$\|\hat{\mathbf{C}}_{:i}-\mathbf{C}_{:i}\|_2 \leq 12\sqrt{2}\frac{w_{\max}}{w_{\min}}\rho^2(K-1)+6\sqrt{2}\frac{\psi}{w_{\min}}+12\sqrt{2}\frac{w_{\max}}{w_{\min}}\tilde{q}$$

We know that $12\sqrt{2}\frac{w_{\max}}{w_{\min}}\tilde{q} \leq 1$ by assumption 2. Thus, iteratively implementing the above result, we have

$$\|\hat{\mathbf{C}}_{:i} - \mathbf{C}_{:i}\|_2 \lesssim \gamma\rho^2(K-1) + \frac{\psi}{w_{\min}}$$

A.1.2 Proof for Corollary 1

Theorem 1 shows that

$$\|\hat{\mathbf{C}}_{:i} - \mathbf{C}_{:i}\|_2 \lesssim \frac{w_{\max}}{w_{\min}}\rho^2(K-1) + \frac{\psi}{w_{\min}}$$

Under the assumption that $\mathcal{E}_{ijk}$ is independent, zero-mean and $\mathbb{E}[e^{t\mathcal{E}_{ijk}}] \leq e^{\frac{\sigma^2 t^2}{2}}$, by Lemma 2, we have

$$\psi \leq \sqrt{8\sigma^2((d_1+d_2+d_3)\log\frac{6}{\log 3/2} + \log\frac{2}{\delta})}$$

holds with probability $1-\delta$. Combined with $w_{\min} \succ \sqrt{\sigma^2[3d\log\frac{6}{\log 3/2} + \log\frac{2}{\delta}]d^2/(K-1)}$, we have

$$\frac{\psi}{w_{\min}} \lesssim \frac{(K-1)}{d}$$

Combined with Assumption 1 that $\rho \leq \frac{\alpha}{\sqrt{d}}$, it's easy to derive that

$$\|\hat{\mathbf{C}}_{:i} - \mathbf{C}_{:i}\|_2 \lesssim \frac{(K-1)}{d}$$

A.2 Proof of Theorem 2

We have shown in Corollary 1 that if elements in error tensor $\mathcal{E}$ are independently and identically sub-Gaussian distributed, we have

$$\|\hat{\mathbf{C}}_{:i} - \mathbf{C}_{:i}\|_2 \lesssim \frac{(K-1)}{d}$$

Based on this result, we obtain the estimation error in clustering

$$\max_i \|\hat{\boldsymbol{\mu}}_{3,i} - \boldsymbol{\mu}^*_{3,i}\|_2 \leq \sqrt{K}\|\hat{\mathbf{C}}_{:i} - \mathbf{C}_{:i}\|_2 \leq \tilde{C}\frac{K^{1.5}}{d}$$

for some constant $\tilde{C}$. If $\min_{i,j} \|\boldsymbol{\mu}^*_{3,i} - \boldsymbol{\mu}^*_{3,j}\|_2 \geq C\frac{K^{1.5}}{d}$ for some constant $C > 4\tilde{C}$, we have, for any two samples $\hat{\boldsymbol{\mu}}_{3,i}, \hat{\boldsymbol{\mu}}_{3,j}$ from two different clusters $\mathfrak{C}^*_i, \mathfrak{C}^*_j$ respectively,

$$\begin{aligned}\|\hat{\boldsymbol{\mu}}_{3,i} - \hat{\boldsymbol{\mu}}_{3,j}\|_2 &= \|\hat{\boldsymbol{\mu}}_{3,i} - \boldsymbol{\mu}^*_{3,i} + \boldsymbol{\mu}^*_{3,i} - \boldsymbol{\mu}^*_{3,j} + \boldsymbol{\mu}^*_{3,j} - \hat{\boldsymbol{\mu}}_{3,j}\|_2 \\ &\geq \|\boldsymbol{\mu}^*_{3,i} - \boldsymbol{\mu}^*_{3,j}\|_2 - \|\hat{\boldsymbol{\mu}}_{3,i} - \boldsymbol{\mu}^*_{3,i}\|_2 - \|\boldsymbol{\mu}^*_{3,j} - \hat{\boldsymbol{\mu}}_{3,j}\|_2 \geq 2\tilde{C}\frac{K^{1.5}}{d}\end{aligned}$$

For any two samples $\hat{\boldsymbol{\mu}}_{3,i}, \hat{\boldsymbol{\mu}}_{3,i'}$ from same cluster $\mathfrak{C}^*_i$,

$$\begin{aligned}\|\hat{\boldsymbol{\mu}}_{3,i} - \hat{\boldsymbol{\mu}}_{3,i'}\|_2 &= \|\hat{\boldsymbol{\mu}}_{3,i} - \boldsymbol{\mu}^*_{3,i} + \boldsymbol{\mu}^*_{3,i} - \hat{\boldsymbol{\mu}}_{3,i'}\|_2 \\ &\leq \|\hat{\boldsymbol{\mu}}_{3,i} - \boldsymbol{\mu}^*_{3,i}\|_2 + \|\boldsymbol{\mu}^*_{3,i} - \hat{\boldsymbol{\mu}}_{3,i'}\|_2 \\ &\leq 2\tilde{C}\frac{K^{1.5}}{d} \\ &\lesssim \frac{K^{1.5}}{d}\end{aligned}$$

Thus, with-in cluster distance is always smaller than the between-cluster distance, and henceforth, we will get the clustering consistency, $\hat{\mathfrak{C}}_i = \mathfrak{C}^*_i, \forall i \in \{1, 2, ..., s_3\}$ with high probability. Analogously, this method can be applied to the first and second mode to get the similar

results.

To see how this bound relate to the cluster size s_3, we will do the following analysis. Recall that Assumption 4 imposes the following restriction

$$\| {}^3\mathbf{\Delta}\mathbf{C}_{:i}\|_1 \leq \frac{w_{\max}(\epsilon_0^2\alpha + 2\epsilon_0\rho(K-1) + \rho^2(K-1)) + \psi}{2Mw_{\min}(1-\epsilon_0^2)}$$

Considering simple case when we have balanced size in each cluster, i.e., there are d/s_3 samples in each cluster. Then, $\| {}^3\mathbf{\Delta}\mathbf{C}_{:i}\|_1$ can be bounded by

$$\begin{aligned}
\| {}^3\mathbf{\Delta}\mathbf{C}_{:i}\|_1 &= \sum_{i,j\in[s_3],i<j} \Big(\frac{d}{s_3}\Big)^2 |\boldsymbol{\mu}^*_{3,i} - \boldsymbol{\mu}^2_{3,j}| \\
&\leq \Big(\frac{d}{s_3}\Big)^2 (s_3-1) \sum_{j\in[s_3]} |\boldsymbol{\mu}^*_{3,j}| \\
&\leq \Big(\frac{d}{s_3}\Big)^2 (s_3-1)\sqrt{s_3}\sqrt{\sum_{j\in[s_3]} |\boldsymbol{\mu}^*_{3,j}|^2} \\
&\leq \Big(\frac{d}{s_3}\Big)^2 (s_3-1)\sqrt{s_3}\sqrt{\frac{s_3}{d}} \leq d^{1.5}(1-\frac{1}{s_3})
\end{aligned} \tag{A.1}$$

where the second inequality is due to Cauchy-Schwarz inequality and the third inequality is derived from the following fact,

$$\begin{aligned}
\|\mathbf{C}_{:i}\|_2 = \sum_{j\in[s_3]} \frac{d}{s_3}|\boldsymbol{\mu}^*_{3,j}|^2 = 1 \\
\Rightarrow \sum_{j\in[s_3]} |\boldsymbol{\mu}^*_{3,j}|^2 = \frac{s_3}{d}
\end{aligned}$$

We impose uniform weight difference operator, i.e., $\gamma^3_{i_1,i_2} = 1$. Due to the special structure for ${}^3\mathbf{\Delta}$, we have the following result for ${}^3\mathbf{\Delta}^\dagger$,

$${}^3\mathbf{\Delta}^\dagger = \frac{1}{d}\,{}^3\mathbf{\Delta}^\top \tag{A.2}$$

This results in $M = 2/d$. Thus, under Assumption 1 - 3, we have

$$\frac{w_{\max}(\epsilon_0^2\alpha + 2\epsilon_0\rho(K-1) + \rho^2(K-1)) + \psi}{2w_{\min}(1-\epsilon_0^2)} \lesssim \frac{K}{d} \tag{A.3}$$

Combining (A.1), (A.2) and (A.3), we have

$$1 - \frac{1}{s_3} \lesssim \frac{K}{\sqrt{d}}$$

Thus, when the cluster size s_3 increases, the clustering task becomes more challenging.

A.3 Supporting Lemmas

In this appendix we provide several supporting lemmas we use:

Lemma 1. *Consider the model $y = \beta^* + \epsilon$ with true parameter $\beta^* \in \mathbb{R}^d$ and any noise ϵ. Denote the fused lasso estimator as $\hat{\beta} := \arg\min_\beta \frac{1}{2}\|y-\beta\|_2^2 + \lambda\|\mathbf{D}\beta\|_1$. Denote $M := \max_j \|[\mathbf{D}^\dagger]_j\|_2$. If $\lambda \geq M\|\epsilon\|_2$, then we have*

$$\|\hat{\beta} - \beta^*\|_2^2 \leq \|\epsilon\|_2^2 + 4\lambda\|\mathbf{D}\beta^*\|_1$$

Lemma 1 provides the error bound of a fused lasso estimator which can be proved by similar arguments to the proof of Theorem 3 in Wang et al. [2016b].

Lemma 2. *Assume that each element in $\mathcal{E}$ is independent, zero-mean and satisfies $\mathbb{E}[e^{t\mathcal{E}_{ijk}}] \leq e^{\frac{\sigma^2 t^2}{2}}$, then spectral norm can be bounded as follows:*

$$\psi := \|\mathcal{E}\| \leq \sqrt{8\sigma^2((d_1+d_2+d_3)\log\frac{6}{\log 3/2} + \log\frac{2}{\delta})}$$

with probability at least $1-\delta$.

Proof of Lemma 2 follows from similar arguments to the proof if Theorem 1 in Tomioka and Suzuki [2014].

Lemma 3. *Consider algorithm Fused-Orth-ALS iterations and suppose the first* $(m-1)$ *columns in each factor matrices have converged. Without loss of generality, let* $\hat{\mathbf{C}}_{:i} = \mathbf{C}_{:i} + \hat{\xi}_i$, *where* $\|\hat{\xi}_i\|_2 \leq 8\sqrt{2}(\gamma\rho^2 K + \frac{\psi}{w_{\min}})$. *Let* $\{\bar{\mathbf{C}}_{:i}, i < m\}$ *denote an orthogonal basis for* $\{\bar{\mathbf{C}}_{:i}, i < m\}$, *then if Assumption 1-4 hold, we have*

$$\bar{\mathbf{C}}_{:i} = \mathbf{C}_{:i} + \xi_i, \;\; \|\xi_i\|_2 \leq 16\sqrt{2}\gamma\alpha K/\sqrt{d}, \;\; \forall i < m$$

Similar results applies for $\mathbf{A}$ *and* $\mathbf{B}$.

Proof. We will use induction to prove this result.

In the first step of proof for Theorem 1, we have shown that after one update,

$$\begin{aligned}\|\hat{\mathbf{C}}_{:1} - \mathbf{C}_{:1}\|^2 &\leq \frac{4\sqrt{2}w_{\max}}{w_{\min}}\rho^2 K + \frac{4\sqrt{2}}{w_{\min}}\psi + \frac{4\sqrt{2}w_{\max}}{w_{\min}}q\epsilon_0 \\ &\leq 4\sqrt{2}(\gamma\rho^2 K + \frac{\psi}{w_{\min}}) + \frac{1}{2}\epsilon_0\end{aligned}$$

If we keep update iteratively, we have $\|\hat{\mathbf{C}}_{:1} - \mathbf{C}_{:1}\|_2 \leq 8\sqrt{2}(\gamma\rho^2 K + \frac{\psi}{w_{\min}})$. Assumption 1 and 3 provide $\rho \leq \alpha/\sqrt{d}$ and $\psi \leq w_{\max}K/d$. Thus, $\|\hat{\mathbf{C}}_{:1} - \mathbf{C}_{:1}\|_2 = \|\hat{\xi}_i\|_2 \leq 16\sqrt{2}\gamma\alpha^2 K/d$. Since $\bar{\mathbf{C}}_{:1} = \hat{\mathbf{C}}_{:1}$, $\|\xi_i\| \leq 16\sqrt{2}\gamma\alpha K/d$. So the base case is correct. Assume the result is correct for the first $m-1$ columns in $\mathbf{C}$. After orthogonalization, the mth basis vector has the following form

$$\bar{\mathbf{C}}_{:m} = \frac{1}{\kappa}((\mathbf{C}_{:m} + \hat{\xi}_m) - \sum_{j<m}((\mathbf{C}_{:m} + \hat{\xi}_m)^\top \bar{\mathbf{C}}_{:j})\bar{\mathbf{C}}_{:j})$$

where κ is the normalizing constant. Define $\mu_{m,j} = \mathbf{C}_{:m}^\top(\mathbf{C}_{:j} + \xi_j) \leq 2\gamma\alpha/\sqrt{d}$ by using the

induction hypothesis. Thus, we can write

$$\kappa\bar{\mathbf{C}}_{:m} = \mathbf{C}_{:m} - \sum_{j<m}(\mathbf{C}_{:m}^{\top}(\mathbf{C}_{:j}+\xi_j))(\mathbf{C}_{:j}+\xi_j) + \hat{\xi}_m - \sum_{j<m}(\hat{\xi}_m^{\top}(\mathbf{C}_{:j}+\xi_j))(\mathbf{C}_{:j}+\xi_j)$$

It's easy to verify that $\|\hat{\xi}_\epsilon\|_2 := \|\hat{\xi}_m - \sum_{j<m}(\hat{\xi}_m^{\top}(\mathbf{C}_{:j}+\xi_j))(\mathbf{C}_{:j}+\xi_j)\|_2 \leq \|\hat{\xi}_m\|_2 \leq 16\sqrt{2}\gamma\alpha^2 K/d$ using projection principle. Then, we have

$$\kappa\bar{\mathbf{C}}_{:m} = \mathbf{C}_{:m} - \sum_{j<m}\mu_{m,j}\mathbf{C}_{:j} - \sum_{j<m}\mu_{m,j}\xi_j + \hat{\xi}_\epsilon := \mathbf{C}_{:m} + \xi'_m$$
$$\|\xi'_m\|_2 \leq 2\gamma\alpha K/\sqrt{d} + 2\gamma\alpha K/\sqrt{d} \times 16\sqrt{2}\gamma\alpha K/d + 16\sqrt{2}\gamma\alpha^2 K/d \leq 3\gamma\alpha K/\sqrt{d}$$

Then, $1 - 3\gamma\alpha K/\sqrt{d} \leq |\kappa| \leq 1 + 3\gamma\alpha K/\sqrt{d}$,

$$\bar{\mathbf{C}}_{:m} = \frac{1}{\kappa}(\mathbf{C}_{:m} + \xi'_m) = \mathbf{C}_{:m} - (1-\frac{1}{\kappa})\mathbf{C}_{:m} + \frac{1}{\kappa}\xi'_m = \mathbf{C}_{:m} + \xi_m$$

Since $1 - 3\gamma\alpha K/\sqrt{d} \leq |1/\kappa| \leq 1 + 6\gamma\alpha K/\sqrt{d}$, we have

$$\|\xi_m\| \leq 6\gamma\alpha K/\sqrt{d} + (1 + 6\gamma\alpha K/\sqrt{d})6\gamma\alpha K/\sqrt{d} \leq 16\sqrt{2}\gamma\alpha K/\sqrt{d}$$

□

A.4 Additional information for real data analysis

A.4.1 Symmetric tensor slices

For HCP dataset, it's obvious that the brain connectivity slice for each individual k, $\mathcal{Y}_{:,:,k}$, is a symmetric matrix. To make sure that the final cluster label assignments are consistent within the first and second modes, we need to make sure that the final estimate from Fused-Orth-ALS algorithm for factor matrices $\mathbf{A}, \mathbf{B}$ are the same. To achieve this, we only need to

impose an additional constrain on the initialization of $\mathbf{A}, \mathbf{B}$ that is $\mathbf{A}^{(0)} = \mathbf{B}^{(0)}$. Since we choose one tuning parameter λ for Fused step in 3 step, it's easy to derive that final estimate $\hat{\mathbf{A}} = \hat{\mathbf{B}}$.

A.4.2 Rank K and number of cluster choice

The rank K for Fused-Orth-ALS algorithm is chosen by elbow method (Figure A.1). We pick $K = 2$ which achieves the lowest recovery error for HCP dataset and $K = 7$, the elbow point for Nations dataset. The number of cluster based on Gap statistics are set to be 5 and

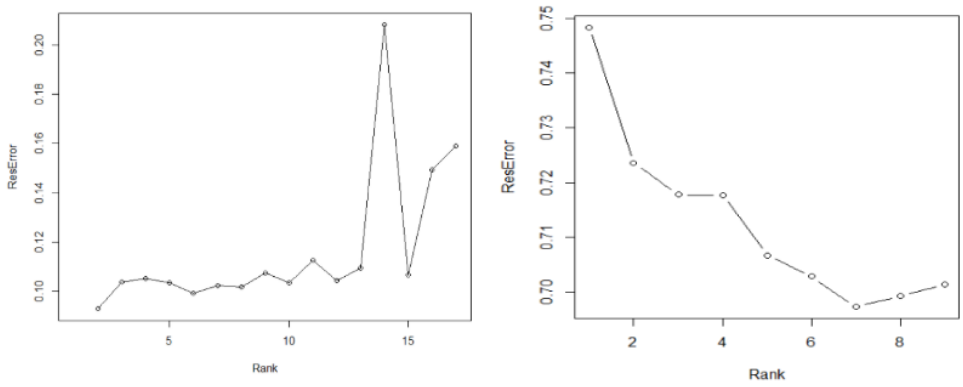


Figure A.1: Rank choice based on Relative error (Left: HCP, Right: Nations)

3 for HCP and nations dataset respectively (Figure A.2).

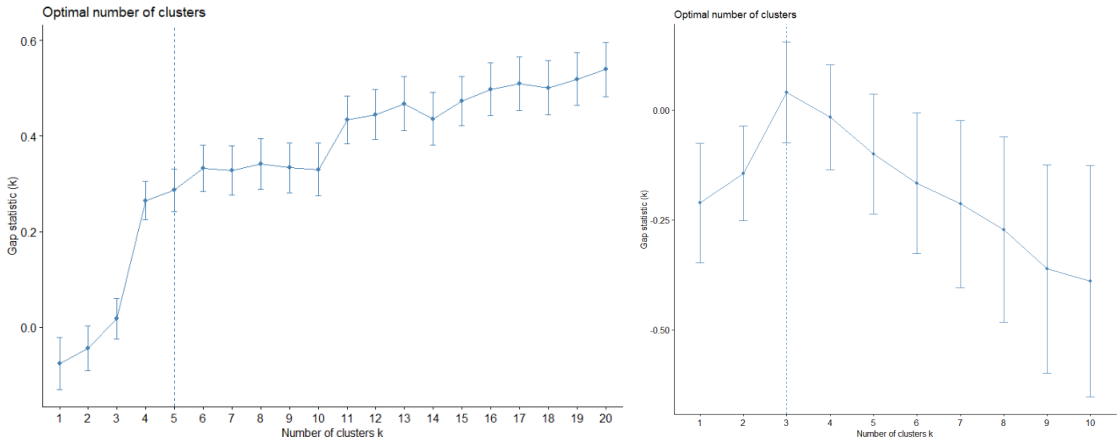


Figure A.2: Number of clusters based on gap statistics (Left: HCP, Right: Nations)

APPENDIX B

TECHNICAL DETAILS OF CHAPTER 3

B.1 Proof of Main Theorems

B.1.1 Proof of Proposition 1

Proof. First, we are going to show that the expectation of $L_{\text{IPS}}(\hat{\mathcal{X}}_{\text{Exp}}|\mathcal{P})$ is $L(\hat{\mathcal{X}}_{\text{Exp}})$.

$$\begin{aligned}
\mathbb{E}_{\mathcal{M}}(L_{\text{IPS}}(\hat{\mathcal{X}}_{\text{Exp}}|\mathcal{P})) &= \mathbb{E}_{\mathcal{M}}\Big(\frac{1}{\prod_{i=1}^{D} d_i} \sum_{(i_1,..,i_{k-1},i_{k+1},...,i_D)\in\mathbb{O}} \sum_{i_k=1}^{d_k} \frac{(\mathcal{X}_{i_1,...,i_D} - \hat{\mathcal{X}}_{i_1,...,i_D})^2}{\mathcal{P}_{i_1,..,i_{k-1},i_{k+1},...,i_D}}\Big) \\
&= \frac{1}{\prod_{i=1}^{D} d_i} \sum_{i_1=1}^{d_1} ... \sum_{i_D=1}^{d_D} \mathbb{E}\Big((i_1,...,i_{k-1},i_{k+1},...,i_D)\in\mathbb{O}\Big) \frac{(\mathcal{X}_{i_1,...,i_D} - \hat{\mathcal{X}}_{i_1,...,i_D})^2}{\mathcal{P}_{i_1,..,i_{k-1},i_{k+1},...,i_D}} \\
&= \frac{1}{\prod_{i=1}^{D} d_i} \sum_{i_1=1}^{d_1} ... \sum_{i_D=1}^{d_D} (\mathcal{X}_{i_1,...,i_D} - \hat{\mathcal{X}}_{i_1,...,i_D})^2 \\
&= L(\hat{\mathcal{X}}_{\text{Exp}})
\end{aligned}$$

Using assumptions of Theorem 1, we can easily find that $\frac{(\mathcal{X}_{i_1,...,i_D} - \hat{\mathcal{X}}_{i_1,...,i_D})^2}{\mathcal{P}_{i_1,..,i_{k-1},i_{k+1},...,i_D}}$ belongs to the range $[0, 4\psi_1^2/\|\mathcal{P}\|_{\min}]$. Now, by implementing Lemma 4, we can successfully derive that with probability at least $1-\delta$,

$$\begin{aligned}
|L_{\text{IPS}}(\hat{\mathcal{X}}_{\text{Exp}}|\mathcal{P}) - L(\hat{\mathcal{X}}_{\text{Exp}})| &\leq \frac{1}{\prod_{i=1}^{D} d_i} \sqrt{\frac{\sum_{i_1=1}^{d_1} \cdots \sum_{i_D=1}^{d_D} (4\psi_1^2/\|\mathcal{P}\|_{\min})^2}{2} \log\frac{2}{\delta}} \\
&= \frac{4\psi_1^2}{\|\mathcal{P}\|_{\min}} \sqrt{\frac{1}{2\prod_{i=1}^{D} d_i} \log\frac{2}{\delta}}
\end{aligned}$$

□

B.1.2 Proof of Theorem 3

Proof. Before we state the proof strategy in detail, we introduce some notations for convenience. Recall, when estimating the propensity score, we denote the optimization objective function as $\ell(\mathcal{S})$, defined in (3.7), which is a log likelihood function of parameter tensor $\mathcal{S}$. Similarly to Schnabel et al. [2016], we use the offset version of log likelihood function and denote it as

$$\begin{aligned}\bar{\ell}(\mathcal{S}) =& \ell(\mathcal{S}) - \ell(\mathbf{0}) \\ =& \sum_{(i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D)\in\mathbb{O}} \Big[\mathbf{1}_{\{\mathcal{M}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D}=1\}} \log \frac{f(\mathcal{S}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D})}{f(0)} + \\ & \mathbf{1}_{\{\mathcal{M}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D}=0\}} \log \frac{1-f(\mathcal{S}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D})}{1-f(0)}\Big]\end{aligned}$$

with a slight abuse of notation, we use $\mathbf{0}$ to represent an order $(D-1)$ tensor with all entries 0.

Next, we define several concepts of discrepancies between tensors and state some facts to illustrate their connections. In particular, we focus on the element-wise notion of discrepancy between two order $(D-1)$ tensor $\mathcal{A}_1$ and $\mathcal{A}_2$. For order $(D-1)$ tensor $\mathcal{A}_1, \mathcal{A}_2 \in \mathbb{R}^{d_1\times\ldots d_{k-1}\times d_{k+1}\ldots\times d_D}$ with entries within $[0,1]$, their Hellinger distance is defined as

$$d_H^2(\mathcal{A}_1;\mathcal{A}_2) = \frac{1}{\prod_{i\neq k} d_i} \sum_{i_1=1}^{d_1} \ldots \sum_{i_{k-1}=1}^{d_{k-1}} \sum_{i_{k+1}=1}^{d_{k+1}} \ldots \sum_{i_D=1}^{d_D} d_H^2(\mathcal{A}_{1,(i_1,\ldots,i_D)};\mathcal{A}_{2,(i_1,\ldots,i_D)})$$

where $d_H^2(a,b) = (\sqrt{a}-\sqrt{b})^2 + (\sqrt{1-a}-\sqrt{1-b})^2, a, b \in [0,1]$. Then, the Kullback-Leibler divergence between two tensor $\mathcal{A}_1, \mathcal{A}_2$ is defined by

$$D_{\mathrm{KL}}(\mathcal{A}_1||\mathcal{A}_2) = \frac{1}{\prod_{i\neq k} d_i} \sum_{i_1=1}^{d_1} \ldots \sum_{i_{k-1}=1}^{d_{k-1}} \sum_{i_{k+1}=1}^{d_{k+1}} \ldots \sum_{i_D=1}^{d_D} K(\mathcal{A}_{1,(i_1,\ldots,i_D)}||\mathcal{A}_{2,(i_1,\ldots,i_D)})$$

where $K(a||b) = a\log(a/b) + (1-a)\log((1-a)/(1-b)), a, b \in [0,1]$. By implementing the Jensen's inequality and an easy fact that $1 - x \leq -\log x, \forall x > 0$, we can establish the relationship between Hellinger distance and Kullback-Leibler divergence

$$d_H^2(\mathcal{A}_1; \mathcal{A}_2) \leq D_{\mathrm{KL}}(\mathcal{A}_1||\mathcal{A}_2)$$

By the optimality of $\hat{\mathcal{S}}_{\mathrm{MLE}}$ with respect to the optimization in 3.8, we can derive

$$\begin{aligned}
0 &\leq \ell(\hat{\mathcal{S}}_{\mathrm{MLE}}) - \ell(\mathcal{S}) = \bar{\ell}(\hat{\mathcal{S}}_{\mathrm{MLE}}) - \bar{\ell}(\mathcal{S}) \\
&= \underbrace{\bar{\ell}(\hat{\mathcal{S}}_{\mathrm{MLE}}) - \mathbb{E}_{\mathcal{M}}\bar{\ell}(\hat{\mathcal{S}}_{\mathrm{MLE}})}_{A1} - \underbrace{(\bar{\ell}(\mathcal{S}) - \mathbb{E}_{\mathcal{M}}\bar{\ell}(\mathcal{S}))}_{A2} + \underbrace{\mathbb{E}_{\mathcal{M}}\bar{\ell}(\hat{\mathcal{S}}_{\mathrm{MLE}}) - \mathbb{E}_{\mathcal{M}}\bar{\ell}(\mathcal{S})}_{A3}
\end{aligned}$$

Under assumption of Theorem 3, $\mathcal{S}, \hat{\mathcal{S}}_{\mathrm{MLE}} \in \mathcal{J}$ and thus, we can bound $A1 - A2$ by $\sup_{\mathcal{S}\in\mathcal{J}}[\bar{\ell}(\mathcal{S}) - \mathbb{E}_{\mathcal{M}}\bar{\ell}(\mathcal{S})]$. The last term can be analyzed as below,

$$\begin{aligned}
A3 = &\sum_{i_1=1}^{d_1} \cdots \sum_{i_{k-1}=1}^{d_{k-1}} \sum_{i_{k+1}=1}^{d_{k+1}} \cdots \sum_{i_D=1}^{d_D} \mathbb{E}_{\mathcal{M}}\Big[\mathcal{M}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D} \log \frac{f(\hat{\mathcal{S}}_{\mathrm{MLE},(i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D)})}{f(\mathcal{S}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D})} \\
&+ (1 - \mathcal{M}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D}) \log \frac{1 - f(\hat{\mathcal{S}}_{\mathrm{MLE},(i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D)})}{1 - f(\mathcal{S}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D})}\Big] \\
= &\sum_{i_1=1}^{d_1} \cdots \sum_{i_{k-1}=1}^{d_{k-1}} \sum_{i_{k+1}=1}^{d_{k+1}} \cdots \sum_{i_D=1}^{d_D} \Big[\mathcal{P}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D} \log \frac{f(\hat{\mathcal{S}}_{\mathrm{MLE},(i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D)})}{f(\mathcal{S}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D})} + \\
&(1 - \mathcal{P}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D}) \log \frac{1 - f(\hat{\mathcal{S}}_{\mathrm{MLE},(i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D)})}{1 - f(\mathcal{S}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D})}\Big] \\
= &\sum_{i_1=1}^{d_1} \cdots \sum_{i_{k-1}=1}^{d_{k-1}} \sum_{i_{k+1}=1}^{d_{k+1}} \cdots \sum_{i_D=1}^{d_D} \Big[\mathcal{P}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D} \log \frac{\hat{\mathcal{P}}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D}}{\mathcal{P}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D}} + \\
&(1 - \mathcal{P}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D}) \log \frac{1 - \hat{\mathcal{P}}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D}}{1 - \mathcal{P}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D}}\Big]
\end{aligned}$$

$$
\begin{aligned}
&= -\sum_{i_1=1}^{d_1} \cdots \sum_{i_{k-1}=1}^{d_{k-1}} \sum_{i_{k+1}=1}^{d_{k+1}} \cdots \sum_{i_D=1}^{d_D} \Big[\mathcal{P}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D} \log \frac{\mathcal{P}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D}}{\hat{\mathcal{P}}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D}} + \\
&\quad (1 - \mathcal{P}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D}) \log \frac{1 - \mathcal{P}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D}}{1 - \hat{\mathcal{P}}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D}} \Big] \\
&= - D_{\mathrm{KL}}(\mathcal{P} || \hat{\mathcal{P}})
\end{aligned}
$$

Under Lemma 5, it's easy to verify that $\|\mathcal{P}\|_\infty, \|\hat{\mathcal{P}}\|_\infty \leq 1$. We can see that if function $g(\cdot)$ is chosen to be $g(x) = x$,

$$
D_{\mathrm{KL}}(\mathcal{P} || \hat{\mathcal{P}}) \geq \frac{1}{\prod_{i \neq k} d_i} d_H^2(\mathcal{P}; \hat{\mathcal{P}}) \geq \frac{1}{2 \prod_{i \neq k} d_i} \|\mathcal{P} - \hat{\mathcal{P}}\|_F^2
$$

Thus, we can derive

$$
\frac{1}{\prod_{i \neq k} d_i} \|\mathcal{P} - \hat{\mathcal{P}}\|_F^2 \leq 4 \sup_{\mathcal{S} \in \mathcal{J}} [\bar{\ell}(\mathcal{S}) - \mathbb{E}_{\mathcal{M}} \bar{\ell}(\mathcal{S})]
$$

Our following analysis explains how to bound $\sup_{\mathcal{S} \in \mathcal{J}} [\bar{\ell}(\mathcal{S}) - \mathbb{E}_{\mathcal{M}} \bar{\ell}(\mathcal{S})]$. By Markov's inequality, for any $h > 0$ and $t > 0$, we have

$$
\begin{aligned}
\mathbb{P}(\sup_{\mathcal{S} \in \mathcal{J}} |\bar{\ell}(\mathcal{S}) - \mathbb{E}_{\mathcal{M}} \bar{\ell}(\mathcal{S})| \geq t) &= \mathbb{P}(\sup_{\mathcal{S} \in \mathcal{J}} |\bar{\ell}(\mathcal{S}) - \mathbb{E}_{\mathcal{M}} \bar{\ell}(\mathcal{S})|^h \geq t^h) \\
&\leq \frac{\mathbb{E}_{\mathcal{M}}[\sup_{\mathcal{S} \in \mathcal{J}} |\bar{\ell}(\mathcal{S}) - \mathbb{E}_{\mathcal{M}} \bar{\ell}(\mathcal{S})|^h]}{t^h}
\end{aligned}
$$

Applying standard symmetrization approach, we introduce $\mathcal{M}' \in \bigotimes_{i \neq k} \mathbb{R}^{d_i}$, which is also an order $(D-1)$ tensor but independently sampled from the same distribution as $\mathcal{M}$. Then

we can bound

$$\begin{aligned}\mathbb{E}_{\mathcal{M}}[\sup_{\mathcal{S}\in\mathcal{J}}|\bar{\ell}(\mathcal{S})-\mathbb{E}_{\mathcal{M}}\bar{\ell}(\mathcal{S})|^h] &= \mathbb{E}_{\mathcal{M}}[\sup_{\mathcal{S}\in\mathcal{J}}|\bar{\ell}(\mathcal{S})-\mathbb{E}_{\mathcal{M}'}[\bar{\ell}_{\mathcal{M}'}(\mathcal{S})]|^h] \\ &= \mathbb{E}_{\mathcal{M}}[\sup_{\mathcal{S}\in\mathcal{J}}|\mathbb{E}_{\mathcal{M}'}[\bar{\ell}(\mathcal{S})-\bar{\ell}_{\mathcal{M}'}(\mathcal{S})]|^h] \\ &\leq \mathbb{E}_{\mathcal{M}}\mathbb{E}_{\mathcal{M}'}[\sup_{\mathcal{S}\in\mathcal{J}}|\bar{\ell}(\mathcal{S})-\bar{\ell}_{\mathcal{M}'}(\mathcal{S})|^h] \\ &= \mathbb{E}_{\mathcal{M}',\mathcal{M}}[\sup_{\mathcal{S}\in\mathcal{J}}|\bar{\ell}(\mathcal{S})-\bar{\ell}_{\mathcal{M}'}(\mathcal{S})|^h]\end{aligned}$$

Here, we subscript $\bar{\ell}$ by $\mathcal{M}'$, representing we change the $\mathcal{M}$ in the original representation of $\bar{\ell}$ by $\mathcal{M}'$ since $\mathcal{M}'$ has the same distribution as $\mathcal{M}$.

$$\begin{aligned}&\bar{\ell}(\mathcal{S})-\bar{\ell}_{\mathcal{M}'}(\mathcal{S}) = \sum_{i_1=1}^{d_1}\dots\sum_{i_{k-1}=1}^{d_{k-1}}\sum_{i_{k+1}=1}^{d_{k+1}}\dots\sum_{i_D=1}^{d_D}[\mathcal{M}_{i_1,\dots,i_{k-1},i_{k+1},\dots,i_D}\log\frac{f(\hat{\mathcal{S}}_{\text{MLE},(i_1,\dots,i_{k-1},i_{k+1},\dots,i_D)})}{f(\mathcal{S}_{i_1,\dots,i_{k-1},i_{k+1},\dots,i_D})}+ \\ &(1-\mathcal{M}_{i_1,\dots,i_{k-1},i_{k+1},\dots,i_D})\log\frac{1-f(\hat{\mathcal{S}}_{\text{MLE},(i_1,\dots,i_{k-1},i_{k+1},\dots,i_D)})}{1-f(\mathcal{S}_{i_1,\dots,i_{k-1},i_{k+1},\dots,i_D})}]- \\ &[\mathcal{M}'_{i_1,\dots,i_{k-1},i_{k+1},\dots,i_D}\log\frac{f(\hat{\mathcal{S}}_{\text{MLE},(i_1,\dots,i_{k-1},i_{k+1},\dots,i_D)})}{f(\mathcal{S}_{i_1,\dots,i_{k-1},i_{k+1},\dots,i_D})}+ \\ &(1-\mathcal{M}'_{i_1,\dots,i_{k-1},i_{k+1},\dots,i_D})\log\frac{1-f(\hat{\mathcal{S}}_{\text{MLE},(i_1,\dots,i_{k-1},i_{k+1},\dots,i_D)})}{1-f(\mathcal{S}_{i_1,\dots,i_{k-1},i_{k+1},\dots,i_D})}] := \sum_{i_1=1}^{d_1}\dots\sum_{i_{k-1}=1}^{d_{k-1}}\sum_{i_{k+1}=1}^{d_{k+1}}\dots\sum_{i_D=1}^{d_D}[B1-B2]\end{aligned}$$

Next, we introduce independently and randomly sampled Rademacher random variables $\epsilon_{i_1,\dots,i_{k-1},i_{k+1},\dots,i_D}\in\{-1,+1\}$ with equal probability and we construct a new random variable

$$\sum_{i_1=1}^{d_1}\dots\sum_{i_{k-1}=1}^{d_{k-1}}\sum_{i_{k+1}=1}^{d_{k+1}}\dots\sum_{i_D=1}^{d_D}\epsilon_{i_1,\dots,i_{k-1},i_{k+1},\dots,i_D}[B1-B2]$$

which has the same distribution as $\bar{\ell}(\mathcal{S}) - \bar{\ell}_{\mathcal{M}'}(\mathcal{S})$. By further analysis, we can derive

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{M}',\mathcal{M}}[\sup_{\mathcal{S}\in\mathcal{J}} |\bar{\ell}(\mathcal{S}) - \bar{\ell}_{\mathcal{M}'}(\mathcal{S})|^h] \\
&= \mathbb{E}_{\mathcal{M}',\mathcal{M}}\Big[\sup_{\mathcal{S}\in\mathcal{J}} \Big| \sum_{i_1=1}^{d_1} ... \sum_{i_{k-1}=1}^{d_{k-1}} \sum_{i_{k+1}=1}^{d_{k+1}} ... \sum_{i_D=1}^{d_D} \epsilon_{i_1,...,i_{k-1},i_{k+1},...,i_D}[B1 - B2]\Big|\Big] \\
&\leq 2^{h-1}\mathbb{E}_{\mathcal{M}',\mathcal{M}}\Big[\sup_{\mathcal{S}\in\mathcal{J}} \Big(|\epsilon_{i_1,...,i_{k-1},i_{k+1},...,i_D}B1|^h + |\epsilon_{i_1,...,i_{k-1},i_{k+1},...,i_D}B2|^h\Big)\Big] \\
&= 2^{h-1}\Big\{\mathbb{E}_{\mathcal{M}',\mathcal{M}}\Big[\sup_{\mathcal{S}\in\mathcal{J}} \Big(|\epsilon_{i_1,...,i_{k-1},i_{k+1},...,i_D}B1|^h\Big)\Big] \\
&+ \mathbb{E}_{\mathcal{M}',\mathcal{M}}\Big[\sup_{\mathcal{S}\in\mathcal{J}} \Big(|\epsilon_{i_1,...,i_{k-1},i_{k+1},...,i_D}B2|^h\Big)\Big]\Big\} \\
&= 2^h\mathbb{E}_{\mathcal{M}}\Big[\sup_{\mathcal{S}\in\mathcal{J}} \Big(|\epsilon_{i_1,...,i_{k-1},i_{k+1},...,i_D}B1|^h\Big)\Big]
\end{aligned}
$$

where the second inequality comes from the fact that $|a+b|^p \leq 2^{p-1}(|a|^p + |b|^p)$ for $p \geq 1$. It's easy to prove that $\log f(x)/f(0)$ and $\log(1-f(x))/(1-f(0))$ are L_{ψ_2}-Lipschitz function over $[-\psi_2, \psi_2]$. Thus, by defining $\bar{\mathcal{M}}_{i_1,...,i_{k-1},i_{k+1},...,i_D} = 2\mathcal{M}_{i_1,...,i_{k-1},i_{k+1},...,i_D} - 1$ and employing Theorem 11.6 in Boucheron et al. [2003], we have

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{M}}\Big[\sup_{\mathcal{S}\in\mathcal{J}} \Big(|\epsilon_{i_1,...,i_{k-1},i_{k+1},...,i_D}B1|^h\Big)\Big] \\
=&(2L_{\psi_2})^h\mathbb{E}_{\mathcal{M}}\Big[\frac{1}{2L_{\psi_2}}\sup_{\mathcal{S}\in\mathcal{J}} \Big(|\epsilon_{i_1,...,i_{k-1},i_{k+1},...,i_D}B1|\Big)\Big]^h \\
\leq&(2L_{\psi_2})^h\mathbb{E}_{\mathcal{M}}\Big[\sup_{\mathcal{S}\in\mathcal{J}} \Big| \sum_{i_1=1}^{d_1} ... \sum_{i_{k-1}=1}^{d_{k-1}} \sum_{i_{k+1}=1}^{d_{k+1}} ... \sum_{i_D=1}^{d_D} \epsilon_{i_1,...,i_{k-1},i_{k+1},...,i_D} \\
&\Big[\mathcal{M}_{i_1,...,i_{k-1},i_{k+1},...,i_D}\mathcal{S}_{i_1,...,i_{k-1},i_{k+1},...,i_D} - (1-\mathcal{M}_{i_1,...,i_{k-1},i_{k+1},...,i_D})\mathcal{S}_{i_1,...,i_{k-1},i_{k+1},...,i_D}\Big]\Big|\Big] \\
=&(2L_{\psi_2})^h\mathbb{E}_{\mathcal{M}}\Big[\sup_{\mathcal{S}\in\mathcal{J}} \Big| \sum_{i_1=1}^{d_1} ... \sum_{i_{k-1}=1}^{d_{k-1}} \sum_{i_{k+1}=1}^{d_{k+1}} ... \sum_{i_D=1}^{d_D} \epsilon_{i_1,...,i_{k-1},i_{k+1},...,i_D} \\
&(\bar{\mathcal{M}}_{i_1,...,i_{k-1},i_{k+1},...,i_D}\mathcal{S}_{i_1,...,i_{k-1},i_{k+1},...,i_D})\Big|^h\Big] \\
=&(2L_{\psi_2})^h\mathbb{E}_{\mathcal{M}}\Big[\sup_{\mathcal{S}\in\mathcal{J}} |\langle \mathcal{G}\circ\bar{\mathcal{M}}, \mathcal{S}\rangle|^h\Big] \leq (2L_{\psi_2})^h\mathbb{E}_{\mathcal{M}}\Big[\sup_{\mathcal{S}\in\mathcal{J}} \|\mathcal{G}\circ\bar{\mathcal{M}}\|^h\|\mathcal{S}\|_*^h\Big]
\end{aligned}
$$

where $\mathcal{G}$ is an order $(D-1)$ tensor with each entry $\epsilon_{i_1,...,i_{k-1},i_{k+1},...,i_D}$ and the last inequality

can be obtained by Lemma 6.

By Lemma 9, we can successfully bound the spectral norm of $\mathcal{G} \circ \bar{\mathcal{M}}$ using

$$\mathbb{P}\Big(\|\mathcal{G} \circ \bar{\mathcal{M}}\|^h \geq \Big(C\sqrt{\sum_{i\neq k} d_i}\Big)^h\Big) \leq 2\exp\Big(-\log(D-1)\sum_{i\neq k} d_i\Big)$$

By Lemma 7, we can bound the nuclear norm of $\mathcal{S}$ by

$$\|\mathcal{S}\|_* \leq \sqrt{r_{2,\max}^{D-2}\prod_{i\neq k} d_i}\psi_2$$

Combining the above two results, we have with probability at least $1 - 2\exp\Big(-\log(D-1)\sum_{i\neq k} d_i\Big)$

$$\mathbb{E}_{\mathcal{M}}\Big(\Big[\sup_{\mathcal{S}\in\mathcal{J}}\Big(|\epsilon_{i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_D}B1|^h\Big]\Big) \leq \Big(2L_{\psi_2}C\psi_2\sqrt{r_{2,\max}^{D-2}\prod_{i\neq k} d_i\sum_{i\neq k} d_i}\Big)^h$$

Thus, by choosing $t = 2eCL_{\psi_2}\psi_2\sqrt{r_{2,\max}^{D-2}\prod_{i\neq k} d_i\sum_{i\neq k} d_i}$ and $h = \log\sum_{i\neq k} d_i$, we have

$$\begin{aligned}
&\mathbb{P}(\sup_{\mathcal{S}\in\mathcal{J}}|\bar{\ell}(\mathcal{S}) - \mathbb{E}_{\mathcal{M}}\bar{\ell}(\mathcal{S})| \geq t)\\
\leq&\frac{\mathbb{E}(\sup_{\mathcal{S}\in\mathcal{J}}|\bar{\ell}(\mathcal{S}) - \mathbb{E}_{\mathcal{M}}\bar{\ell}(\mathcal{S})|^h)}{t^h} \leq \frac{\Big(2L_{\psi_2}C\psi_2\sqrt{r_{2,\max}^{D-2}\prod_{i\neq k} d_i\sum_{i\neq k} d_i}\Big)^h}{\Big(2eCL_{\psi_2}\psi_2\sqrt{r_{2,\max}^{D-2}\prod_{i\neq k} d_i\sum_{i\neq k} d_i}\Big)^h}\\
\leq&\frac{1}{\sum_{i\neq k} d_i}
\end{aligned}$$

Finally, we have with probability at least $1 - 2\exp\Big(-\log(D-1)\sum_{i\neq k} d_i\Big) - 1/\sum_{i\neq k} d_i$

$$\frac{1}{\prod_{i\neq k} d_i}\|\mathcal{P} - \hat{\mathcal{P}}\|_F^2 \leq 4\sup_{\mathcal{S}\in\mathcal{J}}[\bar{\ell}(\mathcal{S}) - \mathbb{E}_{\mathcal{M}}\bar{\ell}(\mathcal{S})] \leq 8eCL_{\psi_2}\psi_2\sqrt{r_{2,\max}^{D-2}\frac{\sum_{i\neq k} d_i}{\prod_{i\neq k} d_i}}$$

□

B.1.3 Proof of Theorem 4

Proof. Using triangular inequality, we have

$$|L_{\text{IPS}}(\hat{\mathcal{X}}_{\text{Obs}}|\hat{\mathcal{P}}) - L(\hat{\mathcal{X}}_{\text{Obs}})| \leq \underbrace{|L_{\text{IPS}}(\hat{\mathcal{X}}_{\text{Obs}}|\hat{\mathcal{P}}) - L_{\text{IPS}}(\hat{\mathcal{X}}_{\text{Obs}}|\mathcal{P})|}_{C1} + \underbrace{|L_{\text{IPS}}(\hat{\mathcal{X}}_{\text{Obs}}|\mathcal{P}) - L(\hat{\mathcal{X}}_{\text{Obs}})|}_{C2}$$

$C2$ can be bounded using the similar strategy as we used in Theorem 1 and thus we should be able to obtain with probability at least $1-\delta$

$$|L_{\text{IPS}}(\hat{\mathcal{X}}_{\text{Obs}}|\mathcal{P}) - L(\hat{\mathcal{X}}_{\text{Obs}})| \leq \frac{4\psi_1'^2}{\|\mathcal{P}\|_{\min}}\sqrt{\frac{1}{2\prod_{i=1}^{D} d_i}\log\frac{2}{\delta}} \leq \frac{4\psi_1'^2}{f(-\psi_2)}\sqrt{\frac{1}{2\prod_{i=1}^{D} d_i}\log\frac{2}{\delta}}$$

Now it suffices to bound $C1$.

$$\begin{aligned}
C1 =& \Big|\frac{1}{\prod_{i=1}^{D} d_i}\sum_{(i_1,\ldots,i_{k-1},i_{k+1},\ldots i_D)\in\mathbb{O}}\sum_{i_k=1}^{d_k}\Big[\frac{(\hat{\mathcal{X}}_{\text{Obs},(i_1,\ldots,i_{k-1},i_{k+1},\ldots i_D)} - \mathcal{X}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots i_D})^2}{\hat{\mathcal{P}}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots i_D}} \\
& - \frac{(\hat{\mathcal{X}}_{\text{Obs},(i_1,\ldots,i_{k-1},i_{k+1},\ldots i_D)} - \mathcal{X}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots i_D})^2}{\mathcal{P}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots i_D}}\Big]\Big| \\
=& \Big|\frac{1}{\prod_{i=1}^{D} d_i}\sum_{(i_1,\ldots,i_{k-1},i_{k+1},\ldots i_D)\in\mathbb{O}}\sum_{i_k=1}^{d_k}\Big[\frac{\mathcal{P}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots i_D} - \hat{\mathcal{P}}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots i_D}}{\hat{\mathcal{P}}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots i_D}\mathcal{P}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots i_D}} \\
& \times (\hat{\mathcal{X}}_{\text{Obs},(i_1,\ldots,i_{k-1},i_{k+1},\ldots i_D)} - \mathcal{X}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots i_D})^2\Big]\Big|
\end{aligned}$$

As we know the following facts, $\forall(i_1,\ldots i_{k-1},i_{k+1},\ldots,i_D)\in\mathbb{O}$,

$$\begin{aligned}
\Big|\hat{\mathcal{X}}_{\text{Obs},(i_1,\ldots,i_{k-1},i_{k+1},\ldots i_D)} - \mathcal{X}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots i_D}\Big| &\leq 2\psi_1' \\
\hat{\mathcal{P}}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots i_D} &\geq f(-\psi_2) \\
\mathcal{P}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots i_D} &\geq f(-\psi_2)
\end{aligned}$$

We can further bound $C1$ by

$$
\begin{aligned}
C1 &\leq \frac{1}{\prod_{i=1}^{D} d_i} \frac{4\psi_1'^2}{f(-\psi_2)^2} \sum_{(i_1,\ldots,i_{k-1},i_{k+1},\ldots i_D)\in\mathbb{O}} \left|\hat{\mathcal{P}}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots i_D} - \mathcal{P}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots i_D}\right| \\
&\leq \frac{4\psi_1'^2}{f(-\psi_2)^2} \sqrt{\frac{|\mathbb{O}|}{\prod_{i\neq k} d_i}} \sqrt{\frac{\sum_{(i_1,\ldots,i_{k-1},i_{k+1},\ldots i_D)\in\mathbb{O}} (\hat{\mathcal{P}}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots i_D} - \mathcal{P}_{i_1,\ldots,i_{k-1},i_{k+1},\ldots i_D})^2}{\prod_{i\neq k} d_i}} \\
&\leq \frac{4\psi_1'^2}{f(-\psi_2)^2} 8eCL_{\psi_2}\psi_2 \sqrt{r_{2,\max}^{D-2} \frac{\sum_{i\neq k} d_i}{\prod_{i\neq k} d_i}}
\end{aligned}
$$

Thus, with a union bound, we can derive with probability at lest $1 - \delta - \frac{1}{\sum_{i\neq k} d_i} - 2\exp\left\{(-\log(D-1)\sum_{i\neq k} d_i)\right\}$,

$$
\begin{aligned}
|L_{\mathrm{IPS}}(\hat{\mathcal{X}}_{\mathrm{Obs}}|\hat{\mathcal{P}}) - L(\hat{\mathcal{X}}_{\mathrm{Obs}})| \leq & \frac{4\psi_1'^2}{f(-\psi_2)} \sqrt{\frac{1}{2\prod_{i=1}^{D} d_i} \log\frac{2}{\delta}} \\
& + \min\left\{\frac{16\psi_1'^2}{f(-\psi_2)^2}, \frac{32eCL_{\psi_2}\psi_1'^2\psi_2}{f(-\psi_2)^2} \sqrt{r_{2,\max}^{D-2} d_k \frac{\sum_{i\neq k} d_i}{\prod_{i=1}^{k} d_i}}\right\}
\end{aligned}
$$

□

B.1.4 Proof of Theorem 5

For notational simplicity, we introduce the following notation

- ω: shorthand for index in an order D tensor $\mathcal{X} \in \mathbb{R}^{d_1\times\ldots d_D}$, i.e, $\mathcal{X}_\omega = \mathcal{X}_{i_1,\ldots i_D}$.
- $\mathcal{O}$: order D tensor, where $\mathcal{O}_\omega = 1$ represents $\mathcal{X}_\omega$ is revealed and 0 vice versa. Obviously, $\mathcal{O}$ can be formulated via stacking observed fiber index, i.e., $\mathbb{O}$ by d_k times along the kth mode.
- $\tilde{\mathcal{P}}$: order D tensor, where $\tilde{\mathcal{P}}_\omega = \mathbb{P}(\mathcal{O}_\omega = 1)$, indicating the probability that the corresponding entry is revealed. Similarly, $\tilde{\mathcal{P}}$ can be derived via stacking $\mathcal{P}$ by d_k

times along the kth mode. Besides, we denote $\tilde{\mathcal{P}}_L$ as the minimum value of $\tilde{\mathcal{P}}$, which is equivalent to $\mathcal{P}_L$.

- $\tilde{\mathcal{S}}$: order D tensor, where $\tilde{\mathcal{P}}_\omega = f(\tilde{\mathcal{S}}_\omega)$, where f is the link function. Similarly, $\tilde{\mathcal{S}}$ can be derived via stacking $\mathcal{S}$ by d_k times along the kth mode.
- $\mathcal{A}^\dagger$: order D tensor with entry component computed as $\mathcal{A}^\dagger_\omega = \mathcal{A}^{-1}_\omega$.
- $\mathcal{A}^\ddagger$: order D tensor with entry component computed as $\mathcal{A}^\ddagger_\omega = \mathcal{A}^{-1/2}_\omega$.
- $\mathfrak{I}_\omega$: with a little bit of abuse of notation, $\mathfrak{I}_\omega$ represents an order D tensor with only entry at index ω equivalent to 1 and all the other entries are 0.

From the definition of $\hat{\mathcal{X}}_{\text{Obs}}$ and underlying assumption of $\mathcal{X}^*$, we have $\hat{\mathcal{X}}_{\text{Obs}}, \mathcal{X}^*$ are all feasible solutions and furthermore, we can write out L_{IPS} in the following format via hadamoard product among tensor (generalization from matrix hadamoard product),

$$\frac{1}{\prod_{i=1}^D d_i}\|\mathcal{O} * \hat{\tilde{\mathcal{P}}}^\ddagger * (\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X})\|_F^2 \leq \frac{1}{\prod_{i=1}^D d_i}\|\mathcal{O} * \hat{\tilde{\mathcal{P}}}^\ddagger * (\mathcal{X}^* - \mathcal{X})\|_F^2$$

for $\hat{\mathcal{X}}_{\text{Obs}}, \mathcal{X}^* \in \mathcal{H}_{\text{Obs}}$. $\hat{\tilde{\mathcal{P}}}$ denotes the estimated propensity score which is derived based on the proposed max-norm rank constrained maximum likelihood approach.

Plugging the fact that $\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X} = \hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^* + \mathcal{X}^* - \mathcal{X}$ into the left hand side of the above inequality, we can derive

$$\begin{aligned}
\frac{1}{\prod_{i=1}^D d_i}\|\mathcal{O} * \hat{\tilde{\mathcal{P}}}^\ddagger * (\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*)\|_F^2 &\leq \frac{2}{\prod_{i=1}^D d_i}\langle \mathcal{O} * \hat{\tilde{\mathcal{P}}}^\ddagger * (\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*), \mathcal{O} * \hat{\tilde{\mathcal{P}}}^\ddagger * (\mathcal{X} - \mathcal{X}^*)\rangle \\
&= \frac{2}{\prod_{i=1}^D d_i}\langle \mathcal{O} * \hat{\tilde{\mathcal{P}}}^\ddagger * (\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*), \mathcal{O} * \hat{\tilde{\mathcal{P}}}^\ddagger * \mathcal{E}\rangle \\
&= \frac{2}{\prod_{i=1}^D d_i}\langle \hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*, \mathcal{O} * \hat{\tilde{\mathcal{P}}}^\dagger * \mathcal{E}\rangle \qquad \text{(B.1)}
\end{aligned}$$

To quantify the distance between $\hat{\mathcal{X}}_{\text{Obs}}$ and $\mathcal{X}^*$, we need take estimation performance of $\hat{\tilde{\mathcal{P}}}$

into consideration and by doing a simple organization, the above inequality can be reformulated as

$$\begin{aligned}\frac{1}{\prod_{i=1}^{D} d_i}\|\mathcal{O} * \hat{\tilde{\mathcal{P}}}^{\ddagger} * (\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*)\|_F^2 &= \frac{1}{\prod_{i=1}^{D} d_i}\langle \hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*, \mathcal{O} * \hat{\tilde{\mathcal{P}}}^{\dagger}(\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*)\rangle \\ =&\frac{1}{\prod_{i=1}^{D} d_i}\Big[\langle \hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*, \mathcal{O} * (\hat{\tilde{\mathcal{P}}}^{\dagger} - \tilde{\mathcal{P}}^{\dagger}) * (\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*)\rangle + \langle \hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*, \mathcal{O} * \tilde{\mathcal{P}}^{\dagger} * (\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*)\rangle\Big]\end{aligned} \tag{B.2}$$

Thus, combining equation (B.2) and inequality (B.1), it's natural to have

$$\begin{aligned}&\frac{1}{\prod_{i=1}^{D} d_i}\|\mathcal{O} * \tilde{\mathcal{P}}^{\ddagger} * (\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*)\|_F^2 \\ &\quad\leq \frac{1}{\prod_{i=1}^{D} d_i}\langle \hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*, \mathcal{O} * (\hat{\tilde{\mathcal{P}}}^{\dagger} - \tilde{\mathcal{P}}^{\dagger}) * (\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*)\rangle + \frac{2}{\prod_{i=1}^{D} d_i}\langle \hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*, \mathcal{O} * \hat{\tilde{\mathcal{P}}}^{\dagger} * \mathcal{E}\rangle \\ &\quad\leq \frac{1}{\prod_{i=1}^{D} d_i}\|\mathcal{O} * (\hat{\tilde{\mathcal{P}}}^{\dagger} - \tilde{\mathcal{P}}^{\dagger}) * (\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*)\|\|\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*\|_* + \frac{2}{\prod_{i=1}^{D} d_i}\|\mathcal{O} * \hat{\tilde{\mathcal{P}}}^{\dagger} * \mathcal{E}\|\|\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*\|_* \\ &\quad\leq (2\Gamma^{(1)} + \Gamma^{(2)})\|\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*\|_* := \Gamma^{(3)}\|\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*\|_*\end{aligned}$$

The last inequality is derived based on Lemma 10 and 11. Utilizing Lemma 7, the relationship between tensor nuclear norm and frobenious norm is stated as

$$\|\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*\|_* \leq \sqrt{\frac{\prod_{i=1}^{D} r'_{1,i}}{\max_i r'_{1,i}}}\|\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*\|_F$$

Denote $r'_{1,\max} = \max_i r'_{1,i}$ and then we have

$$\frac{1}{\prod_{i=1}^{D} d_i}\|\mathcal{O} * \tilde{\mathcal{P}}^{\ddagger} * (\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*)\|_F^2 \leq \Gamma^{(3)}(r'_{1,\max})^{(D-1)/2}\|\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*\|_F$$

We consider the following constrain set, for $\mathbf{r} = [r_1, ..., r_D]$, let $r_{\max} = \max_{i\in\{1,2,...,D\}} r_i$

$$\mathcal{C}(\mathbf{r}) = \Big\{\mathcal{A} \in \bigotimes_{i=1}^{D} \mathbb{R}^{d_i} : \|\mathcal{A}\|_\infty = 1, \mathbb{E}(\|\mathcal{O} * \mathcal{A}\|_F^2) \geq \sqrt{\frac{64\log(\sum_{i=1}^D d_i)}{\log(6/5)|\mathcal{O}|}}, \|\mathcal{A}\|_* \leq r_{\max}^{(D-1)/2}\|\mathcal{A}\|_F\Big\} \tag{B.3}$$

Consider the first case when $\mathbb{E}(\|\mathcal{O} * (\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*)\|_F^2) \leq \|\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*\|_\infty^2 \sqrt{\frac{64\log(\sum_{i=1}^D d_i)}{\log(6/5)|\mathcal{O}|}}$ and $\tilde{\mathcal{P}}_L \geq 1/\mu\prod_{i=1}^D d_i$, utilizing the fact that $\|\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*\|_\infty \leq 2\psi_1'$ implies that

$$\frac{1}{\prod_{i=1}^D d_i}\|\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*\|_F^2 \leq 4(\psi_1')^2\sqrt{\frac{64\log(\sum_{i=1}^D d_i)}{\log(6/5)|\mathcal{O}|}}$$

Otherwise, if $\mathbb{E}(\|\mathcal{O} * (\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*)\|_F^2) \geq \|\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*\|_\infty^2 \sqrt{\frac{64\log(\sum_{i=1}^D d_i)}{\log(6/5)|\mathcal{O}|}}$, following the fact that $1/\|\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*\|_\infty(\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*) \in \mathcal{C}(r_1')$, we can obtain the following inequality using Lemma 13,

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}(\|\mathcal{O} * (\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*)\|_F^2) \leq \frac{1}{\prod_{i=1}^D d_i}\|\mathcal{O} * (\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*)\|_F^2 + C(r_{1,\max}')^{D-1}\mu\prod_{i=1}^D d_i\mathbb{E}(\|\varrho^{(3)}\|)^2 \\
&\leq \frac{1}{\prod_{i=1}^D d_i}\|\mathcal{O} * \tilde{\mathcal{P}}^\ddagger * (\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*)\|_F^2 + C(r_{1,\max}')^{D-1}\mu\prod_{i=1}^D d_i\mathbb{E}(\|\varrho^{(3)}\|)^2 \\
&\leq \Gamma^{(3)}(r_{1,\max}')^{(D-1)/2}\|\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*\|_F + CC_1(r_{1,\max}')^{D-1}\mu\prod_{i=1}^D d_i\frac{\max_i\{d_1, ..., d_D\}\log(\sum_{i=1}^D d_i)}{\prod_{i=1}^D d_i|\mathcal{O}|} \\
&\leq (\Gamma^{(3)})^2(r_{1,\max}')^{D-1}\mu\prod_{i=1}^D d_i + \frac{1}{4\mu\prod_{i=1}^D d_i}\|\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*\|_F^2 \\
&\quad + CC_1(r_{1,\max}')^{D-1}\mu\prod_{i=1}^D d_i\frac{\max_i\{d_1, ..., d_D\}\log(\sum_{i=1}^D d_i)}{\prod_{i=1}^D d_i|\mathcal{O}|}
\end{aligned}$$

The second inequality is derived by the fact $\tilde{\mathcal{P}}_\omega \leq 1$, the third inequality implements Lemma 13 and by noticing if $f(x) = ax - bx^2$, $f(x) \leq a^2/2b$ holds for all x, the fourth inequality can be derived by setting $a = \Gamma^{(3)}(r_{1,\max}')^{(D-1)/2}$ and $b = (4\mu\prod_{i=1}^D d_i)^{-1}$. Replacing the left hand side of the above inequality as $\frac{1}{2\mu\prod_{i=1}^D d_i}\|\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*\|_F^2$ by using the

$\mathbb{E}(\|\mathcal{O} * (\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*)\|_F^2) \geq 1/(\mu \prod_{i=1}^D d_i)\|\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*\|_F^2$, we can further get

$$\begin{aligned}\frac{1}{\prod_{i=1}^D d_i}\|\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*\|_F^2 \leq & 4(\Gamma^{(3)})^2 (r'_{1,\max})^{D-1}\mu^2 \prod_{i=1}^D d_i + \frac{1}{4\mu\prod_{i=1}^D d_i}\|\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*\|_F^2 \\ & + 4CC_1 (r'_{1,\max})^{D-1}\mu^2 \prod_{i=1}^D d_i \frac{\max_i\{d_1,...,d_D\}\log(\sum_{i=1}^D d_i)}{\prod_{i=1}^D d_i |\mathcal{O}|} \\ \leq & 4\mu^2 (r'_{1,\max})^{D-1}\Big(\Gamma + CC_1 \frac{\max_i\{d_1,...,d_D\}\log(\sum_{i=1}^D d_i)}{|\mathcal{O}|}\Big)\end{aligned}$$

where $\Gamma \asymp \max\Big\{\frac{\max\{d_1,...,d_D\}\log(\sum_{i=1}^D d_i)}{\mathcal{P}_L \prod_{i=1}^D d_i}, \frac{(\log(\sum_{i=1}^D d_i))^2}{\mathcal{P}_L^2 \prod_{i=1}^D d_i}\Big\} + r_{2,\max}^{D-1}\frac{\sum_{i=1}^D d_i (\log(\sum_{i=1}^D d_i))^{1/2}}{\prod_{i=1}^D d_i}$.

Thus, we can see that as long as $\frac{\max_i\{d_1,...,d_D\}\log(\sum_{i=1}^D d_i)}{|\mathcal{O}|} \to 0$ as $d_i \to 0$, $\frac{1}{\prod_{i=1}^D d_i}\|\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*\|_F^2 \to 0$.

B.2 Supporting Lemmas

Lemma 4. *(Hoeffding's Inequality) Let $X_1, ..., X_n$ be independent random variables such that $a_i \leq X_i \leq b_i$ almost surely. Consider the sum of these random variables, $S_n = \sum_{i=1}^n X_i$, then for all $t > 0$,*

$$\mathbb{P}(|S_n - \mathbb{E}(S_n)| \geq t) \leq 2\exp\Big(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big)$$

Lemma 5. *Let g be a differentiable function and let $\mathcal{A}, \hat{\mathcal{A}}$ be order D tensor with dimension $d_1 \times d_2 .... \times d_D$ with $\|\mathcal{A}\|_\infty, \|\hat{\mathcal{A}}\|_\infty \leq \alpha$. Then*

$$d_H^2(g(\mathcal{A}), g(\hat{\mathcal{A}})) \geq \inf_{|\xi| \leq \alpha} \frac{(g'(\xi))^2}{8g(\xi)(1-g(\xi))}\frac{\|\mathcal{A} - \hat{\mathcal{A}}\|_F^2}{\prod_{i=1}^D d_i}$$

This Lemma is a straightforward generalization from matrix to tensor case and thus we omit the proof. For proof in matrix case, please check Lemma 2 in Davenport et al. [2014].

Lemma 6. *Let $\mathcal{A}_1, \mathcal{A}_2 \in \mathbb{R}^{d_1 \times ... \times d_D}$ be two order D tensor with the same dimension. Then*

$$|\langle \mathcal{A}_1, \mathcal{A}_2 \rangle| \leq \|\mathcal{A}_1\| \|\mathcal{A}_2\|_*$$

The proof for this Lemma can be found in Friedland and Lim [2018].

Lemma 7. *For order D tensor $\mathcal{A} \in \mathbb{R}^{d_1 \times ... \times d_D}$ with tucker rank rank$(\mathcal{A}) = (r_1, ..., r_D)$, suppose $\|\mathcal{A}\|_\infty \leq \psi_2$, then we have*

$$\|\mathcal{A}\|_* \leq \sqrt{r_{\max}^{D-1} \prod_{i=1}^{D} d_i \psi_2}$$

Proof. By Lemma 3 in Lee and Wang [2020], we can derive

$$\|\mathcal{A}\|_* \leq \sqrt{\frac{\prod_{i=1}^{D} r_i}{r_{\max}}} \|\mathcal{A}\|_F$$

Combining with the assumption that $\|\mathcal{A}\|_\infty \leq \psi_2$, we can easily get

$$\|\mathcal{A}\|_* \leq \sqrt{\frac{\prod_{i=1}^{D} r_i}{r_{\max}}} \sqrt{\prod_{i=1}^{D} d_i \psi_2} \leq \sqrt{r_{\max}^{D-1} \prod_{i=1}^{D} d_i \psi_2}$$

□

Lemma 8. *Suppose that $\mathcal{A} \in \mathbb{R}^{d_1 \times ... \times d_D}$ is an order D tensor with entries as i.i.d independent random variables that satisfies*

$$\mathbb{E}(\mathcal{A}_{i_1,...,i_D}) = 0 \quad \textit{and} \quad \mathbb{E}(e^{t\mathcal{A}_{i_1,...,i_D}}) \leq e^{t^2 L^2/2}$$

Then the spectral norm of $\mathcal{A}$ satisfies

$$\|\mathcal{A}\| \leq \sqrt{8L^2 \log(12D) \sum_{i=1}^{D} d_i + \log(2/\delta)}$$

with probability at least $1 - \delta$.

Please refer to Theorem 1 in [Tomioka and Suzuki, 2014].

Lemma 9. *There exist constant C such that for any random order-D tensor $\mathcal{A} \in \mathbb{R}^{d_1 \times \ldots \times d_D}$ with each component $\mathcal{A}_{i_1,\ldots,i_D}$ are i.i.d symmetric random variables with mean 0 and its absolute value is bounded by 1, then we have*

$$\mathbb{P}\Big(\|\mathcal{A}\| \geq C\sqrt{\sum_{i=1}^{D} d_i}\Big) \leq 2\exp\Big(-\log D \sum_{i=1}^{D} d_i\Big)$$

where $C = 8\log 12 + 9\log D$.

Proof. Note that $\mathcal{A}_{i_1,\ldots,i_D}$ is zero-mean and supported on $[-1, 1]$. Therefore, $\mathcal{A}_{i_1,\ldots,i_D}$ is a sub-Gaussian distributed random variable with parameter 1. Thus

$$\mathbb{E}(\mathcal{A}_{i_1,\ldots,i_D}) = 0 \quad \text{and} \quad \mathbb{E}(e^{t\mathcal{A}_{i_1,\ldots,i_D}}) \leq e^{t^2/2}$$

Then, implementing Lemma 8, with probability $1 - \delta$,

$$\|\mathcal{A}\| \leq \sqrt{8\log(12D) \sum_{i=1}^{D} d_i + \log(2/\delta)}$$

Taking $\delta = 2\exp(-\log D\sum_{i=1}^{D} d_i)$, we have

$$\begin{aligned}\|\mathcal{A}\| &\leq \sqrt{(8\log 12 + 8\log D)\sum_{i=1}^{D} d_i + \log\exp(\log D\sum_{i=1}^{D})d_i} \\ &= \sqrt{(8\log 12 + 9\log D)\sum_{i=1}^{D} d_i}\end{aligned}$$

Setting $C = 8\log 12 + 9\log D$ we will get the final result. □

Lemma 10. *Assume assumptions 5,6,7 and 8 hold and denote*

$$\varrho^{(1)} = \sum_{\omega\in\mathcal{O}} \mathcal{O}_\omega\mathcal{E}_\omega\mathfrak{I}_\omega/\hat{\tilde{\mathcal{P}}}_\omega/\prod_{i=1}^{D} d_i,$$

there exists

$$\Gamma^{(1)} \asymp \max\Big\{\frac{\sqrt{\max\{d_1,...,d_D\}\log(\sum_{i=1}^{D} d_i)}}{\sqrt{\mathcal{P}_L}\prod_{i=1}^{D} d_i}, \frac{\log(\sum_{i=1}^{D} d_i)}{\mathcal{P}_L\prod_{i=1}^{D} d_i}\Big\} + r_{2,\max}^{(D-1)/2}\frac{\sqrt{\sum_{i=1}^{D} d_i}}{\prod_{i=1}^{D} d_i}$$

such that $\|\varrho^{(1)}\| \leq \Gamma^{(1)}$ holds with high probability.

Proof. Denote $\varrho^{(1)} = \frac{1}{\prod_{i=1}^{D} d_i}\sum_{\omega\in\mathcal{O}} \mathcal{O}_\omega\mathcal{E}_\omega\mathfrak{I}_\omega/\hat{\tilde{\mathcal{P}}}_\omega$ and by triangular inequality, it's straightforward to have

$$\begin{aligned}\|\varrho^{(1)}\| &= \frac{1}{\prod_{i=1}^{D} d_i}\|\sum_{\omega\in\mathcal{O}} \mathcal{O}_\omega\mathcal{E}_\omega\mathfrak{I}_\omega/\hat{\tilde{\mathcal{P}}}_\omega\| \\ &\leq \frac{1}{\prod_{i=1}^{D} d_i}\|\sum_{\omega\in\mathcal{O}} \mathcal{O}_\omega\mathcal{E}_\omega\mathfrak{I}_\omega(1/\hat{\tilde{\mathcal{P}}}_\omega - 1/\tilde{\mathcal{P}}_\omega)\| + \frac{1}{\prod_{i=1}^{D} d_i}\|\sum_{\omega} \mathcal{O}_\omega\mathcal{E}_\omega\mathfrak{I}_\omega/\tilde{\mathcal{P}}_\omega\|\end{aligned}$$

At the first step, we consider to bound the second term on the right hand side of the above inequality. Under the assumption that $\mathcal{E}_\omega$ is zero-mean and independent of $\mathcal{O}_\omega$, we can derive

$$\mathbb{E}(\mathcal{O}_\omega\mathcal{E}_\omega\mathfrak{I}_\omega/\tilde{\mathcal{P}}_\omega) = \mathbb{E}(\mathcal{E}_\omega)\mathbb{E}(\mathcal{O}_\omega\mathfrak{I}_\omega/\tilde{\mathcal{P}}_\omega) = 0$$

Furthermore, by implementing law of total expectation, we have $\forall l = 2, 3, 4, ...$

$$\begin{aligned}\mathbb{E}(\mathcal{O}_\omega \mathcal{E}_\omega / \tilde{\mathcal{P}}_\omega)^l = \mathbb{E}\Big(\mathbb{E}(\mathcal{O}_\omega \mathcal{E}_\omega / \tilde{\mathcal{P}}_\omega | \mathcal{O}_\omega)^l\Big) &\leq \mathbb{E}\Big\{\frac{l!}{2}\Big(\frac{c_\sigma \mathcal{O}_\omega}{\tilde{\mathcal{P}}_\omega}\Big)^2 \Big(\frac{\eta \mathcal{O}_\omega}{\tilde{\mathcal{P}}_\omega}\Big)^{l-2}\Big\} \\ &\leq \frac{l!}{2}(c_\sigma / \sqrt{\tilde{\mathcal{P}}_L})^2 \eta_H^{l-2}\end{aligned}$$

where $\eta_H = \eta / \tilde{\mathcal{P}}_L$. Next, we will adopt the proof strategy of Theorem 6.2 in Tropp [2012] which is the matrix Bernstein inequality for the sub-exponential case. Similarly, we generalize the self-adjoint dilation of matrix to tensor which concatenate the original tensor into a larger dimension with $\sum_{i=1}^D d_i$ along each mode and denote it as $\mathcal{L}$. Under the assumption on error tensor $\mathcal{E}_\omega$, we have

$$\mathbb{E}\Big\{\mathcal{L}(\mathcal{O}_\omega \mathcal{E}_\omega \mathcal{I}_\omega / \tilde{\mathcal{P}}_\omega)^l\Big\} \leq \frac{l!}{2}\eta_H^{l-2}\mathcal{L}(\frac{c_\sigma}{\sqrt{\tilde{\mathcal{P}}_\omega}}\mathcal{I}_\omega)^2$$

One step further gives us

$$\sigma_H^2 = \|\sum_\omega \mathcal{L}(\frac{c_\sigma}{\sqrt{\tilde{\mathcal{P}}_\omega}}\mathcal{I}_\omega)^2\| = c_\sigma^2 \|\sum_\omega \mathcal{L}(\mathcal{I}_\omega)^2\| / \tilde{\mathcal{P}}_L \leq c_\sigma^2 \max_i\{d_1, ..., d_D\} / \tilde{\mathcal{P}}_L$$

where the last inequality comes from proof of Lemma S4.1 in Mao et al. [2019]. Thus, implementing Theorem 6.2 in Tropp [2012], we could show that,

$$\begin{aligned}&\mathbb{P}(\|\sum_\omega \mathcal{O}_\omega \mathcal{E}_\omega \mathcal{I}_\omega / \tilde{\mathcal{P}}_\omega\| \geq t) \\ \leq &\sum_{i=1}^D d_i \exp\Big\{\frac{-t^2/2}{c_\sigma^2 \max_i\{d_1, ..., d_D\}/\tilde{\mathcal{P}}_L + \eta_H t}\Big\} \\ = &\begin{cases}\sum_{i=1}^D d_i \exp\Big\{\frac{-t^2}{4c_\sigma^2 \max_i\{d_1, ..., d_D\}/\tilde{\mathcal{P}}_L}\Big\} & t \leq c_\sigma^2 \max_i\{d_1, ..., d_D\}/(\tilde{\mathcal{P}}_L \eta_H) \\ \sum_{i=1}^D d_i \exp\Big\{\frac{-t}{4\eta_H}\Big\} & t > c_\sigma^2 \max_i\{d_1, ..., d_D\}/(\tilde{\mathcal{P}}_L \eta_H)\end{cases}\end{aligned}$$

In other words, with probability at least $1-\exp\{-s\}$, we have

$$\|\sum_{\omega}\mathcal{O}_{\omega}\mathcal{E}_{\omega}\mathfrak{I}_{\omega}/\tilde{\mathcal{P}}_{\omega}\|\leq\max\left\{2c_{\sigma}\sqrt{\frac{\max_i\{d_1,...,d_D\}(s+\log(\sum_{i=1}^{D}d_i))}{\tilde{\mathcal{P}}_L}},4\eta_H(s+\log(\sum_{i=1}^{D}d_i))\right\}$$

By setting $s=\log(\sum_{i=1}^{D}d_i)$, the above inequality can be simplified as with probability at least $1-1/\sum_{i=1}^{D}d_i$,

$$\frac{1}{\prod_{i=1}^{D}d_i}\|\sum_{\omega}\mathcal{O}_{\omega}\mathcal{E}_{\omega}\mathfrak{I}_{\omega}/\tilde{\mathcal{P}}_{\omega}\|\leq\max\left\{2c_{\sigma}\frac{\sqrt{2\max_i\{d_1,...,d_D\}\log(\sum_{i=1}^{D}d_i)}}{\sqrt{\tilde{\mathcal{P}}_L}\prod_{i=1}^{D}d_i},\frac{8\eta_H\log(\sum_{i=1}^{D}d_i)}{\prod_{i=1}^{D}d_i}\right\}$$

Next, it suffices to show $\frac{1}{\prod_{i=1}^{D}d_i}\|\sum_{\omega}\mathcal{O}_{\omega}\mathcal{E}_{\omega}\mathfrak{I}_{\omega}/(1/\hat{\tilde{\mathcal{P}}}_{\omega}-1/\tilde{\mathcal{P}}_{\omega})\|$ is bounded. Similarly, we can show that $\forall t>0$, the following inequality holds

$$\begin{aligned}\|\sum_{\omega}\mathcal{O}_{\omega}\mathcal{E}_{\omega}\mathfrak{I}_{\omega}(1/\hat{\tilde{\mathcal{P}}}_{\omega}-1/\tilde{\mathcal{P}}_{\omega})\|^2&\leq\|\sum_{\omega}\mathcal{O}_{\omega}\mathcal{E}_{\omega}\mathfrak{I}_{\omega}(1/\hat{\tilde{\mathcal{P}}}_{\omega}-1/\tilde{\mathcal{P}}_{\omega})\|_F^2=\sum_{\omega}\mathcal{E}_{\omega}^2\mathcal{O}_{\omega}^2(1/\hat{\tilde{\mathcal{P}}}_{\omega}-1/\tilde{\mathcal{P}}_{\omega})^2\\&\leq\sum_{\omega}\mathcal{E}_{\omega}^2(1/\hat{\tilde{\mathcal{P}}}_{\omega}-1/\tilde{\mathcal{P}}_{\omega})^2\leq\max_{\omega}\{\mathcal{E}_{\omega}^2\}\sum_{\omega}(1/\hat{\tilde{\mathcal{P}}}_{\omega}-1/\tilde{\mathcal{P}}_{\omega})^2\\&=\max_{\omega}\{\mathcal{E}_{\omega}^2\}\|\hat{\tilde{\mathcal{P}}}^{\dagger}-\tilde{\mathcal{P}}^{\dagger}\|_F^2\end{aligned}\tag{B.4}$$

Combining the Markov inequality and assumptions on error tensor $\mathcal{E}$, we can derive for all $a>0$,

$$\mathbb{P}(\max_{\omega}\{\mathcal{E}_{\omega}^2\}\geq a)=\mathbb{P}(\max_{\omega}\{\mathcal{E}_{\omega}^4\}\geq a^2)\leq\frac{\max_{\omega}\mathbb{E}(\mathcal{E}_{\omega}^4)}{a^2}\leq\frac{12c_{\sigma}^2\eta^2}{a^2}\tag{B.5}$$

By choosing $a^2=\log(\sum_{i=1}^{D}d_i)$, we have $\max_{\omega}\{\mathcal{E}_{\omega}^2\}\leq\sqrt{\log(\sum_{i=1}^{D}d_i)}$ with probability at least $1-12c_{\sigma}^2\eta^2/\log(\sum_{i=1}^{D}d_i)$.

Implementing Taylor's theorem, we have there exists some ξ between $\hat{\tilde{\mathcal{P}}}_{\omega}$ and $\tilde{\mathcal{P}}_{\omega}$ such that

$$1/\hat{\tilde{\mathcal{P}}}_{\omega}-1/\tilde{\mathcal{P}}_{\omega}=-\frac{f'(\xi)}{f(\xi)^2}(\hat{\tilde{\mathcal{S}}}_{\omega}-\tilde{\mathcal{S}}_{\omega})$$

Recall the estimation of $\hat{\tilde{\mathcal{S}}}$ requires $|\xi| \leq \psi_2$, we have

$$\frac{f'(\xi)}{f(\xi)^2} \leq \frac{1-f(-\psi_2)}{f(-\psi_2)} \sup_{|\xi|\leq\psi_2} \frac{|f'(\xi)|}{f(\xi)(1-f(\xi))} = \frac{1-f(-\psi_2)}{f(-\psi_2)} L_{\psi_2}$$

Thus, implementing result in Lemma 12, with probability at least $1-\exp\{-c_2\sum_{i=1}^D d_i\}$,

$$\|\hat{\tilde{\mathcal{P}}}^\dagger - \tilde{\mathcal{P}}^\dagger\|_F^2 \leq \Big(\frac{1-f(-\psi_2)}{f(-\psi_2)}\Big)^2 L_{\psi_2}^2 \|\hat{\tilde{\mathcal{S}}} - \tilde{\mathcal{S}}\|_F^2 \leq \Big(\frac{1-f(-\psi_2)}{f(-\psi_2)}\Big)^2 L_{\psi_2}^2 c_1 r_{2,\max}^{D-1} \frac{L_{\psi_2}^2}{U_{\psi_2}^2} \sum_{i=1}^D d_i \tag{B.6}$$

Through inequality in (B.5) and (B.6), (B.4) can be upper bounded by

$$\|\sum_\omega \mathcal{O}_\omega \mathcal{E}_\omega \mathfrak{I}_\omega (1/\hat{\tilde{\mathcal{P}}}_\omega - 1/\tilde{\mathcal{P}}_\omega)\|^2 \leq \sqrt{\log(\sum_{i=1}^D d_i)} \Big(\frac{1-f(-\psi_2)}{f(-\psi_2)}\Big)^2 L_{\psi_2}^2 c_1 r_{2,\max}^{D-1} \frac{L_{\psi_2}^2}{U_{\psi_2}^2} \sum_{i=1}^D d_i$$

In summary, we have with probability at least $1 - 1/\sum_{i=1}^D d_i - 12c_\sigma^2\eta^2/\log(\sum_{i=1}^D d_i) - \exp\{-c_2\sum_{i=1}^D d_i\}$,

$$\begin{aligned}
\|\varrho^{(1)}\| \leq & \frac{1}{\prod_{i=1}^D d_i} \|\sum_\omega \mathcal{O}_\omega \mathcal{E}_\omega \mathfrak{I}_\omega / (1/\hat{\tilde{\mathcal{P}}}_\omega - 1/\tilde{\mathcal{P}}_\omega)\| + \frac{1}{\prod_{i=1}^D d_i} \|\sum_\omega \mathcal{O}_\omega \mathcal{E}_\omega \mathfrak{I}_\omega / \tilde{\mathcal{P}}_\omega\| \\
\lesssim & \max\Big\{ \frac{\sqrt{\max\{d_1,...,d_D\}\log(\sum_{i=1}^D d_i)}}{\sqrt{\mathcal{P}_L}\prod_{i=1}^D d_i}, \frac{\log(\sum_{i=1}^D d_i)}{\mathcal{P}_L \prod_{i=1}^D d_i} \Big\} \\
& + r_{2,\max}^{(D-1)/2} \frac{\sqrt{\sum_{i=1}^D d_i}(\log^{1/4}(\sum_{i=1}^D d_i))}{\prod_{i=1}^D d_i}
\end{aligned}$$

□

Lemma 11. *Assume conditions 5,6,7 and 8 hold, denote* $\varrho^{(2)} = \sum_\omega (\hat{\mathcal{X}}_{obs,\omega} - \mathcal{X}_\omega^*)(\mathcal{O}_\omega/\hat{\tilde{\mathcal{P}}}_\omega - \mathcal{O}_\omega/\tilde{\mathcal{P}}_\omega)\mathfrak{I}_\omega/\prod_{i=1}^D d_i$*, then there exists*

$$\Gamma^{(2)} \asymp \frac{\psi_1' \|\hat{\tilde{\mathcal{P}}}^\dagger - \tilde{\mathcal{P}}^\dagger\|_F}{\prod_{i=1}^D d_i}$$

such that $\|\varrho^{(2)}\| \leq \Gamma^{(2)}$ *holds.*

Proof. By $\max_\omega\{|\hat{\mathcal{X}}_{\text{obs},\omega} - \mathcal{X}^*_\omega|\} \leq 2\psi_1'$ and relationship between spectral norm and Frobenious norm, we have

$$\prod_{i=1}^{D} d_i \|\varrho^{(2)}\| \leq 2\psi_1' \|\hat{\tilde{\mathcal{P}}}^\dagger - \tilde{\mathcal{P}}^\dagger\|_F$$

□

Lemma 12. *Consider the underlying signal tensor for propensity score and we use proposed maximum norm low rank constrained maximum likelihood approach to get the estimate* $\hat{\mathcal{S}}$ *as defined in 3.8. Assume 5 holds and we can obtain with probability at least* $1 - \exp\{-c_3 \sum_{i\neq k} d_i\}$

$$\frac{1}{\prod_{i\neq k} d_i}\|\hat{\mathcal{S}} - \mathcal{S}\|_F^2 \leq \frac{c_1 L_{\psi_2}^2 \max_i r_{2,i}^{D-1}}{U_{\psi_2}^2} \frac{\sum_{i\neq k} d_i}{\prod_{i\neq k} d_i}$$

where L_{ψ_2} *and* U_{ψ_2} *are defined as* $L_{\psi_2} = \max_{|x|\leq\psi_2} |f(x)'|/f(x)(1-f(x))$ *and* $U_{\psi_2} = \min_{|x|\leq\psi_2}[(f(x)')^2/f^2(x) - f(x)''/f(x)]$*, controling the "steepness" and "convexity" of the link function* f *and* c_1, C_3 *are constants only dependent on* D*.*

Proof. The proof can be completed by following the proof of theorem 4.1 in Lee and Wang [2020]. □

Lemma 13. *Let* ϵ_ω *be i.i.d Rademacher random variables and denote* $\varrho^{(3)} = \frac{1}{|\mathcal{O}|}\sum_{\omega\in\mathcal{O}} \epsilon_\omega \mathcal{O}_\omega$*. If we assume* $\tilde{\mathcal{P}}_L \geq \mu \prod_{i=1}^D d_i$*, for* $|\mathcal{O}| \geq c \min_i\{d_1, ..., d_D\} \log^3(\sum_{i=1}^D d_i)$ *where* c *is some constant, there exists an absolute constant* C_1 *such that*

$$\mathbb{E}\|\varrho^{(3)}\| \leq C_1 \sqrt{\frac{\max_i\{d_1, ..., d_D\} \log(\sum_{i=1}^D d_i)}{\prod_{i=1}^D d_i |\mathcal{O}|}}$$

Furthermore, for $\mathcal{A} \in \mathcal{C}(\mathbf{r})$ where $\mathcal{C}(\mathbf{r})$ is defined in B.3, we have

$$\frac{1}{\prod_{i=1}^{D} d_i}\|\mathcal{O} * \mathcal{A}\|_F^2 \geq \frac{1}{2}\mathbb{E}(\|\mathcal{O} * \mathcal{A}\|_F^2) - C(\max_i r_i)^{D-1}\mu \prod_{i=1}^{D} d_i \mathbb{E}\|\varrho^{(3)}\|^2$$

with probability at least $1 - 1/\sum_{i=1}^{D} d_i$ for some constant C.

Proof. The proof can be completed by following the proof of Lemma 12 in Klopp [2014]. □

B.3 More simulation experiments

In this section, we analyze how the algorithm 3 and 4 perform when the tensor dimensions and Tucker rank along different modes are imbalanced. Combining with the results in 3.6, we can conclude that our proposed algorithm are stable enough to perform well under different scenarios.

First, we validate the performance of algorithm 3 for propensity score when the underlying parameter tensor $\mathcal{S}$ is imbalanced in terms of dimension and Tucker rank. We follow the same data generation framework as illustrated in 3.6, where $\mathcal{S} = \mathcal{C} \times_1 \mathbf{S}_1 \times_2 \mathbf{S}_2 \times_3 \mathbf{S}_3$ with tucker low rank $\mathbf{r} = (r_1, r_2, r_3)$. Two dimension choices are performed in our experiments, one representing small size tensor with $d_1 = 10, d_2 = 20, d_3 = 30$ and one for large size with $d_1 = 40, d_2 = 50, d_3 = 60$. To investigate how imbalanced tucker rank affects the algorithm performance, we consider low rank $r_1 = 2, r_2 = 5, r_3 = 8$ and high rank $r_1 = 5, r_2 = 8, r_3 = 10$. Similar to the square tensor case, we vary the choice of $\psi \in \{1, 10\}$ to evaluate the consequence of max norm ψ. As we can tell from Figure B.1, the results are almost identical to those in square tensor case.

Analogously, we verify algorithm 4 works when true underlying tensor $\mathcal{X}_{\text{True}}$ is imbalanced in terms of dimension and Tucker rank. Assuming the same Tucker low rank decomposition structure of $\mathcal{X}_{\text{True}}$ as 3.6, experiments with two sample size $d_1 = d_2 = 10, d_3 = d_4 = 20$ and $d_1 = d_2 = 20, d_3 = d_4 = 50$ are performed for varying perturbation level

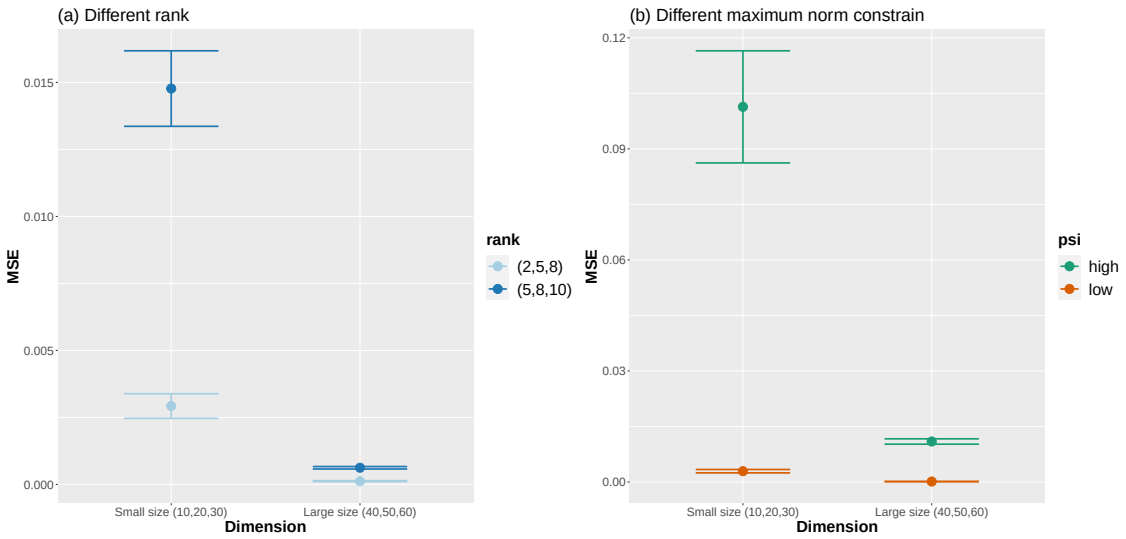


Figure B.1: MSE= $\|\hat{\mathcal{P}} - \mathcal{P}\|_F^2/(d_1 d_2 d_3)$ of propensity score estimate via Algorithm 3 under different choice of dimension, tucker rank and max norm ψ. (a) Performance of experiments with various dimensions and tucker rank. ψ is set to be 1. (b) Performance of experiments with various dimensions and $\psi = \{1(\text{low}), 10(\text{high})\}$.

$\sigma \in \{0, 0.5, 1, 5\}$, rank $(r_1, r_2, r_3, r_4) \in \{(2,2,5,5), (5,5,8,8), (8,8,20,20)\}$ and $\psi \in \{1, 10\}$.

Please refer to Figure B.2 for results summary.

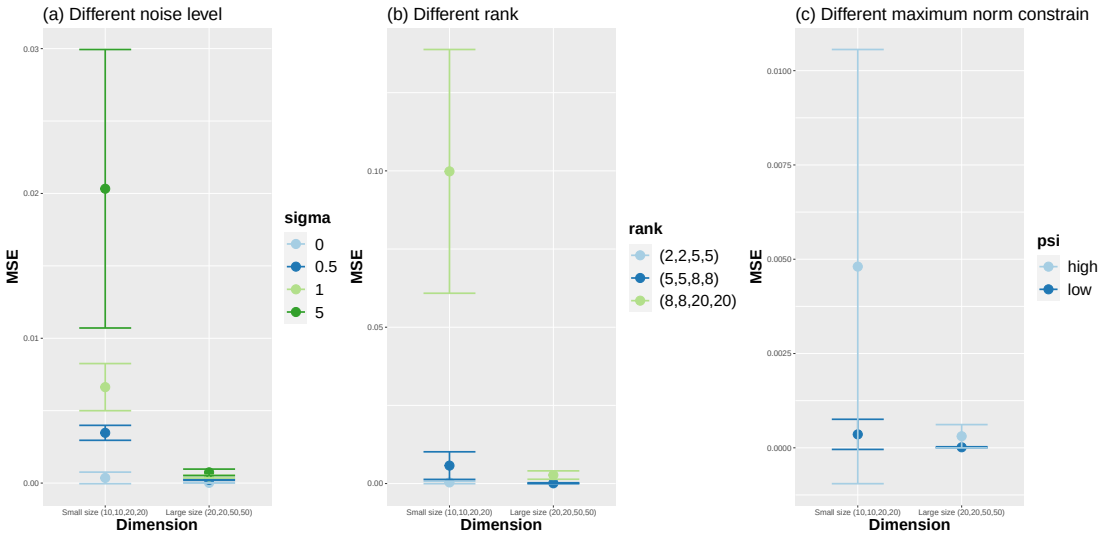


Figure B.2: MSE= $\|\hat{\mathcal{X}}_{\text{Obs}} - \mathcal{X}^*\|_F^2/(d_1d_2d_3d_4)$ of propensity score estimate via Algorithm 3 under different choice of dimension, tucker rank, perturbation level and max norm ψ. (a) Performance of experiments with various dimensions and perturbation level. ψ is set to be 1 and $(r_1, r_2, r_3, r_4) = (2, 2, 5, 5)$. (b) Performance of experiments with various dimensions and tucker rank, $\psi = 1, \sigma = 0$. (c) Performance of experiments with various dimensions and $\psi = \{1(\text{low}), 10(\text{high})\}$, $(r_1, r_2, r_3, r_4) = (2, 2, 5, 5), \sigma = 0$.

APPENDIX C

TECHNICAL DETAILS OF CHAPTER 4

C.1 Proof of Theorem 6

We first introduce more notations utilized in appendix: $\gamma = \max\{\|\mathbf{\Sigma}_X^{-1}\|, \|\mathbf{\Sigma}_Y^{-1}\|, \|\mathbf{\Sigma}_Z^{-1}\|\}$, $d_s = d_1 + d_2 + d_3$, $d_p = d_1 d_2 d_3$, $n_p = n_1 n_2 n_3$.

C.1.1 Proof of exact recovery property

Lemma 14. *If the following assumptions hold, $(\mathcal{G}_0, \mathcal{Q}_0)$ are the unique minimizer to the optimization problem 4.4.*

Assumption 9. *For any non-zero $\mathcal{M} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ such that $P_\Omega(\mathcal{M}) = 0$ and $\mathcal{M} = [\mathcal{M}; P_X, P_Y, P_Z]$, we have with $\xi \leq \sqrt{5/2}$*

$$\|P_{\mathcal{R}}(\mathcal{M})\|_F \leq \xi \|P_{\mathcal{R}^\perp}(\mathcal{M})\|_F$$

Assumption 10. *Suppose $\exists \mathcal{H} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ such that*

$$P_\Omega(\mathcal{H}) = \mathcal{H}, \quad \|P_{\mathcal{R}}(\mathcal{H}) - [\mathcal{I}; \mathbf{U}, \mathbf{V}, \mathbf{W}]\|_F \leq \frac{1}{4}\sqrt{\frac{|\Omega|}{2n_1 n_2 n_3}}, \quad \|P_{\mathcal{R}^\perp}(\mathcal{H})\| \leq 1/2$$

where $\mathcal{I}$ represents a cubical tensor with ones along the superdiagonal.

Proof. Our goal is to prove $(\mathcal{G}_0, \mathcal{Q}_0)$ are the unique optimizer and we can assume there exists another minimizer to the optimization problem 4.4, i.e., $(\mathcal{G}_0 + \Delta\mathcal{G}, \mathcal{Q}_0 + \Delta\mathcal{Q})$ such that $\Delta\mathcal{G}, \Delta\mathcal{Q} \neq 0$. We need to show the contradiction that $\lambda_G\|\mathcal{G}_0 + \Delta\mathcal{G}\|_1 + \lambda_Q\|\mathcal{Q}_0 + \Delta\mathcal{Q}\|_* > \lambda_G\|\mathcal{G}_0\|_1 + \lambda_Q\|\mathcal{Q}_0\|_*$.

Before we state detailed proof procedure, there are some important facts that will be extensively implemented:

1. $P_\Omega([\mathcal{G}_0 + \Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}]) = P_\Omega([\mathcal{G}_0; \mathbf{X}, \mathbf{Y}, \mathbf{Z}])$ which indicates that $P_\Omega([\Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}]) = 0$.

2. $[\mathcal{G}_0 + \Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}] = [\mathcal{G}_0 + \Delta\mathcal{G}; \mathbf{X}P_X, \mathbf{Y}P_Y, \mathbf{Z}P_Z]$ which can be proved by noticing $\mathbf{X}P_X = \mathbf{U}_X\mathbf{\Sigma}_X\mathbf{V}_X^\top\mathbf{U}_X\mathbf{U}_X^\top = \mathbf{X}\mathbf{X}^\top\mathbf{U}_X\mathbf{\Sigma}_X^{-2}\mathbf{U}_X^\top\mathbf{X} = \mathbf{U}_X\mathbf{\Sigma}_X^2\mathbf{U}_X^\top\mathbf{U}_X\mathbf{\Sigma}_X^{-2}\mathbf{U}_X^\top\mathbf{X} = \mathbf{X}$ and similarly, $\mathbf{Y}P_Y = \mathbf{Y}, \mathbf{Z}P_Z = \mathbf{Z}$.

3. Combining the facts that $[\mathcal{G}_0 + \Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}] = \mathcal{Q}_0 + \Delta\mathcal{Q}$ and $[\mathcal{G}_0; \mathbf{X}, \mathbf{Y}, \mathbf{Z}] = \mathcal{Q}_0$, we have $[\Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}] = \Delta\mathcal{Q} \neq 0$. $\|P_{\mathcal{R}}([\Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}])\|_F \leq \xi\|P_{\mathcal{R}^\perp}([\Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}])\|_F \leq \xi\|P_{\mathcal{R}^\perp}([\Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}])\|_*$ by Assumption 9.

We have

$$
\begin{aligned}
&\lambda_G\|\mathcal{G}_0 + \Delta\mathcal{G}\|_1 + \lambda_Q\|\mathcal{Q}_0 + \Delta\mathcal{Q}\|_* \\
=&\lambda_Q\|[\mathcal{G}_0 + \Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}]\|_* + \lambda_G\|\mathcal{G}_0 + \Delta\mathcal{G}\|_1 \\
=&\lambda_Q\|[\mathcal{G}_0 + \Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}]\|_*\|[\mathcal{I}; (\mathbf{U}, \mathbf{u}_\perp), (\mathbf{V}, \mathbf{v}_\perp), (\mathbf{W}, \mathbf{w}_\perp)]\| + \lambda_G\|\mathcal{G}_0 + \Delta\mathcal{G}\|_1 \\
\geq&\lambda_Q\Big\langle[\mathcal{G}_0 + \Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}], [\mathcal{I}; (\mathbf{U}, \mathbf{u}_\perp), (\mathbf{V}, \mathbf{v}_\perp), (\mathbf{W}, \mathbf{w}_\perp)]\Big\rangle + \lambda_G\|\mathcal{G}_0 + \Delta\mathcal{G}\|_1 \\
=&\lambda_Q\Big\langle[\mathcal{G}_0; \mathbf{X}, \mathbf{Y}, \mathbf{Z}], [\mathcal{I}; (\mathbf{U}, \mathbf{u}_\perp), (\mathbf{V}, \mathbf{v}_\perp), (\mathbf{W}, \mathbf{w}_\perp)]\Big\rangle \\
&+ \Big\langle[\Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}], [\mathcal{I}; (\mathbf{U}, \mathbf{u}_\perp), (\mathbf{V}, \mathbf{v}_\perp), (\mathbf{W}, \mathbf{w}_\perp)]\Big\rangle + \lambda_G\|\mathcal{G}_0 + \Delta\mathcal{G}\|_1 \\
=&\lambda_Q\Big\langle[\mathcal{G}_0; \mathbf{X}, \mathbf{Y}, \mathbf{Z}], [\mathcal{I}; (\mathbf{U}, \mathbf{u}_\perp), (\mathbf{V}, \mathbf{v}_\perp), (\mathbf{W}, \mathbf{w}_\perp)]\Big\rangle \\
&+ \Big\langle[\Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}], [\mathcal{I}; (\mathbf{U}, \mathbf{u}_\perp), (\mathbf{V}, \mathbf{v}_\perp), (\mathbf{W}, \mathbf{w}_\perp)]\Big\rangle + \lambda_G\|\mathcal{G}_0 + \Delta\mathcal{G}\|_1 \\
=&\lambda_Q\Big(\|\mathcal{Q}_0\|_* + \Big\langle[\Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}], [\mathcal{I}; (\mathbf{U}, \mathbf{u}_\perp), (\mathbf{V}, \mathbf{v}_\perp), (\mathbf{W}, \mathbf{w}_\perp)] - \mathcal{H}\Big\rangle\Big) + \lambda_G\|\mathcal{G}_0 + \Delta\mathcal{G}\|_1
\end{aligned}
$$

where $\mathbf{u}_\perp, \mathbf{v}_\perp, \mathbf{w}_\perp$ are singular vectors of $P_{\mathcal{R}^\perp}([\Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}])$ such that $\mathbf{U}^\top\mathbf{u}_\perp = 0, \mathbf{V}^\top\mathbf{v}_\perp = 0, \mathbf{W}^\top\mathbf{w}_\perp = 0$ and $(\mathbf{U}, \mathbf{u}_\perp)$ represent a concatenated matrix which binds $\mathbf{U}$ and $\mathbf{u}_\perp$ by columns and similarly for $(\mathbf{V}, \mathbf{v}_\perp), (\mathbf{W}, \mathbf{w}_\perp)$. The second equality comes from the fact that the spectral norm of $[\mathcal{I}; (\mathbf{U}, \mathbf{u}_\perp), (\mathbf{V}, \mathbf{v}_\perp), (\mathbf{W}, \mathbf{w}_\perp)]$ is exactly 1 and the first inequal-

ity is derived based on the matrix norm inequality and note that $\Big\langle [\Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}], \mathcal{H} \Big\rangle = \Big\langle [\Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}], P_\Omega(\mathcal{H}) \Big\rangle = \Big\langle P_\Omega([\Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}]), \mathcal{H} \Big\rangle = 0$ by Assumption 10 and fact 1.

By now, it suffices to show $\lambda_Q \Big\langle [\Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}], [\mathcal{I}; (\mathbf{U}, \mathbf{u}_\perp), (\mathbf{V}, \mathbf{v}_\perp), (\mathbf{W}, \mathbf{w}_\perp)] - \mathcal{H} \Big\rangle + \lambda_G \|\mathcal{G}_0 + \Delta\mathcal{G}\|_1 \geq \lambda_G \|\mathcal{G}_0\|_1$. We can do the following derivations,

$$
\begin{aligned}
&\lambda_Q \Big\langle [\Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}], [\mathcal{I}; (\mathbf{U}, \mathbf{u}_\perp), (\mathbf{V}, \mathbf{v}_\perp), (\mathbf{W}, \mathbf{w}_\perp)] - \mathcal{H} \Big\rangle + \lambda_G \|\mathcal{G}_0 + \Delta\mathcal{G}\|_1 \\
=&\lambda_Q \Big(\Big\langle P_{\mathcal{R}}([\Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}]), [\mathcal{I}; \mathbf{U}, \mathbf{V}, \mathbf{W}] - P_{\mathcal{R}}(\mathcal{H}) \Big\rangle \\
&+ \Big\langle P_{\mathcal{R}}([\Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}]), \mathbf{u}_\perp \circ \mathbf{v}_\perp \circ \mathbf{w}_\perp - P_{\mathcal{R}^\perp}(\mathcal{H}) \Big\rangle \\
&+ \Big\langle P_{\mathcal{R}^\perp}([\Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}]), [\mathcal{I}; \mathbf{U}, \mathbf{V}, \mathbf{W}] - P_{\mathcal{R}}(\mathcal{H}) \Big\rangle \\
&+ \Big\langle P_{\mathcal{R}^\perp}([\Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}]), \mathbf{u}_\perp \circ \mathbf{v}_\perp \circ \mathbf{w}_\perp - P_{\mathcal{R}^\perp}(\mathcal{H}) \Big\rangle \Big) + \lambda_G \|\mathcal{G}_0 + \Delta\mathcal{G}\|_1 \\
=&\lambda_Q \Big(\Big\langle P_{\mathcal{R}}([\Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}]), [\mathcal{I}; \mathbf{U}, \mathbf{V}, \mathbf{W}] - P_{\mathcal{R}}(\mathcal{H}) \Big\rangle \\
&+ \Big\langle P_{\mathcal{R}^\perp}([\Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}]), \mathbf{u}_\perp \circ \mathbf{v}_\perp \circ \mathbf{w}_\perp - P_{\mathcal{R}^\perp}(\mathcal{H}) \Big\rangle \Big) + \lambda_G \|\mathcal{G}_0 + \Delta\mathcal{G}\|_1 \\
\geq&\lambda_Q \Big(\|P_{\mathcal{R}^\perp}([\Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}])\|_* - \|P_{\mathcal{R}}([\Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}])\|_F \|[\mathcal{I}; \mathbf{U}, \mathbf{V}, \mathbf{W}] - P_{\mathcal{R}}(\mathcal{H})\|_F \\
&- \|P_{\mathcal{R}^\perp}(\mathcal{H})\| \|P_{\mathcal{R}^\perp}([\Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}])\|_* \Big) + \lambda_G \|\mathcal{G}_0 + \Delta\mathcal{G}\|_1 \\
\geq&\lambda_Q \Big(\|P_{\mathcal{R}^\perp}([\Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}])\|_* (1 - \frac{1}{2} - \frac{\xi}{4} \sqrt{\frac{|\Omega|}{2n_1 n_2 n_3}}) \Big) + \lambda_G \|\mathcal{G}_0 + \Delta\mathcal{G}\|_1
\end{aligned}
$$

where the second equality comes from the orthogonality between $P_{\mathcal{R}}$ and $P_{\mathcal{R}^\perp}$ and combining matrix norm inequality, Assumption 10 and fact 3, we will obtain the last two inequalities.

Recall the fact that $\xi \leq \sqrt{\frac{5}{2}}$, we can obtain $\frac{\xi}{4} \sqrt{\frac{|\Omega|}{2n_1 n_2 n_3}} \leq 1/2$ holds automatically since after simple reorganization, we have $|\Omega| \leq 32 n_1 n_2 n_3 / 5$ which is a plain truth. Through the above analysis, we are able to derive that under the following 2 scenarios, we can prove $\lambda_Q \Big\langle [\Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}], [\mathcal{I}; (\mathbf{U}, \mathbf{u}_\perp), (\mathbf{V}, \mathbf{v}_\perp), (\mathbf{W}, \mathbf{w}_\perp)] - \mathcal{H} \Big\rangle + \lambda_G \|\mathcal{G}_0 + \Delta\mathcal{G}\|_1 \geq \lambda_G \|\mathcal{G}_0\|_1$.

1. Scenario 1: $\|\mathcal{G}_0 + \Delta\mathcal{G}\|_1 \geq \|\mathcal{G}_0\|_1$

   If $\|\mathcal{G}_0 + \Delta\mathcal{G}\|_1 \geq \|\mathcal{G}_0\|_1$ holds, it's obvious our proof has been finished. Otherwise,

assume $\|\mathcal{G}_0 + \Delta\mathcal{G}\|_1 < \|\mathcal{G}_0\|_1$ and if the minimizer $\mathcal{G}$'s exist in a small ϵ-ball $B_\epsilon(\mathcal{G}_0)$ as a continuous neighbor of $\mathcal{G}_0$ such that $\epsilon < \min_{i,j,k} |\mathcal{G}_{0,(i,j,k)}|$. Then for each $\mathcal{G}_s = \{\mathcal{G}_0 + \Delta\mathcal{G} \in B_\epsilon(\mathcal{G}_0)\}$, $\|\mathcal{G}_s\|_1 \geq \|\mathcal{G}_0\|_1 - d_1 d_2 d_3 \epsilon$. Thus, $\|\mathcal{G}_0 + \Delta\mathcal{G}_0\|_1 \geq \|\mathcal{G}_0\|_1 - d_1 d_2 d_3 \epsilon$. Since ϵ is arbitrarily small, we can obtain $\|\mathcal{G}_0 + \Delta\mathcal{G}\|_1 \geq \|\mathcal{G}_0\|_1$. On the contrary, if the minimizer $\mathcal{G}$'s lies outside this small ϵ-ball $B_\epsilon(\mathcal{G}_0)$, i.e., $\mathcal{G}_0$ is an isolated minimizer, we can show that by choosing appropriate tunning parameters, scenario 2 will hold automatically.

2. Scenario 2: $\lambda_Q\Big(\|P_{\mathcal{R}^\perp}([\Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}])\|_*(1/2 - \frac{\xi}{4}\sqrt{\frac{|\Omega|}{2n_1 n_2 n_3}})\Big) + \lambda_G\|\mathcal{G}_0 + \Delta\mathcal{G}\|_1 \geq \lambda_G\|\mathcal{G}_0\|_1$. Thus, we need to prove $\lambda_Q\Big(\|P_{\mathcal{R}^\perp}([\Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}])\|_*(1/2 - \frac{\xi}{4}\sqrt{\frac{|\Omega|}{2n_1 n_2 n_3}})\Big) + \lambda_G\|\Delta\mathcal{G}\|_1 \geq 0$. Assume there exists a constant $C' > 0$ such that for all $\Delta\mathcal{G}$, $\|P_{\mathcal{R}^\perp}([\Delta\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}])\|_* / \|\Delta\mathcal{G}\|_1 \geq C'$, then by choosing tuning parameters satisfying $\lambda_G/\lambda_Q < C'(1/2 - \frac{\xi}{4}\sqrt{\frac{|\Omega|}{2n_1 n_2 n_3}})$ will be enough.

   If such $C' \geq 0$ does not exist, it implies that there exists infinite minimizers and there exists a sub-sequence $\{\Delta\mathcal{G}_{s_i}\}_{i=1}^\infty$ such that $\lim_{i\to\infty} \|P_{\mathcal{R}^\perp}([\Delta\mathcal{G}_{s_i}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}])\|_F \to 0$ since $\Delta\mathcal{G} \neq 0$, in other words, $\|\Delta\mathcal{G}\|_1 \neq 0$. Using fact 3, we could obtain

$$0 \leq \lim_{i\to\infty} \|P_{\mathcal{R}}([\Delta\mathcal{G}_{s_i}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}])\|_F \leq \lim_{i\to\infty} \|P_{\mathcal{R}^\perp}([\Delta\mathcal{G}_{s_i}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}])\|_F \to 0$$

   Thus, we could derive $\lim_{i\to\infty} \|[\Delta\mathcal{G}_{s_i}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}]\|_F = 0$ which implies $\{\mathcal{G} + \Delta\mathcal{G}_{s_i}\} \subset B_\epsilon(\mathcal{G}_0)$, contradicting the assumption in Scenario 1 that $\mathcal{G}_0$ is an isolated minimizer.

Thus, we have successfully shown that $(\mathcal{G}_0, \mathcal{Q}_0)$ are the unique minimizer to the optimization problem 4.6. □

Next, we will show Assumption 9 & 10 in Lemma 14 hold with high probability when sample size $|\Omega|$ satisfies certain conditions.

C.1.2 Assumption 9 holds with high probability

Lemma 15. *If the sample size satisfies*

$$|\Omega| \geq \max\Big\{\frac{32}{3}\log\frac{2(d_s-2r)r^2}{\psi_1}\gamma\mu_{xyz}\mu_0^2 r^2 d_s, \\ \frac{128}{3}\log\frac{2(d_p-d_s r^2+2r^3)}{\psi_2}(\gamma^2\mu_{xyz}^3 d_p-\gamma\mu_{xyz}r^2 d_s+2r^3)\Big\}$$

Assumption 9 holds with probability at least $1-\psi_1-\psi_2$.

Proof. Before we state the detailed proof, there are two key observations as follows:

1. $P_\Omega P_{\mathcal{R}}(\mathcal{M}) = -P_\Omega P_{\mathcal{R}^\perp}(\mathcal{M})$. This can be derived through the following procedure: $P_\Omega P_{\mathcal{R}}(\mathcal{M}) + P_\Omega P_{\mathcal{R}^\perp}(\mathcal{M}) = P_\Omega(P_{\mathcal{R}}(\mathcal{M}) + P_{\mathcal{R}^\perp}(\mathcal{M})) = P_\Omega([\mathcal{M}; P_X, P_Y, P_Z]) = P_\Omega(\mathcal{M}) = 0$
2. $\Big\langle \mathcal{M}, P_{\mathcal{R}}P_\Omega P_{\mathcal{R}}(\mathcal{M})\Big\rangle = \Big\langle \mathcal{M}, P_{\mathcal{R}^\perp}P_\Omega P_{\mathcal{R}^\perp}(\mathcal{M})\Big\rangle$. This can be proved by using observation 1, which leads to the following $\Big\langle \mathcal{M}, P_{\mathcal{R}}P_\Omega P_{\mathcal{R}}(\mathcal{M})\Big\rangle - \Big\langle \mathcal{M}, P_{\mathcal{R}^\perp}P_\Omega P_{\mathcal{R}^\perp}(\mathcal{M})\Big\rangle = \Big\langle \mathcal{M}, P_{\mathcal{R}}P_\Omega P_{\mathcal{R}}(\mathcal{M})\Big\rangle + \Big\langle \mathcal{M}, P_{\mathcal{R}^\perp}P_\Omega P_{\mathcal{R}}(\mathcal{M})\Big\rangle = \Big\langle \mathcal{M}, (P_{\mathcal{R}}+P_{\mathcal{R}^\perp})P_\Omega P_{\mathcal{R}}(\mathcal{M})\Big\rangle = 0$

Then, by using Lemma 16 & 17 and observation 2, we can show with probability at least $1-\psi_1-\psi_2$, if

$$|\Omega| \geq \max\Big\{\frac{32}{3}\log\frac{2(d_s-2r)r^2}{\psi_1}\gamma\mu_{xyz}\mu_0^2 r^2 d_s, \\ \frac{128}{3}\log\frac{2(d_p-d_s r^2+2r^3)}{\psi_2}(\gamma^2\mu_{xyz}^3 d_p-\gamma\mu_{xyz}r^2 d_s+2r^3)\Big\}$$

then we have

$$\frac{1}{2}\|P_{\mathcal{R}}(\mathcal{M})\|_F^2 \leq \frac{n_1n_2n_3}{|\Omega|}\Big\langle \mathcal{M}, P_{\mathcal{R}}P_\Omega P_{\mathcal{R}}(\mathcal{M})\Big\rangle \\ = \frac{n_1n_2n_3}{|\Omega|}\Big\langle \mathcal{M}, P_{\mathcal{R}^\perp}P_\Omega P_{\mathcal{R}^\perp}(\mathcal{M})\Big\rangle \leq \frac{5}{4}\|P_{\mathcal{R}^\perp}(\mathcal{M})\|_F^2$$

which we can easily derive $\|P_{\mathcal{R}}(\mathcal{M})\|_F \leq \sqrt{\frac{5}{2}}\|P_{\mathcal{R}^\perp}(\mathcal{M})\|_F$. □

Lemma 16. *With probability at least* $1-\psi_1$*, we have*

$$\|P_{\mathcal{R}} - \frac{n_1n_2n_3}{|\Omega|}P_{\mathcal{R}}P_{\Omega}P_{\mathcal{R}}\| \leq \sqrt{\frac{8}{3}\log\frac{2(d_s-2r)r^2}{\psi_1}\frac{\gamma\mu_{xyz}\mu_0^2r^2d_s}{|\Omega|}}$$

Moreover, if

$$|\Omega| \geq \frac{32}{3}\log\frac{2(d_s-2r)r^2}{\psi_1}\gamma\mu_{xyz}\mu_0^2r^2d_s$$

then for any $\mathcal{M} \in \mathbb{R}^{n_1\times n_2\times n_3}$*,*

$$\frac{1}{2}\|P_{\mathcal{R}}(\mathcal{M})\|_F^2 \leq \left\langle \mathcal{M}, \frac{n_1n_2n_3}{|\Omega|}P_{\mathcal{R}}P_{\Omega}P_{\mathcal{R}}(\mathcal{M})\right\rangle$$

Proof. For $\mathcal{M} \in \mathbb{R}^{n_1\times n_2\times n_3}$,

$$\begin{aligned}\frac{n_1n_2n_3}{|\Omega|}P_{\mathcal{R}}P_{\Omega}P_{\mathcal{R}}(\mathcal{M}) &= \frac{n_1n_2n_3}{|\Omega|}\sum_{(i,j,k)\in\Omega}\left\langle P_{\mathcal{R}}(\mathcal{M}), \mathbf{e}_i\circ\mathbf{e}_j\circ\mathbf{e}_k\right\rangle P_{\mathcal{R}}(\mathbf{e}_i\circ\mathbf{e}_j\circ\mathbf{e}_k)\\ &= \frac{n_1n_2n_3}{|\Omega|}\sum_{(i,j,k)\in\Omega}P_{\mathcal{R}}P_{(i,j,k)}P_{\mathcal{R}}(\mathcal{M}) := \frac{n_1n_2n_3}{|\Omega|}\sum_{(i,j,k)\in\Omega}\mathcal{T}_{i,j,k}(\mathcal{M})\end{aligned}$$

Suppose $(i,j,k) \in \Omega$ are uniformly sampled from $[n_1]\times[n_2]\times[n_3]$, we can calculate the expectation for $\mathcal{T}_{i,j,k}(\mathcal{M})$ which is

$$\mathbb{E}(\mathcal{T}_{i,j,k}(\mathcal{M})) = \mathbb{E}(P_{\mathcal{R}}(\mathcal{M})_{i,j,k}\mathbf{e}_i\circ\mathbf{e}_j\circ\mathbf{e}_k) = \mathbb{E}(\sum_{(i,j,k)\in\Omega}P_{\mathcal{R}}(\mathcal{M})_{i,j,k}\mathcal{E}_{i,j,k}) = \frac{1}{n_1n_2n_3}P_{\mathcal{R}}(\mathcal{M})$$

Thus, we can show $P_{\mathcal{R}}(\mathcal{M}) - \frac{n_1n_2n_3}{|\Omega|}P_{\mathcal{R}}P_{\Omega}P_{\mathcal{R}}(\mathcal{M})$ is the sum of a sequence of mean 0,

dependent tensors, i.e.,

$$\mathbb{E}(P_{\mathcal{R}}(\mathcal{M}) - \frac{n_1n_2n_3}{|\Omega|}P_{\mathcal{R}}P_{\Omega}P_{\mathcal{R}}(\mathcal{M})) = \mathbb{E}(P_{\mathcal{R}}(\mathcal{M}) - \frac{n_1n_2n_3}{|\Omega|}\sum_{(i,j,k)\in\Omega}\frac{1}{n_1n_2n_3}P_{\mathcal{R}}(\mathcal{M})) = 0$$

To bound $\|P_{\mathcal{R}} - \frac{n_1n_2n_3}{|\Omega|}P_{\mathcal{R}}P_{\Omega}P_{\mathcal{R}}\|$, it's natural to consider using concentration inequality such as Bernstein inequality. However, since this sequence of mean 0 tensors are dependent instead of i.i.d, martingale type Bernstein inequality [Tropp, 2011] is implemented.

Let (a_l, b_l, c_l) be sequentially uniformly sampled from Ω without replacement, $\mathcal{S}_l = \{(a_j, b_j, c_j) : j \leq l\}$ and $m_k = n_1n_2n_3 - k$. Given $\mathcal{S}_l$, the conditional expectation of $P_{(a_{l+1},b_{l+1},c_{l+1})}$ (with a little bit abuse of notation, $P_{P_{\mathcal{G}}}$ is the mapping giving a tensor whose entries in set $P_{\mathcal{G}}$ as the original corresponding value in tensor and 0 for all other entries) is

$$\mathbb{E}(P_{(a_{l+1},b_{l+1},c_{l+1})}|P_{\mathcal{S}_l}) = \frac{P_{\mathcal{S}_l^c}}{m_l}$$

Then we can define the martingale differences as

$$D_l = n_1n_2n_3\frac{m_{|\Omega|}}{m_l}P_{\mathcal{R}}(P_{(a_l,b_l,c_l)} - \frac{P_{\mathcal{S}_{l-1}^c}}{m_{l-1}})P_{\mathcal{R}}$$

since naturally, it's easy to show $P_{\mathcal{S}_0^c} = \mathbf{I}$ and $S_{|\Omega|} = \Omega$ and furthermore we have

$$\begin{aligned}
P_{\mathcal{R}}P_{\Omega}P_{\mathcal{R}}/m_{|\Omega|} &= \frac{D_{\Omega}}{n_1n_2n_3m_{|\Omega|}} + P_{\mathcal{R}}\frac{P_{\mathcal{S}_{|\Omega|-1}^c}}{m_{|\Omega|-1}}P_{\mathcal{R}}/m_{|\Omega|} + P_{\mathcal{R}}P_{\mathcal{S}_{|\Omega|-1}}P_{\mathcal{R}}/m_{|\Omega|} \\
&= \sum_{l=1}^{|\Omega|}\frac{D_l}{n_1n_2n_3m_{|\Omega|}} + P_{\mathcal{R}}(1/m_{|\Omega|} - 1/m_0) \\
&= \sum_{l=1}^{|\Omega|}\frac{D_l}{n_1n_2n_3m_{|\Omega|}} + \frac{|\Omega|}{n_1n_2n_3m_{|\Omega|}}P_{\mathcal{R}}
\end{aligned}$$

Thus, from the above analysis, we have obtained

$$\frac{n_1n_2n_3}{|\Omega|}P_{\mathcal{R}}P_{\Omega}P_{\mathcal{R}} - P_{\mathcal{R}} = \frac{1}{|\Omega|}\sum_{l=1}^{|\Omega|} D_l$$

then by applying martingale type Bernstein inequality, we have

$$\mathbb{P}(\frac{1}{|\Omega|}\|\sum_{l=1}^{|\Omega|} D_l\| \geq \tau) \leq 2\text{rank}(P_{\mathcal{R}})\exp\Big\{-\frac{|\Omega|^2\tau^2/2}{\sigma^2 + |\Omega|\tau M/3}\Big\} \tag{C.1}$$

M is upper bound of $\|D_k\|$ and σ^2 is upper bound of $\|\sum_{l=1}^{|\Omega|}\mathbb{E}(D_lD_l|\mathcal{S}_{l-1})\|$ and now we should be able to say D_l's are random self-adjoint operators. Then we will bound $\text{rank}(P_{\mathcal{R}}), M, \sigma^2$ separately.

From the formulation of $P_{\mathcal{R}}$, $\text{rank}(P_{\mathcal{R}})$ is

$$\begin{aligned}
\text{rank}(P_{\mathcal{R}}) =& \text{rank}(P_{\mathcal{R}}^0 + P_{\mathcal{R}}^1 + P_{\mathcal{R}}^2 + P_{\mathcal{R}}^3) \\
=& \text{rank}(P_U \otimes P_V \otimes P_W) + \text{rank}(P_U \otimes P_{V^\perp} \otimes P_W) \\
&+ \text{rank}(P_{U^\perp} \otimes P_V \otimes P_W) + \text{rank}(P_U \otimes P_V \otimes P_{W^\perp}) \\
=& d_1r^2 + (d_2 - r)r^2 + (d_3 - r)r^2 = (d_s - 2r)r^2
\end{aligned}$$

By the definition of operator norm, we have

$$\begin{aligned}
\|\sum_{l=1}^{|\Omega|}\mathbb{E}[D_lD_l|\mathcal{S}_{l-1}]\| =& \max_{\|P_{\mathcal{R}}(\mathcal{M})\|_F=1}\sum_{l=1}^{|\Omega|}\mathbb{E}[\langle D_l(\mathcal{M}), D_l(\mathcal{M})\rangle|\mathcal{S}_{l-1}] \\
\leq& \sum_{l=1}^{|\Omega|}\frac{(n_1n_2n_3(m_{|\Omega|}/m_l))^2}{m_{|\Omega|-1}}\max_{\|P_{\mathcal{R}}(\mathcal{M})\|_F=1}\sum_{(i,j,k)\in\Omega}\langle P_{\mathcal{R}}P_{(i,j,k)}P_{\mathcal{R}}(\mathcal{M}), \mathcal{M}\rangle \\
\leq& |\Omega|n_1n_2n_3\max_{i,j,k}\|P_{\mathcal{R}}P_{(i,j,k)}P_{\mathcal{R}}\| \\
\leq& |\Omega|n_1n_2n_3\gamma\mu_{xyz}\mu_0^2r^2\frac{d_1+d_2+d_3}{n_1n_2n_3} = |\Omega|\gamma\mu_{xyz}\mu_0^2r^2d_s
\end{aligned}$$

The second inequality comes from the fact that $m_{|\Omega|} \leq m_l$ and $\sum_{l=1}^{|\Omega|} m_{|\Omega|}/(m_l m_{l-1}) = |\Omega|/n_1 n_2 n_3$. The third inequality is derived from Lemma 18.

Similarly, we can obtain

$$M \leq \max_l n_1 n_2 n_3 (m_{|\Omega|}/m_l) 2 \max_{i,j,k} \|P_{\mathcal{R}} P_{(i,j,k)} P_{\mathcal{R}}\| \leq 2\gamma \mu_{xyz} \mu_0^2 r^2 d_s$$

Thus, the Bernstein inequality in C.1 can be written as

$$\mathbb{P}(\frac{1}{|\Omega|}\|\sum_{l=1}^{|\Omega|} D_l\| \geq \tau) \leq 2(d_s - 2r)r^2 \exp\Big\{ - \frac{\tau^2/2}{1 + 2\tau/3} \times \frac{|\Omega|}{\gamma \mu_{xyz} \mu_0^2 r^2 d_s} \Big\} \tag{C.2}$$

Furthermore, if we try to simplify the Bernstein inequality , we can find as long as $\tau \leq \frac{1}{M}\sigma^2$, inequality C.1 can be reorganized as

$$\mathbb{P}(\frac{1}{|\Omega|}\|\sum_{l=1}^{|\Omega|} D_l\| \geq \tau) \leq 2\text{rank}(P_{\mathcal{R}}) \exp\Big\{ - 3\tau^2/(8\sigma^2) \Big\}$$

So we can get for inequality C.2, if

$$M^2 \log \frac{2(d_s - 2r)r^2}{\psi_1} \leq \frac{3}{8}|\Omega|\sigma^2$$

which is equivalent to put constraint on sample size $|\Omega|$, i.e.,

$$|\Omega| \geq \sqrt{\frac{32}{3} \log \frac{2(d_s - 2r)r^2}{\psi_1} \gamma \mu_{xyz} \mu_0^2 r^2 d_s} \tag{C.3}$$

then with probability at least $1 - \psi_1$,

$$\|P_{\mathcal{R}} - \frac{n_1 n_2 n_3}{|\Omega|} P_{\mathcal{R}} P_\Omega P_{\mathcal{R}}\| \leq \sqrt{\frac{8}{3} \log \frac{2(d_s - 2r)r^2}{\psi_1} \frac{\gamma \mu_{xyz} \mu_0^2 r^2 d_s}{|\Omega|}} \leq \frac{1}{2}$$

where the last inequality holds when $|\Omega| \geq \frac{32}{3}\log\frac{2(d_s-2r)r^2}{\psi_1}\gamma\mu_{xyz}\mu_0^2 r^2 d_s$, resulting in the condition in C.3 satisfied automatically.

The above analysis indicates as an operator in the range of $P_{\mathcal{R}}$, the spectral norm of $\frac{n_1n_2n_3}{|\Omega|}P_{\mathcal{R}}P_{\Omega}P_{\mathcal{R}}$ is contained in $[1/2, 3/2]$ and it's easy to show

$$\frac{n_1n_2n_3}{|\Omega|}\|P_{\mathcal{R}}P_{\Omega}P_{\mathcal{R}}(\mathcal{M})\|_F^2 = \left\langle \mathcal{M}, \frac{n_1n_2n_3}{|\Omega|}P_{\mathcal{R}}P_{\Omega}P_{\mathcal{R}}(\mathcal{M})\right\rangle \geq \frac{1}{2}\|P_{\mathcal{R}}(\mathcal{M})\|_F^2$$

□

Lemma 17. *With probability at least $1-\psi_2$, we have*

$$\|P_{\mathcal{R}^\perp} - \frac{n_1n_2n_3}{|\Omega|}P_{\mathcal{R}^\perp}P_{\Omega}P_{\mathcal{R}^\perp}\| \leq \sqrt{\frac{8}{3}\log\frac{2(d_p-d_sr^2+2r^3)}{\psi_2}\frac{\gamma^2\mu_{xyz}^3 d_p - \gamma\mu_{xyz}r^2d_s + 2r^3}{|\Omega|}}$$

Moreover, if

$$|\Omega| \geq \frac{128}{3}\log\frac{2(d_p-d_sr^2+2r^3)}{\psi_2}(\gamma^2\mu_{xyz}^3 d_p - \gamma\mu_{xyz}r^2d_s + 2r^3)$$

then for any $\mathcal{M} \in \mathbb{R}^{n_1\times n_2\times n_3}$,

$$\left\langle \mathcal{M}, \frac{n_1n_2n_3}{|\Omega|}P_{\mathcal{R}^\perp}P_{\Omega}P_{\mathcal{R}^\perp}(\mathcal{M})\right\rangle \leq \frac{5}{4}\|P_{\mathcal{R}^\perp}(\mathcal{M})\|_F^2$$

Proof. Following the similar proof logic as Lemma 16, we have

$$\begin{aligned}
\text{rank}(P_{\mathcal{R}^\perp}) =& \text{rank}(P^0_{\mathcal{R}^\perp} + P^1_{\mathcal{R}^\perp} + P^2_{\mathcal{R}^\perp} + P^3_{\mathcal{R}^\perp}) \\
=& \text{rank}(P_{U^\perp}\otimes P_{V^\perp}\otimes P_{W^\perp}) + \text{rank}(P_U\otimes P_{V^\perp}\otimes P_{W^\perp}) \\
&+ \text{rank}(P_{U^\perp}\otimes P_V\otimes P_{W^\perp}) + \text{rank}(P_{U^\perp}\otimes P_{V^\perp}\otimes P_W) \\
=& d_1(d_2-r)(d_3-r) + (d_1-r)r(d_3-r) + (d_1-r)(d_2-r)r \\
=& d_p - d_sr^2 + 2r^3
\end{aligned}$$

By utilizing Lemma 18, we can obtain with probability at least $1-\psi_2$,

$$\begin{aligned}\|P_{\mathcal{R}^\perp} - \frac{n_1n_2n_3}{|\Omega|}P_{\mathcal{R}^\perp}P_\Omega P_{\mathcal{R}^\perp}\| &\leq \sqrt{\frac{8}{3}\log\frac{2(d_p-d_sr^2+2r^3)}{\psi_2}\frac{\gamma^2\mu_{xyz}^3d_p-\gamma\mu_{xyz}r^2d_s+2r^3}{|\Omega|}} \\ &\leq \frac{1}{4}\end{aligned}$$

where the last inequality holds when $|\Omega| \geq \frac{128}{3}\log\frac{2(d_p-d_sr^2+2r^3)}{\psi_2}(\gamma^2\mu_{xyz}^3d_p - \gamma\mu_{xyz}r^2d_s + 2r^3)$.

Furthermore, we have

$$\left\langle \mathcal{M}, \frac{n_1n_2n_3}{|\Omega|}P_{\mathcal{R}^\perp}P_\Omega P_{\mathcal{R}^\perp}(\mathcal{M}) - P_{\mathcal{R}^\perp}(\mathcal{M})\right\rangle \leq \frac{1}{4}\|P_{\mathcal{R}^\perp}(\mathcal{M})\|_F^2$$

which we can easily derive that

$$\left\langle \mathcal{M}, \frac{n_1n_2n_3}{|\Omega|}P_{\mathcal{R}^\perp}P_\Omega P_{\mathcal{R}^\perp}(\mathcal{M})\right\rangle \leq \frac{5}{4}\|P_{\mathcal{R}^\perp}(\mathcal{M})\|_F^2$$

□

Lemma 18.

$$\begin{aligned}\max_{i,j,k}\|P_{\mathcal{R}}(\mathbf{e}_i\circ\mathbf{e}_j\circ\mathbf{e}_k)\|^2 &\leq \frac{\gamma\mu_{xyz}\mu_0^2r^2d_s}{n_1n_2n_3} \\ \max_{i,j,k}\|P_{\mathcal{R}^\perp}(\mathbf{e}_i\circ\mathbf{e}_j\circ\mathbf{e}_k)\|^2 &\leq \frac{\gamma^2\mu_{xyz}^3d_p-\gamma\mu_{xyz}r^2d_s+2r^3}{n_1n_2n_3}\end{aligned}$$

Proof. Consider the formulation of $P_{\mathcal{R}}$ and $P_{\mathcal{R}^\perp}$,

$$\begin{aligned}
\|P_{\mathcal{R}}(\mathbf{e}_i \circ \mathbf{e}_j \circ \mathbf{e}_k)\|^2 =& \sum_{m,n=0}^{3} \left\langle P_{\mathcal{R}}^m(\mathbf{e}_i \circ \mathbf{e}_j \circ \mathbf{e}_k), P_{\mathcal{R}}^n(\mathbf{e}_i \circ \mathbf{e}_j \circ \mathbf{e}_k) \right\rangle \\
=& \sum_{m=0}^{3} \|P_{\mathcal{R}}^m(\mathbf{e}_i \circ \mathbf{e}_j \circ \mathbf{e}_k)\|^2 \\
=& \|P_U \mathbf{e}_i\|^2 \|P_V \mathbf{e}_j\|^2 \|P_W \mathbf{e}_k\|^2 + \|P_{U_\perp} \mathbf{e}_i\|^2 \|P_V \mathbf{e}_j\|^2 \|P_W \mathbf{e}_k\|^2 \\
&+ \|P_U \mathbf{e}_i\|^2 \|P_{V_\perp} \mathbf{e}_j\|^2 \|P_W \mathbf{e}_k\|^2 + \|P_U \mathbf{e}_i\|^2 \|P_V \mathbf{e}_j\|^2 \|P_{W_\perp} \mathbf{e}_k\|^2 \\
\leq& \|P_U \mathbf{e}_i\|^2 \|P_V \mathbf{e}_j\|^2 \|P_W \mathbf{e}_k\|^2 + \|P_{U_\perp} \mathbf{e}_i\|^2 \|P_V \mathbf{e}_j\|^2 \|P_W \mathbf{e}_k\|^2 \\
&+ \|P_U \mathbf{e}_i\|^2 \|P_V \mathbf{e}_j\|^2 \|P_W \mathbf{e}_k\|^2 + \|P_U \mathbf{e}_i\|^2 \|P_{V_\perp} \mathbf{e}_j\|^2 \|P_W \mathbf{e}_k\|^2 \\
&+ \|P_U \mathbf{e}_i\|^2 \|P_V \mathbf{e}_j\|^2 \|P_W \mathbf{e}_k\|^2 + \|P_U \mathbf{e}_i\|^2 \|P_V \mathbf{e}_j\|^2 \|P_{W_\perp} \mathbf{e}_k\|^2 \\
\leq& \gamma \mu_{xyz} \mu^2 \frac{d_1 r^2}{n_1 n_2 n_3} + \gamma \mu_{xyz} \mu^2 \frac{d_2 r^2}{n_1 n_2 n_3} + \gamma \mu_{xyz} \mu^2 \frac{d_3 r^2}{n_1 n_2 n_3} \\
=& \gamma \mu_{xyz} \mu^2 \frac{d_s r^2}{n_1 n_2 n_3}
\end{aligned}$$

where the second equality utilizes the orthogonality among $P_{\mathcal{R}}^0, P_{\mathcal{R}}^1, P_{\mathcal{R}}^2$ and $P_{\mathcal{R}}^3$. The second inequality is derived based on

$$\begin{aligned}
\|P_U \mathbf{e}_i\|^2 + \|P_{U^\perp} \mathbf{e}_i\|^2 = \|P_X \mathbf{e}_i\|^2 =& \|\mathbf{U}_X \mathbf{U}_X^\top \mathbf{e}_i\|^2 = \|\mathbf{X} \mathbf{V}_X \mathbf{\Sigma}_X^{-2} V_X^\top \mathbf{X}^\top \mathbf{e}_i\|^2 \\
=& \mathbf{x}_i^\top \mathbf{V}_X \mathbf{\Sigma}_X^{-2} \mathbf{V}_X^\top \mathbf{x}_i = \|\mathbf{\Sigma}_X^{-1} \mathbf{V}_X^\top \mathbf{x}_i\|^2 \leq \gamma \mu_{xyz} \frac{d_1}{n_1}
\end{aligned}$$

Similarly, $\|P_Y \mathbf{e}_j\|^2 \leq \gamma \mu_{xyz} d_2/n_2$ and $\|P_Z \mathbf{e}_k\|^2 \leq \gamma \mu_{xyz} d_3/n_3$. Recall that $\mu_0 \geq 1$, thus we could obtain $\|P_{U_\perp} \mathbf{e}_i\|^2 = \|P_X \mathbf{e}_i\|^2 - \|P_U \mathbf{e}_i\|^2 \leq \sigma \mu_{xyz} d_1/n_1 - r/n_1$ and $\|P_{V_\perp} \mathbf{e}_j\|^2 \leq \sigma \mu_{xyz} d_2/n_2 - r/n_2, \|P_{W_\perp} \mathbf{e}_k\|^2 \leq \sigma \mu_{xyz}(d_1 - r)/n_1$ in a similar way. Following analogical

proof strategy, we have

$$
\begin{aligned}
&\|P_{\mathcal{R}^\perp}(\mathbf{e}_i \circ \mathbf{e}_j \circ \mathbf{e}_k)\|^2 \\
=&\sum_{m,n=0}^{3} \left\langle P^m_{\mathcal{R}^\perp}(\mathbf{e}_i \circ \mathbf{e}_j \circ \mathbf{e}_k), P^n_{\mathcal{R}^\perp}(\mathbf{e}_i \circ \mathbf{e}_j \circ \mathbf{e}_k) \right\rangle \\
=&\sum_{m=0}^{3} \|P^m_{\mathcal{R}^\perp}(\mathbf{e}_i \circ \mathbf{e}_j \circ \mathbf{e}_k)\|^2 \\
=&\|P_{U^\perp}\mathbf{e}_i\|^2\|P_{V^\perp}\mathbf{e}_j\|^2\|P_{W^\perp}\mathbf{e}_k\|^2 + \|P_{U_\perp}\mathbf{e}_i\|^2\|P_V\mathbf{e}_j\|^2\|P_{W^\perp}\mathbf{e}_k\|^2 \\
&+ \|P_U\mathbf{e}_i\|^2\|P_{V_\perp}\mathbf{e}_j\|^2\|P_{W^\perp}\mathbf{e}_k\|^2 + \|P_{U_\perp}\mathbf{e}_i\|^2\|P_{V_\perp}\mathbf{e}_j\|^2\|P_W\mathbf{e}_k\|^2 \\
\leq&(\frac{\gamma\mu_{xyz}d_1 - r}{n_1})(\frac{\gamma\mu_{xyz}d_2 - r}{n_2})(\frac{\gamma\mu_{xyz}d_3 - r}{n_3}) + (\frac{\gamma\mu_{xyz}d_1 - r}{n_1})(\frac{r}{n_2})(\frac{\gamma\mu_{xyz}d_3 - r}{n_3}) \\
&+ (\frac{r}{n_1})(\frac{\gamma\mu_{xyz}d_2 - r}{n_2})(\frac{\gamma\mu_{xyz}d_3 - r}{n_3}) + (\frac{\gamma\mu_{xyz}d_1 - r}{n_1})(\frac{\gamma\mu_{xyz}d_2 - r}{n_2})(\frac{r}{n_3}) \\
\leq&\frac{\gamma^2\mu^3_{xyz}d_p - \gamma\mu_{xyz}r^2d_s + 2r^3}{n_1n_2n_3}
\end{aligned}
$$

□

C.1.3 Assumption 10 holds with high probability

Lemma 19. *Assumption 10 holds with probability at least* $1 - 2q_2(d_s - 2r)r^2 \exp\left\{ - \frac{\tau^2/2}{1+2\tau/3} \frac{q_1}{\gamma\mu_{xyz}\mu_0^2 r^2 d_s} \right\} - 2q_2 n_1 n_2 n_3 \exp\left\{ - \frac{3q_1}{32\gamma\mu_{xyz}\mu_0^2 r^2 d_s} \right\} - q_2(n_1n_2n_3)^{-\beta-1}$*, where* q_1, q_2 *are constants which separate* $|\Omega|$ *into* q_2 *subsequences with each length as* q_1 *and* $q_2 \geq -(1/\log\tau)\log(\sqrt{32}n_1n_2n_3|\Omega|^{-1/2})$.

Proof. To show assumption 2 holds with high probability, we need to borrow 'golfing scheme' from Gross [2011] to construct Ω as uniformly sampled subset from $[n_1] \times [n_2] \times [n_3]$. We try to divide the all the observations in Ω, i.e., $\{(a_i, b_i, c_i)\}$ into q_2 sub-sequences with each

length as q_1 and furthermore, we define

$$\Omega_l = \{(a_i, b_i, c_i) | (l-1)q_1 \leq i \leq lq_1\} \tag{C.4}$$

where $q_1 q_2 \leq |\Omega|$. $[\mathcal{I}; \mathbf{U}, \mathbf{V}, \mathbf{W}]$ enjoys following property $[\mathcal{I}; \mathbf{U}, \mathbf{V}, \mathbf{W}] = P^0_{\mathcal{R}}([\mathcal{I}; \mathbf{U}, \mathbf{V}, \mathbf{W}])$ and $\|[\mathcal{I}; \mathbf{U}, \mathbf{V}, \mathbf{W}]\| = 1$ due to the fact that $\mathbf{U}, \mathbf{V}, \mathbf{W}$ are factor matrices with columnwise unit-1 norm. Set $P_{\mathbf{Q}_l} = P_{\mathcal{I}} - \frac{1}{q_1} \sum_{i=(l-1)q_1+1}^{lq_1} (n_1 n_2 n_3) P_{(a_i, b_i, c_i)}$ where $P_{\mathcal{I}}$ represents the identity operator on tensors and $P_{(a_i, b_i, c_i)}$ represents the operator maintaining the (a_i, b_i, c_i)-th entry and substituting all the other entries as 0. Through the above preparation, we are able to construct $\mathcal{H}$ as follows

$$\mathcal{H}_l = \sum_{s=1}^{l} (P_{\mathcal{I}} - P_{\mathbf{Q}_s}) P_{\mathcal{R}} P_{\mathbf{Q}_{s-1}} P_{\mathcal{R}} ... P_{\mathcal{R}} P_{\mathbf{Q}_1} P_{\mathcal{R}} [\mathcal{I}; \mathbf{U}, \mathbf{V}, \mathbf{W}]$$

and $\mathcal{H}_{q_2} = \mathcal{H}$. Recall that $P_\Omega(P_{\mathcal{I}} - P_{\mathbf{Q}_l}) = P_{\mathcal{I}} - P_{\mathbf{Q}_l}$, the constructed $\mathcal{H}$ has the property $\mathcal{H} = P_\Omega(\mathcal{H})$. Moreover, we have

$$\begin{aligned} P_{\mathcal{R}}(\mathcal{H}) &= \sum_{s=1}^{l} (P_{\mathcal{R}} - P_{\mathcal{R}} P_{\mathbf{Q}_s} P_{\mathcal{R}}) P_{\mathcal{R}} P_{\mathbf{Q}_{s-1}} P_{\mathcal{R}} ... P_{\mathcal{R}} P_{\mathbf{Q}_1} P_{\mathcal{R}} [\mathcal{I}; \mathbf{U}, \mathbf{V}, \mathbf{W}] \\ &= [\mathcal{I}; \mathbf{U}, \mathbf{V}, \mathbf{W}] - (P_{\mathcal{R}} P_{\mathbf{Q}_l} P_{\mathcal{R}}) ... P_{\mathcal{R}} P_{\mathbf{Q}_1} P_{\mathcal{R}} [\mathcal{I}; \mathbf{U}, \mathbf{V}, \mathbf{W}] \end{aligned}$$

Thus, to show $\|P_{\mathcal{R}}(\mathcal{H}) - [\mathcal{I}; \mathbf{U}, \mathbf{V}, \mathbf{W}]\|_F \leq \frac{1}{4}\sqrt{\frac{|\Omega|}{2n_1 n_2 n_3}}$, it suffices to prove

$$\frac{1}{\sqrt{n_1 n_2 n_3}} \|(P_{\mathcal{R}} P_{\mathbf{Q}_{q_2}} P_{\mathcal{R}}) ... P_{\mathcal{R}} P_{\mathbf{Q}_1} P_{\mathcal{R}} [\mathcal{I}; \mathbf{U}, \mathbf{V}, \mathbf{W}]\|_F \leq \frac{1}{n_1 n_2 n_3} \sqrt{\frac{|\Omega|}{32}}$$

which is equivalent to

$$\max_{1 \leq l \leq q_2} \|P_{\mathcal{R}} P_{\mathbf{Q}_l} P_{\mathcal{R}}\| \leq \tau \text{ and } \tau^{q_2} \leq \frac{1}{n_1 n_2 n_3} \sqrt{\frac{|\Omega|}{32}}$$

due to $1/\sqrt{n_1n_2n_3}\|[\mathcal{I};\mathbf{U},\mathbf{V},\mathbf{W}]\|_F \leq \|[\mathcal{I};\mathbf{U},\mathbf{V},\mathbf{W}]\|\infty \leq \|[\mathcal{I};\mathbf{U},\mathbf{V},\mathbf{W}]\| \leq 1$. Thus, as long as when we split Ω into q_2 subsequences with $q_2 \geq -(1/\log\tau)\log(\sqrt{32}n_1n_2n_3|\Omega|^{-1/2})$, the remaining thing to show $\|P_{\mathcal{R}}P_{\mathbf{Q}_l}P_{\mathcal{R}}\|$ is uniformly lower bounded by τ with high probability. By utilizing the result from Lemma 20 and uniform bound property, we have

$$\begin{aligned}
&\mathbb{P}\Big(\|P_{\mathcal{R}}(\mathcal{H}) - [\mathcal{I};\mathbf{U},\mathbf{V},\mathbf{W}]\|_F \leq \frac{1}{4}\sqrt{\frac{|\Omega|}{2n_1n_2n_3}}\Big)\\
\geq& 1 - \mathbb{P}(\max_{1\leq l\leq q_2}\|P_{\mathcal{R}}P_{\mathbf{Q}_l}P_{\mathcal{R}}\| \geq \tau)\\
\geq& 1 - 2q_2(d_s-2r)r^2\exp\Big\{-\frac{\tau^2/2}{1+2\tau/3}\frac{q_1}{\gamma\mu_{xyz}\mu_0^2r^2d_s}\Big\}
\end{aligned}$$

Next, we consider to prove $\|P_{\mathcal{R}^\perp}(\mathcal{H})\| \leq \frac{1}{2}$ which is equivalent to show the inequality $\|\sum_{s=1}^{q_2}P_{\mathbf{Q}_l}(P_{\mathcal{R}}P_{\mathbf{Q}_{l-1}}P_{\mathcal{R}})$
$...(P_{\mathcal{R}}P_{\mathbf{Q}_1}P_{\mathcal{R}})[\mathcal{I};\mathbf{U},\mathbf{V},\mathbf{W}]\| \leq 1/2$ holds.

Let's define a sequence of tensors $[\mathcal{I};\mathbf{U},\mathbf{V},\mathbf{W}]_s = P_{\mathcal{R}}P_{\mathbf{Q}_s}P_{\mathcal{R}}[\mathcal{I};\mathbf{U},\mathbf{V},\mathbf{W}]_{s-1}, \forall s \geq 1$ and $[\mathcal{I};\mathbf{U},\mathbf{V},\mathbf{W}]_0 = [\mathcal{I};\mathbf{U},\mathbf{V},\mathbf{W}]$. As follows, we are able to derive

$$\begin{aligned}
&\mathbb{P}(\|\sum_{s=1}^{q_2}P_{\mathbf{Q}_s}([\mathcal{I};\mathbf{U},\mathbf{V},\mathbf{W}]_{s-1})\| \geq 1/2)\\
\leq&\mathbb{P}(\|P_{\mathbf{Q}_1}([\mathcal{I};\mathbf{U},\mathbf{V},\mathbf{W}]_0)\| \geq 1/4) + \mathbb{P}(\|[\mathcal{I};\mathbf{U},\mathbf{V},\mathbf{W}]_1\|_\infty \geq \|[\mathcal{I};\mathbf{U},\mathbf{V},\mathbf{W}]_0\|_\infty/2)\\
&+\mathbb{P}(\|\sum_{s=2}^{q_2}P_{\mathbf{Q}_s}([\mathcal{I};\mathbf{U},\mathbf{V},\mathbf{W}]_{s-1})\| \geq 1/4, \|[\mathcal{I};\mathbf{U},\mathbf{V},\mathbf{W}]_1\|_\infty < \|[\mathcal{I};\mathbf{U},\mathbf{V},\mathbf{W}]_0\|_\infty/2)\\
&...\\
\leq&\sum_{s=1}^{q_2-1}\mathbb{P}(\|P_{\mathcal{R}}P_{\mathbf{Q}_s}P_{\mathcal{R}}([\mathcal{I};\mathbf{U},\mathbf{V},\mathbf{W}]_{s-1})\|_\infty \geq \|[\mathcal{I};\mathbf{U},\mathbf{V},\mathbf{W}]\|_\infty/2^s, \|[\mathcal{I};\mathbf{U},\mathbf{V},\mathbf{W}]_{s-1}\|_\infty <\\
&\|[\mathcal{I};\mathbf{U},\mathbf{V},\mathbf{W}]_0\|_\infty/2^{s-1}) + \sum_{s-1}^{q_2}\mathbb{P}(\|P_{\mathbf{Q}_s}([\mathcal{I};\mathbf{U},\mathbf{V},\mathbf{W}]_{s-1})\| \geq 2^{-1-s}, \|[\mathcal{I};\mathbf{U},\mathbf{V},\mathbf{W}]_{s-1}\|_\infty\\
&<\|[\mathcal{I};\mathbf{U},\mathbf{V},\mathbf{W}]_0\|_\infty/2^{s-1})
\end{aligned}$$

$$
\begin{aligned}
&\leq q_2 \max_{\mathcal{X}=P_{\mathcal{R}}(\mathcal{X}),\|\mathcal{X}\|_\infty\leq 1} \Big[\mathbb{P}\Big(P_{\mathcal{R}}P_{\mathbf{Q}_1}P_{\mathcal{R}}(\mathcal{X}) \geq 1/2\Big) + q_2\mathbb{P}\Big(\|P_{\mathbf{Q}_1}(\mathcal{X})\| \geq 1/(4\|[\mathcal{I};\mathbf{U},\mathbf{V},\mathbf{W}]\|_\infty)\Big)\Big] \\
&\leq 2q_2 n_1n_2n_3 \exp\Big\{-\frac{3q_1}{32\gamma\mu_{xyz}\mu_0^2 r^2 d_s}\Big\} + q_2(n_1n_2n_3)^{-\beta-1}
\end{aligned}
$$

where the second to last inequality set $\mathcal{X} = [\mathcal{I};\mathbf{U},\mathbf{V},\mathbf{W}]_{s-1}/\|[\mathcal{I};\mathbf{U},\mathbf{V},\mathbf{W}]_{s-1}\|_\infty$ and the last inequality is based on Lemma 21 by setting $\tau = 1/2$ and concentration inequality for sparse tensors in [Yuan and Zhang, 2016]. □

Lemma 20.

$$
\mathbb{P}\Big(\|P_{\mathcal{R}}P_{\mathbf{Q}_l}P_{\mathcal{R}}\| \geq \tau\Big) \leq 2(d_s - 2r)r^2 \exp\Big\{-\frac{\tau^2/2}{1+2\tau/3}\frac{q_1}{\gamma\mu_{xyz}\mu_0^2 r^2 d_s}\Big\}
$$

Proof. Different from Lemma 16, this section is interested in a sequence of i.i.d tensors. We will use similar proof strategy with matrix Bernstein inequality. Define

$$
D_i = n_1n_2n_3 P_{\mathcal{R}}P_{(a_i,b_i,c_i)} - P_{\mathcal{R}}
$$

Since $\|P_{\mathcal{R}}P_{(i,j,k)}P_{\mathcal{R}}\| \leq \gamma\mu_{xyz}\mu_0^2 r^2 d_s/(n_1n_2n_3)$, it's easy to show $\|D_i\| \leq 2\gamma\mu_{xyz}\mu_0^2 r^2 d_s$. Then the total variation is bounded by

$$
\begin{aligned}
\max_{\|\mathcal{X}\|_F=1} \mathbb{E}(\langle D_i(\mathcal{X}), D_i(\mathcal{X})\rangle) &\leq (n_1n_2n_3)^2 \max_{\|\mathcal{X}\|_F=1} \mathbb{E}\langle (P_{\mathcal{R}}P_{(a_i,b_i,c_i)}P_{\mathcal{R}})^2\mathcal{X},\mathcal{X}\rangle \\
&\leq n_1n_2n_3 \max_{i,j,k} \|P_{\mathcal{R}}(\mathbf{e}_i\circ\mathbf{e}_j\circ\mathbf{e}_k)\|_F^2 \\
&\leq \gamma\mu_{xyz}\mu_0^2 r^2 d_s
\end{aligned}
$$

Thus, implementing Bernstein inequality, we have

$$\begin{aligned}\mathbb{P}(\|\frac{1}{q_1}\sum_{i=1}^{q_1} D_i\| \geq \tau) \leq & 2\text{rank}(P_{\mathcal{R}})\exp\Big\{\frac{\tau^2/2}{\gamma\mu_{xyz}\mu_0^2 r^2 d_s + 2\tau\gamma\mu_{xyz}\mu_0^2 r^2 d_s/3}\Big\} \\ = & 2(d_s - 2r)r^2\exp\Big\{\frac{\tau^2/2}{1+2\tau/3}\frac{q_1}{\gamma\mu_{xyz}\mu_0^2 r^2 d_s}\Big\}\end{aligned}$$

□

Lemma 21.

$$\max_{\|\mathcal{X}\|_\infty \leq 1} \mathbb{P}\Big(\|P_{\mathcal{R}}P_{\mathbf{Q}_s}P_{\mathcal{R}}(\mathcal{X})\|_\infty \geq \tau\Big) \leq 2n_1n_2n_3\exp\Big\{-\frac{\tau^2/2}{1+2\tau/3}\frac{q_1}{\gamma\mu_{xyz}\mu_0^2 r^2 d_s}\Big\}$$

Proof. The proof for this Lemma is quite similar to proof of Lemma 20 and we will borrow the notations D from it. For each $(i,j,k) \in [n_1]\times[n_2]\times[n_3]$, we should derive the following fact,

$$\begin{aligned}& \max_{\|\mathcal{X}\|_\infty\leq 1}\frac{1}{n_1n_2n_3}|\langle D_l(\mathcal{X}), \mathbf{e}_i\circ\mathbf{e}_j\circ\mathbf{e}_k\rangle| \\ = & \max_{\|\mathcal{X}\|_\infty\leq 1}\frac{1}{n_1n_2n_3}|\langle P_{\mathcal{R}}(\mathbf{e}_i\circ\mathbf{e}_j\circ\mathbf{e}_k), P_{\mathcal{R}}(\mathbf{e}_{a_l}\circ\mathbf{e}_{b_l}\circ\mathbf{e}_{c_l})\rangle\mathcal{X}_{a_l,b_l,c_l} \\ & - \langle P_{\mathcal{R}}(\mathbf{e}_i\circ\mathbf{e}_j\circ\mathbf{e}_k), P_{\mathcal{R}}(\mathcal{X})\rangle| \\ = & 2\max_{i,j,k}\|P_{\mathcal{R}}(\mathbf{e}_i\circ\mathbf{e}_j\circ\mathbf{e}_k)\|^2\|\mathcal{X}\|_\infty \\ \leq & 2\frac{\gamma\mu_{xyz}\mu_0^2 r^2 d_s}{n_1n_2n_3}\end{aligned}$$

where the last inequality is derived based on Lemma 18. Similarly, we can derive the bound for its total variation as

$$\mathbb{E}(\frac{1}{n_1n_2n_3}|\langle \mathbf{e}_i\circ\mathbf{e}_j\circ\mathbf{e}_k, P_{\mathcal{R}}D_l(\mathcal{X})\rangle|)^2 \leq \frac{1}{n_1n_2n_3}\|P_{\mathcal{R}}(\mathbf{e}_i\circ\mathbf{e}_j\circ\mathbf{e}_k)\|_F^2 \leq \frac{\gamma\mu_{xyz}\mu_0^2 r^2 d_s}{(n_1n_2n_3)^2}$$

It's simple to show $\langle \mathbf{e}_i \circ \mathbf{e}_j \circ \mathbf{e}_k, P_{\mathcal{R}} D_l(\mathcal{X}) \rangle$ are i.i.d mean 0 simple random variables and by implementing Bernstein inequality, we can obtain

$$\mathbb{P}(\frac{1}{q_1}\sum_{l=1}^{q_1}\langle \mathbf{e}_i \circ \mathbf{e}_j \circ \mathbf{e}_k, P_{\mathcal{R}} D_l(\mathcal{X}) \rangle \geq \tau) \leq 2\exp\Big\{ -\frac{\tau^2/2}{1+2\tau/3}\frac{q_1}{\gamma \mu_{xyz}\mu_0^2 r^2 d_s}\Big\}$$

Then, it completes the proof by implementing the union bound.

$$\max_{\|\mathcal{X}\|_\infty \leq 1} \mathbb{P}\Big(\|P_{\mathcal{R}} P_{\mathbf{Q}_s} P_{\mathcal{R}}(\mathcal{X})\|_\infty \geq \tau \Big) \leq 2n_1 n_2 n_3 \exp\Big\{ -\frac{\tau^2/2}{1+2\tau/3}\frac{q_1}{\gamma \mu_{xyz}\mu_0^2 r^2 d_s}\Big\}$$

□

C.2 Proof of Theorem 7

Proof. If the perfect feature matrices $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ are corrupted by some noise $\Delta\mathbf{X}, \Delta\mathbf{Y}, \Delta\mathbf{Z}$ which are bounded by $\|\Delta\mathbf{X}\|_F \leq \Delta_x, \|\Delta\mathbf{Y}\|_F \leq \Delta_y, \|\Delta\mathbf{Z}\|_F \leq \Delta_z$, the optimization problem changes to

$$\begin{aligned} \min_{\mathcal{Q},\mathcal{G}} \ & \|P_\Omega([\mathcal{G}; \mathbf{X}+\Delta\mathbf{X}, \mathbf{Y}+\Delta\mathbf{Y}, \mathbf{Z}+\Delta\mathbf{Z}] - \mathcal{R})\|_F^2 \\ \text{s.t.} \ & \mathcal{Q} - [\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}] \in B_\phi(0) \\ & \|\mathcal{G}\|_1 \leq \alpha_1, \|\mathcal{Q}\|_* \leq \alpha_2 \end{aligned} \tag{C.5}$$

where $B_\phi(0)$ is a small ball with radius ϵ at center 0.

Following the classical proof procedure, we will use Rademacher complexity, which is a commonly used learning theoretic tool to measure the complexity of a function class, to derive the recovery property of our model under corrupted feature settings. Thus, by Lemma 22, we can see that to bound $\text{Risk}_\ell(f)$, we need to bound the model empirical ℓ risk $\widehat{\text{Risk}}_\ell(f)$ and model complexity measured by Rademacher complexity of the feasible function class

F_Θ at the same time. Thus, in the following proof, we will show how to bound the model complexity by the feature settings.

By using the Rademacher contraction principle, we can derive

$$\begin{aligned}
&\mathbb{E}_\sigma(\mathfrak{R}(F_\Theta)) = \mathbb{E}_\sigma[\sup_{f\in F_\Theta}\frac{1}{|\Omega|}\sum_{\alpha=1}^{|\Omega|}\sigma_\alpha \ell(f_\theta(i_\alpha,j_\alpha,k_\alpha),\mathcal{R}_{i_\alpha,j_\alpha,k_\alpha})] \\
\leq&\frac{L_\ell}{|\Omega|}\mathbb{E}_\sigma\Big[\sup_{\|\mathcal{G}\|_1\leq\alpha_1}\sum_{\alpha=1}^{|\Omega|}\sigma_\alpha[\mathcal{G};\mathbf{x}_{i_\alpha},\mathbf{y}_{j_\alpha},\mathbf{z}_{k_\alpha}]+\sup_{\|\mathcal{G}\|_1\leq\alpha_1}\sum_{\alpha=1}^{|\Omega|}\sigma_\alpha[\mathcal{G};\mathbf{x}_{i_\alpha},\mathbf{y}_{j_\alpha},\Delta\mathbf{z}_{k_\alpha}] \\
&+\sup_{\|\mathcal{G}\|_1\leq\alpha_1}\sum_{\alpha=1}^{|\Omega|}\sigma_\alpha[\mathcal{G};\mathbf{x}_{i_\alpha},\Delta\mathbf{y}_{j_\alpha},\mathbf{z}_{k_\alpha}]+\sup_{\|\mathcal{G}\|_1\leq\alpha_1}\sum_{\alpha=1}^{|\Omega|}\sigma_\alpha[\mathcal{G};\mathbf{x}_{i_\alpha},\Delta\mathbf{y}_{j_\alpha},\Delta\mathbf{z}_{k_\alpha}] \\
&+\sup_{\|\mathcal{G}\|_1\leq\alpha_1}\sum_{\alpha=1}^{|\Omega|}\sigma_\alpha[\mathcal{G};\Delta\mathbf{x}_{i_\alpha},\mathbf{y}_{j_\alpha},\mathbf{z}_{k_\alpha}]+\sup_{\|\mathcal{G}\|_1\leq\alpha_1}\sum_{\alpha=1}^{|\Omega|}\sigma_\alpha[\mathcal{G};\Delta\mathbf{x}_{i_\alpha},\mathbf{y}_{j_\alpha},\Delta\mathbf{z}_{k_\alpha}] \\
&+\sup_{\|\mathcal{G}\|_1\leq\alpha_1}\sum_{\alpha=1}^{|\Omega|}\sigma_\alpha[\mathcal{G};\Delta\mathbf{x}_{i_\alpha},\Delta\mathbf{y}_{j_\alpha},\mathbf{z}_{k_\alpha}]+\sup_{\|\mathcal{G}\|_1\leq\alpha_1}\sum_{\alpha=1}^{|\Omega|}\sigma_\alpha[\mathcal{G};\Delta\mathbf{x}_{i_\alpha},\Delta\mathbf{y}_{j_\alpha},\Delta\mathbf{z}_{k_\alpha}]\Big]
\end{aligned}$$

Denote $C_x = \max_i \|\mathbf{x}_i\|_2, C_y = \max_j \|\mathbf{y}_j\|_2$ and $C_z = \max_k \|\mathbf{z}_k\|_2$, $\tilde{\Delta} = C_xC_y\Delta_z + C_xC_z\Delta_y + C_x\Delta_y\Delta_z + \Delta_xC_yC_z + \Delta_x\Delta_zC_y + C_z\Delta_x\Delta_y + \Delta_x\Delta_y\Delta_z$. Since there exists a fixed C_0 such that $\|\mathcal{G}\|_* \leq C_0\|\mathcal{G}\|_1$, we bound the last seven terms in the above inequality with probability at least $1-\exp(-(d_1d_2d_3\log^2(d_{\max}))/2)$as below:

$$\begin{aligned}
&d_1d_2d_3\log dL_\ell\alpha_1C_0/|\Omega|\Big(\max_{i,j,k}\|[\mathcal{I};\mathbf{x}_{i_\alpha},\mathbf{y}_{j_\alpha},\Delta\mathbf{z}_{k_\alpha}]\|+\max_{i,j,k}\|[\mathcal{I};\mathbf{x}_{i_\alpha},\Delta\mathbf{y}_{j_\alpha},\mathbf{z}_{k_\alpha}]\| \\
&+\max_{i,j,k}\|[\mathcal{I};\mathbf{x}_{i_\alpha},\Delta\mathbf{y}_{j_\alpha},\Delta\mathbf{z}_{k_\alpha}]\|+\max_{i,j,k}\|[\mathcal{I};\Delta\mathbf{x}_{i_\alpha},\mathbf{y}_{j_\alpha},\mathbf{z}_{k_\alpha}]\|+\max_{i,j,k}\|[\mathcal{I};\Delta\mathbf{x}_{i_\alpha},\mathbf{y}_{j_\alpha},\Delta\mathbf{z}_{k_\alpha}] \\
&+\max_{i,j,k}\|[\mathcal{I};\Delta\mathbf{x}_{i_\alpha},\Delta\mathbf{y}_{j_\alpha},\mathbf{z}_{k_\alpha}]+\max_{i,j,k}\|[\mathcal{I};\Delta\mathbf{x}_{i_\alpha},\Delta\mathbf{y}_{j_\alpha},\Delta\mathbf{z}_{k_\alpha}]\Big)\leq d_1d_2d_3\log d_{\max}L_\ell\alpha_1C_0\tilde{\Delta}/|\Omega|.
\end{aligned}$$

Next we bound the term $\frac{L_\ell}{|\Omega|}\mathbb{E}_\sigma\Big[\sup_{\|\mathcal{G}\|_1\leq\alpha_1}\sum_{\alpha=1}^{|\Omega|}\sigma_\alpha[\mathcal{G};\mathbf{x}_{i_\alpha},\mathbf{y}_{j_\alpha},\mathbf{z}_{k_\alpha}]\Big]$ following two scenarios:

- When $\alpha_1C_xC_yC_z \leq (\alpha_2+\phi)$, applying Lemma 23, we have with probability at least $1-\gamma_2$, $\frac{L_\ell}{|\Omega|}\mathbb{E}_\sigma\Big[\sup_{\|\mathcal{G}\|_1\leq\alpha_1}\sum_{\alpha=1}^{|\Omega|}\sigma_\alpha[\mathcal{G};\mathbf{x}_{i_\alpha},\mathbf{y}_{j_\alpha},\mathbf{z}_{k_\alpha}]\Big] \leq C_0L_\ell\alpha_1C_xC_yC_z\sqrt{\frac{\log(1/\gamma_2)}{|\Omega|}}$.

- When $\alpha_1 C_x C_y C_z \geq (\alpha_2 + \phi)$, by implementing the constrain $\mathcal{Q} - [\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}] \in B_\phi(0)$ and the fact that tensor nuclear norm is the dual norm of its spectral norm, we can obtain with probability at least $1 - \gamma_2$,

$$
\begin{aligned}
&\frac{L_\ell}{|\Omega|}\mathbb{E}_\sigma\Big[\sup_{\|\mathcal{Q}\|_* \leq \alpha_2} \sum_{\alpha=1}^{|\Omega|} \sigma_\alpha[\mathcal{Q}; \mathbf{e}_{i_\alpha}, \mathbf{e}_{j_\alpha}, \mathbf{e}_{k_\alpha}] + \sup_{\|\mathcal{Q}-[\mathcal{G};\mathbf{X},\mathbf{Y},\mathbf{Z}]\|_F \leq \phi} \sum_{\alpha=1}^{|\Omega|} \sigma_\alpha[\mathcal{Q} - [\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}]; \mathbf{e}_{i_\alpha}, \mathbf{e}_{j_\alpha}, \mathbf{e}_{k_\alpha}]\Big] \\
\leq & L_\ell \max_{i,j,k} \|[\mathcal{I}; \mathbf{e}_{i_\alpha}, \mathbf{e}_{j_\alpha}, \mathbf{e}_{k_\alpha}]\| \|\mathcal{Q}\|_* \sqrt{\frac{\log(1/\gamma_2)}{|\Omega|}} \\
& + L_\ell \max_{i,j,k} \|[\mathcal{I}; \mathbf{e}_{i_\alpha}, \mathbf{e}_{j_\alpha}, \mathbf{e}_{k_\alpha}]\|_F \|\mathcal{Q} - [\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}]\|_F \sqrt{\frac{\log(1/\gamma_2)}{|\Omega|}} \\
\leq & L_\ell(\alpha_2 + \phi)\sqrt{\frac{\log(1/\gamma_2)}{|\Omega|}}
\end{aligned}
$$

Recall the definition of μ_{xyz}, we have $C_x C_y C_z \leq \mu_{xyz}^{3/2}$. Then, combining the above results with Lemma 22 and let $\gamma_2 = 1/n_{\max}$, we have

$$
\mathrm{Risk}_\ell(f) \lesssim L_\ell\Big(\max\{(\alpha_2 + \phi), \alpha_1 \mu_{xyz}^{3/2}\}\sqrt{\frac{\log n_{\max}}{|\Omega|}} + \frac{\alpha_1 d_1 d_2 d_3 \log d_{\max} \tilde{\Delta}}{|\Omega|}\Big) + \mathcal{B}\sqrt{\frac{\log(1/\psi_3)}{2|\Omega|}}
$$

Thus, Theorem 7 can be attained with high probability directly by setting $\mathrm{Risk}_\ell(f) \leq \varepsilon$. □

Lemma 22. *Let ℓ be a loss function with Lipschitz constant L_ℓ bounded by $\mathcal{B}$ with respect to its first argument and ψ_3 be a constant in $(0, 1)$. Let $\mathfrak{R}(F_\Theta)$ be the Rademacher complexity of the function class F_Θ (with respect to Ω and loss function ℓ) defined as*

$$
\mathfrak{R}(F_\Theta) = \mathbb{E}_\sigma[\sup_{f \in F_\Theta} \frac{1}{|\Omega|} \sum_{\alpha=1}^{|\Omega|} \sigma_\alpha \ell(f_\theta(i, j, k), \mathcal{R}_{i,j,k})]
$$

where σ_α takes the value $\{\pm 1\}$ with probability $1/2$ respectively. Then with probability $1 - \psi_3$, for all $f \in F_\Theta$ we have

$$
Risk_\ell(f) \leq \widehat{Risk}_\ell(f) + 2\mathbb{E}_\Omega(\mathfrak{R}(F_\Theta)) + \mathcal{B}\sqrt{\frac{\log(1/\psi_3)}{2|\Omega|}}
$$

Lemma 23. *Let* $S_w = \{\mathcal{E} \in \mathbb{R}^{n_1 \times n_2 \times n_3} | \|\mathcal{E}\|_* \leq w\}$, $a = \max_i \|\mathcal{A}_i\|$ *where* $\mathcal{A}_i = \mathbf{x}_i \circ \mathbf{y}_i \circ \mathbf{z}_i \in \mathbb{R}^{n_1 \times n_2 \times n_3}, \forall i \in \{1, 2, ..., |\Omega|\}$, *we have with probability at least* $1 - \gamma$,

$$\mathbb{E}_\sigma[\sup_{\mathcal{E} \in S_w} \frac{1}{|\Omega|} \sum_{\alpha=1}^{|\Omega|} \langle \mathcal{E}, \mathcal{A}_i \rangle] \lesssim O(aw\sqrt{\frac{\log(1/\gamma)}{|\Omega|}})$$

This Lemma can be proved following Lemma 2.8 in Barak and Moitra [2016] and we omit its proof here.

C.3 Proof of Theorem 8

Proof. Before we state the analysis for the convergence property of nested double ADMM algorithm, we do some organization on the augmented Lagrangian 4.8. By setting

$$P_{\mathcal{G}}(\mathcal{G}) = \begin{bmatrix} -[\mathcal{G}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}] \\ 0 \end{bmatrix} \quad P_{\mathcal{Q}}(\mathcal{Q}) = \begin{bmatrix} \mathcal{Q} \\ P_\Omega(\mathcal{Q}) \end{bmatrix} \quad P_{\mathcal{F}}(\mathcal{F}) = \begin{bmatrix} \mathcal{F} \\ 0 \end{bmatrix}$$

$$\mathcal{D}(\mathcal{R}) = \begin{bmatrix} 0 \\ P_\Omega(\mathcal{R}) \end{bmatrix} \quad \mathcal{M} = \begin{bmatrix} \mathcal{M}_1 \\ \mathcal{M}_2 \end{bmatrix}$$

Then Lagrangian function 4.8 can be written down as

$$\begin{aligned} \mathcal{L}_\beta(\mathcal{Q}, \mathcal{G}, \mathcal{F}, \mathcal{M}_1, \mathcal{M}_2) =& \frac{1}{2}\|\mathcal{F}\|_F^2 + \frac{\lambda_Q}{3}\sum_{i=1}^{3}\|\mathcal{Q}_{(i)}\|_* + \lambda_G\|\mathcal{G}\|_1 \\ &+ \Big\langle \mathcal{M}, P_{\mathcal{G}}(\mathcal{G}) + P_{\mathcal{Q}}(\mathcal{Q}) + P_{\mathcal{F}}(\mathcal{F}) - \mathcal{D}(\mathcal{R}) \Big\rangle + \\ &\frac{\beta}{2}\|P_{\mathcal{G}}(\mathcal{G}) + P_{\mathcal{Q}}(\mathcal{Q}) + P_{\mathcal{F}}(\mathcal{F}) - \mathcal{D}(\mathcal{R})\|_F^2 \end{aligned}$$

To derive $(\mathcal{G}^t, \mathcal{Q}^t, \mathcal{F}^t, \mathcal{M}^t)$ converges to the KKT point, we will show two things:

1. $(\mathcal{G}^\infty, \mathcal{Q}^\infty, \mathcal{F}^\infty, \mathcal{M}^\infty)$ is a KKT point of our optimization problem.

2. The subsequence $(\mathcal{G}^t, \mathcal{Q}^t, \mathcal{F}^t, \mathcal{M}^t)$ convergences to $(\mathcal{G}^\infty, \mathcal{Q}^\infty, \mathcal{F}^\infty, \mathcal{M}^\infty)$

We know that there are 4 KKT conditions to check, stationarity condition, complementary slackness, primal feasibility and dual feasibility. Since our optimization problem only has equality constrains, the number of KKT conditions to check shrinks to 2, which are stationarity condition and primal feasibility.

We know that $P_{\mathcal{G}}(\mathcal{G}^t) + P_{\mathcal{Q}}(\mathcal{Q}^t) + P_{\mathcal{F}}(\mathcal{F}^t) - \mathcal{D}(\mathcal{R}) = -\beta_t^{-1}(\mathcal{M}^{t+1} - \mathcal{M}) \to 0$, indicating that any accumulation point of sequence $(\mathcal{G}^t, \mathcal{Q}^t, \mathcal{F}^t, \mathcal{M}^t)$ will be feasible. By Lemma 24, we know that $(\mathcal{G}^t, \mathcal{Q}^t, \mathcal{F}^t, \mathcal{M}^t)$ is bounded and there exists a subsequence $(\mathcal{G}^{t_i}, \mathcal{Q}^{t_i}, \mathcal{F}^{t_i}, \mathcal{M}^{t_i})$ converges to $(\mathcal{G}^\infty, \mathcal{Q}^\infty, \mathcal{F}^\infty, \mathcal{M}^\infty)$. Suppose $\lambda_G = \lambda_Q = 1/2$, we have

$$
\begin{aligned}
&\|\mathcal{G}^{t_i}\|_1 + \frac{1}{3}\sum_{j=1}^{3}\|\mathcal{Q}_{(j)}^{t_i}\|_* + \|\mathcal{F}^{t_i}\|_F^2 \\
\leq &\|\mathcal{G}^*\|_1 + \frac{1}{3}\sum_{j=1}^{3}\|\mathcal{Q}_{(j)}^*\|_* + \|\mathcal{F}^*\|_F^2 + \left\langle \mathcal{G}^{t_i} - \mathcal{G}^*, -\beta_{t_i-1}\rho_1(\mathcal{G}^{t_i} - \mathcal{G}^{t_i-1}) - P_{\mathcal{G}}^*(\bar{\mathcal{M}}^{t_i})\right\rangle \\
&+ \left\langle \mathcal{Q}^{t_i} - \mathcal{Q}^*, -\beta_{t_i-1}\rho_2(\mathcal{Q}^{t_i} - \mathcal{Q}^{t_i-1}) - P_{\mathcal{Q}}^*(\hat{\mathcal{M}}^{t_i})\right\rangle + \left\langle \mathcal{F}^{t_i} - \mathcal{F}^*, -\beta_{t_i-1}(\mathcal{F}^{t_i} - \mathcal{F}^{t_i-1}) - P_{\mathcal{F}}^*(\mathcal{M}^{t_i})\right\rangle
\end{aligned}
$$

Let $i \to \infty$, we have $\mathcal{G}^{t_i} - \mathcal{G}^{t_i-1}, \mathcal{Q}^{t_i} - \mathcal{Q}^{t_i-1}, \mathcal{F}^{t_i} - \mathcal{F}^{t_i-1}$ all go to 0 and we could have

$$
\begin{aligned}
&\|\mathcal{G}^\infty\|_1 + \frac{1}{3}\sum_{j=1}^{3}\|\mathcal{Q}_{(j)}^\infty\|_* + \|\mathcal{F}^\infty\|_F^2 \\
\leq &\|\mathcal{G}^*\|_1 + \frac{1}{3}\sum_{j=1}^{3}\|\mathcal{Q}_{(j)}^*\|_* + \|\mathcal{F}^*\|_F^2 + \left\langle \mathcal{G}^\infty - \mathcal{G}^*, -P_{\mathcal{G}}^*(\bar{\mathcal{M}}^\infty)\right\rangle + \left\langle \mathcal{Q}^\infty - \mathcal{Q}^*, -P_{\mathcal{Q}}^*(\hat{\mathcal{M}}^\infty)\right\rangle \\
&+ \left\langle \mathcal{F}^\infty - \mathcal{F}^*, -P_{\mathcal{F}}^*(\mathcal{M}^\infty)\right\rangle \\
= &\|\mathcal{G}^*\|_1 + \frac{1}{3}\sum_{j=1}^{3}\|\mathcal{Q}_{(j)}^*\|_* + \|\mathcal{F}^*\|_F^2 \\
&- \left\langle P_{\mathcal{G}}(\mathcal{G}^\infty - \mathcal{G}^*) + P_{\mathcal{Q}}(\mathcal{Q}^\infty - \mathcal{Q}^*) + P_{\mathcal{F}}(\mathcal{F}^\infty - \mathcal{F}^*), \mathcal{M}^\infty\right\rangle \\
= &\|\mathcal{G}^*\|_1 + \frac{1}{3}\sum_{j=1}^{3}\|\mathcal{Q}_{(j)}^*\|_* + \|\mathcal{F}^*\|_F^2
\end{aligned}
$$

where the last equality comes from the fact that $(\mathcal{G}^\infty, \mathcal{Q}^\infty, \mathcal{F}^\infty)$ and $(\mathcal{G}^*, \mathcal{Q}^*, \mathcal{F}^*)$ are both

feasible solutions. Therefore, from the above analysis, we can obtain that $(\mathcal{G}^\infty, \mathcal{Q}^\infty, \mathcal{F}^\infty)$ is optimal solution to our optimization problem. Similarly, by proof of Lemma 24, it's not difficult to check the stationarity condition is satisfied. Thus, we can say that $(\mathcal{G}^\infty, \mathcal{Q}^\infty, \mathcal{F}^\infty)$ is KKT point. The remaining work is to show the subsequence $(\mathcal{G}^t, \mathcal{Q}^t, \mathcal{F}^t, \mathcal{M}^t)$ convergences to $(\mathcal{G}^\infty, \mathcal{Q}^\infty, \mathcal{F}^\infty, \mathcal{M}^\infty)$. By setting $(\mathcal{G}^*, \mathcal{Q}^*, \mathcal{F}^*)$ as $(\mathcal{G}^\infty, \mathcal{Q}^\infty, \mathcal{F}^\infty)$, implementing Lemma 24, we have

$$\begin{aligned}&\rho_1\|\mathcal{G}^t - \mathcal{G}^\infty\|_F^2 - \|P_{\mathcal{G}}(\mathcal{G}^t - \mathcal{G}^\infty)\|_F^2 + \rho_2\|\mathcal{Q}^t - \mathcal{Q}^\infty\|_F^2 - \|P_{\mathcal{Q}}(\mathcal{Q}^t - \mathcal{Q}^\infty)\|_F^2 \\ &+ \|\mathcal{F}^t - \mathcal{F}^*\|_F^2 - \|P_{\mathcal{F}}(\mathcal{F}^t - \mathcal{F}^\infty)\|_F^2 + \beta_t^{-2}\|\mathcal{M}^t - \mathcal{M}^\infty\|_F^2 \to 0\end{aligned}$$

Thus, we have $\mathcal{G}^t \to \mathcal{G}^\infty, \mathcal{Q}^t \to \mathcal{Q}^\infty, \mathcal{F}^t \to \mathcal{F}^\infty$ and since a KKT point is the unique optimizer of a convex optimization problem, the whole proof is finished. □

Lemma 24. *Given β_t is non-decreasing and upper bounded, if $\rho_1 \geq \|P_{\mathcal{G}}\|^2, \rho_2 \geq \|P_{\mathcal{Q}}\|^2$ and $(\mathcal{G}^*, \mathcal{Q}^*, \mathcal{F}^*, \mathcal{M}_1^*, \mathcal{M}_2^*)$ be any KKT point of 4.7, then we have the following item*

$$\begin{aligned}\Gamma(\mathcal{G}^t, \mathcal{Q}^t, \mathcal{F}^t, \mathcal{M}^t) =&\rho_1\|\mathcal{G}^t - \mathcal{G}^*\|_F^2 - \|P_{\mathcal{G}}(\mathcal{G}^t - \mathcal{G}^*)\|_F^2 + \rho_2\|\mathcal{Q}^t - \mathcal{Q}^*\|_F^2 - \|P_{\mathcal{Q}}(\mathcal{Q}^t - \mathcal{Q}^*)\|_F^2 \\ &+ \|\mathcal{F}^t - \mathcal{F}^*\|_F^2 - \|P_{\mathcal{F}}(\mathcal{F}^t - \mathcal{F}^*)\|_F^2 + \beta_t^{-2}\|\mathcal{M}^t - \mathcal{M}^*\|_F^2\end{aligned}$$

is non-decreasing and moreover,

$$\|\mathcal{G}^{t+1} - \mathcal{G}^t\| \to 0, \|\mathcal{Q}^{t+1} - \mathcal{Q}^t\| \to 0, \|\mathcal{F}^{t+1} - \mathcal{F}^t\| \to 0, \|\mathcal{M}^{t+1} - \mathcal{M}^t\| \to 0$$

Proof. First, we introduce some facts that are utilized later

1. By nested double ADMM algorithm updating rule for $\mathcal{M}$, we have $\mathcal{M}^{t+1} = \mathcal{M}^t + \beta_t(P_{\mathcal{G}}(\mathcal{G}) + P_{\mathcal{Q}}(\mathcal{Q}) + P_{\mathcal{F}}(\mathcal{F}) - \mathcal{D}(\mathcal{R}))$
2. By simple linear algebra, we can obtain $2\langle\mathcal{G}^{t+1} - \mathcal{G}^*, \mathcal{G}^{t+1} - \mathcal{G}^t\rangle = \|\mathcal{G}^{t+1} - \mathcal{G}^*\|_F^2 -$

$$\|\mathcal{G}^t - \mathcal{G}^*\|_F^2 + \|\mathcal{G}^{t+1} - \mathcal{G}^t\|_F^2$$

3. Since $(\mathcal{G}^*, \mathcal{Q}^*, \mathcal{F}^*)$ is the KKT point, we have $P_{\mathcal{G}}(\mathcal{G}^*) + P_{\mathcal{Q}}(\mathcal{Q}^*) + P_{\mathcal{F}}(\mathcal{F}^*) - \mathcal{D}(\mathcal{R}) = 0$

Next, we will show $\Gamma(\mathcal{G}^{t+1}, \mathcal{Q}^{t+1}, \mathcal{F}^{t+1}, \mathcal{M}^{t+1}) - \Gamma(\mathcal{G}^t, \mathcal{Q}^t, \mathcal{F}^t, \mathcal{M}^t) \leq 0$.

$$\begin{aligned}
&\Gamma(\mathcal{G}^{t+1}, \mathcal{Q}^{t+1}, \mathcal{F}^{t+1}, \mathcal{M}^{t+1}) - \Gamma(\mathcal{G}^t, \mathcal{Q}^t, \mathcal{F}^t, \mathcal{M}^t) \\
=&2\rho_1\langle \mathcal{G}^{t+1} - \mathcal{G}^t, \mathcal{G}^{t+1} - \mathcal{G}^*\rangle - \rho_1\|\mathcal{G}^{t+1} - \mathcal{G}^t\|_F^2 + 2\langle P_{\mathcal{G}}(\mathcal{G}^{t+1} - \mathcal{G}^t), P_{\mathcal{G}}(\mathcal{G}^{t+1} - \mathcal{G}^*)\rangle \\
&- \|P_{\mathcal{G}}(\mathcal{G}^{t+1} - \mathcal{G}^t)\|_F^2 + 2\rho_2\langle \mathcal{Q}^{t+1} - \mathcal{Q}^t, \mathcal{Q}^{t+1} - \mathcal{Q}^*\rangle - \rho_2\|\mathcal{Q}^{t+1} - \mathcal{Q}^t\|_F^2 \\
&+ 2\langle P_{\mathcal{Q}}(\mathcal{Q}^{t+1} - \mathcal{Q}^t), P_{\mathcal{Q}}(\mathcal{Q}^{t+1} - \mathcal{Q}^*)\rangle - \|P_{\mathcal{Q}}(\mathcal{Q}^{t+1} - \mathcal{Q}^t)\|_F^2 \\
&+ 2\langle \mathcal{F}^{t+1} - \mathcal{F}^t, \mathcal{F}^{t+1} - \mathcal{F}^*\rangle - \|\mathcal{F}^{t+1} - \mathcal{F}^t\|_F^2 + 2\langle P_{\mathcal{F}}(\mathcal{F}^{t+1} - \mathcal{F}^t), P_{\mathcal{F}}(\mathcal{F}^{t+1} - \mathcal{F}^*)\rangle \\
&- \|P_{\mathcal{F}}(\mathcal{F}^{t+1} - \mathcal{F}^t)\|_F^2 + 2\beta_t^{-2}\langle \mathcal{M}^{t+1} - \mathcal{M}^t, \mathcal{M}^{t+1} - \mathcal{M}^*\rangle - \beta_t^{-2}\|\mathcal{M}^{t+1} - \mathcal{M}^t\|_F^2 \\
=&-(\rho_1\|\mathcal{G}^{t+1} - \mathcal{G}^t\|_F^2 - \|P_{\mathcal{G}}(\mathcal{G}^{t+1} - \mathcal{G}^t)\|_F^2) - (\rho_2\|\mathcal{Q}^{t+1} - \mathcal{Q}^t\|_F^2 - \|P_{\mathcal{Q}}(\mathcal{Q}^{t+1} - \mathcal{Q}^t)\|_F^2) \\
&- (\|\mathcal{F}^{t+1} - \mathcal{F}^t\|_F^2 - \|P_{\mathcal{F}}(\mathcal{F}^{t+1} - \mathcal{F}^t)\|_F^2) - \beta_t^{-2}\|\mathcal{M}^{t+1} - \mathcal{M}^t\|_F^2 \\
&+ 2\rho_1\langle \mathcal{G}^{t+1} - \mathcal{G}^t, \mathcal{G}^{t+1} - \mathcal{G}^*\rangle + 2\langle P_{\mathcal{G}}(\mathcal{G}^{t+1} - \mathcal{G}^t), P_{\mathcal{G}}(\mathcal{G}^{t+1} - \mathcal{G}^*)\rangle \\
&+ 2\rho_2\langle \mathcal{Q}^{t+1} - \mathcal{Q}^t, \mathcal{Q}^{t+1} - \mathcal{Q}^*\rangle + 2\langle P_{\mathcal{Q}}(\mathcal{Q}^{t+1} - \mathcal{Q}^t), P_{\mathcal{Q}}(\mathcal{Q}^{t+1} - \mathcal{Q}^*)\rangle \\
&+ 2\langle \mathcal{F}^{t+1} - \mathcal{F}^t, \mathcal{F}^{t+1} - \mathcal{F}^*\rangle + 2\langle P_{\mathcal{F}}(\mathcal{F}^{t+1} - \mathcal{F}^t), P_{\mathcal{F}}(\mathcal{F}^{t+1} - \mathcal{F}^*)\rangle \\
&+ 2\beta_t^{-2}\langle \mathcal{M}^{t+1} - \mathcal{M}^t, \mathcal{M}^{t+1} - \mathcal{M}^*\rangle
\end{aligned}$$

where the second equality is derived based on fact 2. Since $\rho_1 \geq \|P_{\mathcal{G}}\|^2, \rho_2 \geq \|P_{\mathcal{Q}}\|^2$, it's easy to show $-(\rho_1\|\mathcal{G}^{t+1} - \mathcal{G}^t\|_F^2 - \|P_{\mathcal{G}}(\mathcal{G}^{t+1} - \mathcal{G}^t)\|_F^2), -(\rho_2\|\mathcal{Q}^{t+1} - \mathcal{Q}^t\|_F^2 - \|P_{\mathcal{Q}}(\mathcal{Q}^{t+1} - \mathcal{Q}^t)\|_F^2), -(\|\mathcal{F}^{t+1} - \mathcal{F}^t\|_F^2 - \|P_{\mathcal{F}}(\mathcal{F}^{t+1} - \mathcal{F}^t)\|_F^2), -\beta_t^{-2}\|\mathcal{M}^{t+1} - \mathcal{M}^t\|_F^2$ are all less than 0. It suffices to show $2\rho_1\langle \mathcal{G}^{t+1} - \mathcal{G}^t, \mathcal{G}^{t+1} - \mathcal{G}^*\rangle + 2\langle P_{\mathcal{G}}(\mathcal{G}^{t+1} - \mathcal{G}^t), P_{\mathcal{G}}(\mathcal{G}^{t+1} - \mathcal{G}^*)\rangle + 2\rho_2\langle \mathcal{Q}^{t+1} - \mathcal{Q}^t, \mathcal{Q}^{t+1} - \mathcal{Q}^*\rangle + 2\langle P_{\mathcal{Q}}(\mathcal{Q}^{t+1} - \mathcal{Q}^t), P_{\mathcal{Q}}(\mathcal{Q}^{t+1} - \mathcal{Q}^*)\rangle + 2\langle \mathcal{F}^{t+1} - \mathcal{F}^t, \mathcal{F}^{t+1} - \mathcal{F}^*\rangle + 2\langle P_{\mathcal{F}}(\mathcal{F}^{t+1} - \mathcal{F}^t), P_{\mathcal{F}}(\mathcal{F}^{t+1} - \mathcal{F}^*)\rangle + 2\beta_t^{-2}\langle \mathcal{M}^{t+1} - \mathcal{M}^t, \mathcal{M}^{t+1} - \mathcal{M}^*\rangle \leq 0$. Due to the fact that $\mathcal{G}^t, \mathcal{Q}^t, \mathcal{F}^t$ are the optimal solution for each iterative steps at the tth iteration, we should be able to

obtain

$$-\beta_t\rho_1(\mathcal{G}^{t+1}-\mathcal{G}^t)-P_{\mathcal{G}}^*(\bar{\mathcal{M}}^{t+1})\in\partial\|\mathcal{G}^{t+1}\|_1$$
$$-\beta_t\rho_2(\mathcal{Q}_{(i)}^{t+1}-\mathcal{Q}_{(i)}^t)-P_{\mathcal{Q}}^*(\hat{\mathcal{M}}^{t+1})_{(i)}\in\partial\|\mathcal{Q}_{(i)}^{t+1}\|_*$$

where $\bar{\mathcal{M}}^{t+1}=\mathcal{M}^t+\beta_t(P_{\mathcal{G}}(\mathcal{G})+P_{\mathcal{Q}}(\mathcal{Q}^{t+1})+P_{\mathcal{F}}(\mathcal{F}^{t+1})-\mathcal{D}(\mathcal{R}))$, $\hat{\mathcal{M}}^{t+1}=\mathcal{M}^t+\beta_t(P_{\mathcal{G}}(\mathcal{G}^{t+1})+P_{\mathcal{Q}}(\mathcal{Q}^t)+P_{\mathcal{F}}(\mathcal{F}^{t+1})-\mathcal{D}(\mathcal{R}))$ and $P_{\mathcal{G}}^*, P_{\mathcal{Q}}^*$ are the adjoint operators. Furthermore, we can do the following derivation,

$$\begin{aligned}
&-2\beta_t^{-1}\Big\langle\mathcal{G}^{t+1}-\mathcal{G}^*,[-\beta_t\rho_1(\mathcal{G}^{t+1}-\mathcal{G}^t)-P_{\mathcal{G}}^*(\bar{\mathcal{M}}^{t+1})]+P_{\mathcal{G}}^*(\mathcal{M}^*)\Big\rangle\\
=&2\rho_1\Big\langle\mathcal{G}^{t+1}-\mathcal{G}^*,\mathcal{G}^{t+1}-\mathcal{G}^t\Big\rangle+2\Big\langle P_{\mathcal{G}}(\mathcal{G}^{t+1}-\mathcal{G}^*),\bar{\mathcal{M}}^{t+1}-\mathcal{M}^*\Big\rangle\\
=&2\rho_1\Big\langle\mathcal{G}^{t+1}-\mathcal{G}^*,\mathcal{G}^{t+1}-\mathcal{G}^t\Big\rangle-2\Big\langle P_{\mathcal{G}}(\mathcal{G}^{t+1}-\mathcal{G}^*),\mathcal{M}^{t+1}-\mathcal{M}^*\Big\rangle+2\Big\langle P_{\mathcal{G}}(\mathcal{G}^{t+1}-\mathcal{G}^*),P_{\mathcal{G}}(\mathcal{G}^{t+1}-\mathcal{G}^t)\Big\rangle
\end{aligned}$$

By analogy, we can obtain

$$\begin{aligned}
&-2\beta_t^{-1}\Big\langle\mathcal{Q}^{t+1}-\mathcal{Q}^*,[-\beta_t\rho_2(\mathcal{Q}^{t+1}-\mathcal{Q}^t)-P_{\mathcal{Q}}^*(\hat{\mathcal{M}}^{t+1})]+P_{\mathcal{Q}}^*(\mathcal{M}^*)\Big\rangle\\
=&2\rho_2\Big\langle\mathcal{Q}^{t+1}-\mathcal{Q}^*,\mathcal{Q}^{t+1}-\mathcal{Q}^t\Big\rangle-2\Big\langle P_{\mathcal{Q}}(\mathcal{Q}^{t+1}-\mathcal{Q}^*),\mathcal{M}^{t+1}-\mathcal{M}^*\Big\rangle+2\Big\langle P_{\mathcal{Q}}(\mathcal{Q}^{t+1}-\mathcal{Q}^*),P_{\mathcal{Q}}(\mathcal{Q}^{t+1}-\mathcal{Q}^t)\Big\rangle\\
&-2\beta_t^{-1}\Big\langle\mathcal{F}^{t+1}-\mathcal{F}^*,[-\beta_t(\mathcal{F}^{t+1}-\mathcal{F}^t)-P_{\mathcal{F}}^*(\mathcal{M}^{t+1})]+P_{\mathcal{F}}^*(\mathcal{M}^*)\Big\rangle\\
=&2\Big\langle\mathcal{F}^{t+1}-\mathcal{F}^*,\mathcal{F}^{t+1}-\mathcal{F}^t\Big\rangle-2\Big\langle P_{\mathcal{F}}(\mathcal{F}^{t+1}-\mathcal{F}^*),\mathcal{M}^{t+1}-\mathcal{M}^*\Big\rangle+2\Big\langle P_{\mathcal{F}}(\mathcal{F}^{t+1}-\mathcal{F}^*),P_{\mathcal{F}}(\mathcal{F}^{t+1}-\mathcal{F}^t)\Big\rangle
\end{aligned}$$

Also, based on fact 1 and fact 3, we have

$$\begin{aligned}
&2\beta_t^{-2}\Big\langle\mathcal{M}^{t+1}-\mathcal{M}^t,\mathcal{M}^{t+1}-\mathcal{M}^*\Big\rangle\\
=&2\beta_t^{-1}\Big\langle\mathcal{M}^{t+1}-\mathcal{M}^*,P_{\mathcal{G}}(\mathcal{G}^{t+1})+P_{\mathcal{Q}}(\mathcal{Q}^{t+1})+P_{\mathcal{F}}(\mathcal{F}^{t+1})-\mathcal{D}(\mathcal{R})\Big\rangle\\
=&2\beta_t^{-1}\Big\langle\mathcal{M}^{t+1}-\mathcal{M}^*,P_{\mathcal{G}}(\mathcal{G}^{t+1})+P_{\mathcal{Q}}(\mathcal{Q}^{t+1})+P_{\mathcal{F}}(\mathcal{F}^{t+1})-(P_{\mathcal{G}}(\mathcal{G}^*)+P_{\mathcal{Q}}(\mathcal{Q}^*)+P_{\mathcal{F}}(\mathcal{F}^*))\Big\rangle\\
=&2\beta_t^{-1}\Big\langle\mathcal{M}^{t+1}-\mathcal{M}^*,P_{\mathcal{G}}(\mathcal{G}^{t+1})-\mathcal{G}^*\Big\rangle+2\beta_t^{-1}\Big\langle\mathcal{M}^{t+1}-\mathcal{M}^*,P_{\mathcal{Q}}(\mathcal{Q}^{t+1}-\mathcal{Q}^*)\Big\rangle\\
&+2\beta_t^{-1}\Big\langle\mathcal{M}^{t+1}-\mathcal{M}^*,P_{\mathcal{F}}(\mathcal{F}^{t+1}-\mathcal{F}^*)\Big\rangle
\end{aligned}$$

Therefore,

$$
\begin{aligned}
&2\rho_1\langle \mathcal{G}^{t+1}-\mathcal{G}^t, \mathcal{G}^{t+1}-\mathcal{G}^*\rangle + 2\langle P_{\mathcal{G}}(\mathcal{G}^{t+1}-\mathcal{G}^t), P_{\mathcal{G}}(\mathcal{G}^{t+1}-\mathcal{G}^*)\rangle \\
&+ 2\rho_2\langle \mathcal{Q}^{t+1}-\mathcal{Q}^t, \mathcal{Q}^{t+1}-\mathcal{Q}^*\rangle + 2\langle P_{\mathcal{Q}}(\mathcal{Q}^{t+1}-\mathcal{Q}^t), P_{\mathcal{Q}}(\mathcal{Q}^{t+1}-\mathcal{Q}^*)\rangle \\
&+ 2\langle \mathcal{F}^{t+1}-\mathcal{F}^t, \mathcal{F}^{t+1}-\mathcal{F}^*\rangle + 2\langle P_{\mathcal{F}}(\mathcal{F}^{t+1}-\mathcal{F}^t), P_{\mathcal{F}}(\mathcal{F}^{t+1}-\mathcal{F}^*)\rangle \\
&+ 2\beta_t^{-2}\langle \mathcal{M}^{t+1}-\mathcal{M}^t, \mathcal{M}^{t+1}-\mathcal{M}^*\rangle \\
=&-2\beta_t^{-1}\Big\langle \mathcal{G}^{t+1}-\mathcal{G}^*, [-\beta_t\rho_1(\mathcal{G}^{t+1}-\mathcal{G}^t) - P^*_{\mathcal{G}}(\bar{\mathcal{M}}^{t+1})] + P^*_{\mathcal{G}}(\mathcal{M}^*)\Big\rangle \\
&-2\beta_t^{-1}\Big\langle \mathcal{Q}^{t+1}-\mathcal{Q}^*, [-\beta_t\rho_2(\mathcal{Q}^{t+1}-\mathcal{Q}^t) - P^*_{\mathcal{Q}}(\hat{\mathcal{M}}^{t+1})] + P^*_{\mathcal{Q}}(\mathcal{M}^*)\Big\rangle \\
&-2\beta_t^{-1}\Big\langle \mathcal{F}^{t+1}-\mathcal{F}^*, [-\beta_t(\mathcal{F}^{t+1}-\mathcal{F}^t) - P^*_{\mathcal{F}}(\mathcal{M}^{t+1})] + P^*_{\mathcal{F}}(\mathcal{M}^*)\Big\rangle \le 0
\end{aligned}
$$

where the last inequality comes from the monotonicity of subgradient mapping. By now, we have successfully prove the non-increasing property of $\Gamma(\mathcal{G}^t, \mathcal{Q}^t, \mathcal{F}^t, \mathcal{M}^t)$. Combining with its non-negativity, we will have as $t \to \infty$, we have

$$
\|\mathcal{G}^{t+1}-\mathcal{G}^t\| \to 0, \|\mathcal{Q}^{t+1}-\mathcal{Q}^t\| \to 0, \|\mathcal{F}^{t+1}-\mathcal{F}^t\| \to 0, \|\mathcal{M}^{t+1}-\mathcal{M}^t\| \to 0
$$

hold naturally. □

C.4 Tensor nuclear norm approximation

In this section, we discuss some approximation bounds for nuclear norm of $\mathcal{Q} \in \mathbb{R}^{n_1\times n_2\times n_3}$ that are computable in polynomial time. We adopted sum of nuclear norm, which is the arithmetic mean of the (matrix) nuclear norm of the flattenings:

$$
\|\mathcal{Q}\|_{\#} = (\|\mathcal{Q}_{(1)}\|_* + \|\mathcal{Q}_{(2)}\|_* + \|\mathcal{Q}_{(3)}\|_*)/3
$$

Theoretically, if $n_1 \leq n_2 \leq n_3$, $\|\mathcal{Q}\|_*$ can be bounded by

$$\|\mathcal{Q}\|_{\#} \leq \|\mathcal{Q}\|_* \leq \sqrt{n_2}\|\mathcal{Q}\|_{\#}$$

Another alternative characterizations for nuclear norm of order three tensor is proposed in Friedland and Lim [2018],

$$\max\{\|\mathcal{Q}_{(1)}\|_*, \|\mathcal{Q}_{(2)}\|_*, \|\mathcal{Q}_{(3)}\|_*\} \leq \|\mathcal{Q}\|_* \leq \min\{\sqrt{\min\{r_2, r_3\}}\|\mathcal{Q}_{(1)}\|_*, \sqrt{\min\{r_1, r_3\}}\|\mathcal{Q}_{(2)}\|_*, \sqrt{\min\{r_1, r_2\}}\|\mathcal{Q}_{(3)}\|_*\}$$

where r_1, r_2, r_3 represents tucker rank of $\mathcal{Q}$. We try to use $\max\{\|\mathcal{Q}_{(1)}\|_*, \|\mathcal{Q}_{(2)}\|_*, \|\mathcal{Q}_{(3)}\|_*\}$ as an approximation in LISTAI and compare its performance with nested double ADMM algorithm. We randomly simulate core tensor $\mathcal{G} \in \mathbb{R}^{d\times d\times d}$ and auxiliary information $X, Y, Z \in \mathbb{R}^{n\times d}$ from Uniform [0,1] distribution. The true tensor is formulated via $\mathcal{R} = [\mathcal{G}; X, Y, Z]$. 60% entries in $\mathcal{G}$ are set to 0, which mimics the sparsity in $\mathcal{G}$. The missing proportion p indicates p entries in $\mathcal{R}$ are unrevealed and we treat those unrevealed entries as testing set. The dimension n, d and missing proportion p are set with multiple choices to validate the outperformance of nested double ADMM.

	$\|\mathcal{Q}\|_{\#}$		$\max\{\|\mathcal{Q}_{(1)}\|_*, \|\mathcal{Q}_{(2)}\|_*, \|\mathcal{Q}_{(3)}\|_*\}$	
	Train	Test	Train	Test
$n = 10, d = 2, p = 0.1$	0.0623(0.01360)	0.1285(0.02911)	0.0645(0.01998)	0.2106(0.06310)
$n = 20, d = 2, p = 0.1$	0.0417(0.00824)	0.0701(0.00603)	0.0574(0.01601)	0.173(0.10517)
$n = 10, d = 4, p = 0.1$	0.0268(0.08899)	0.4338(0.02713)	0.0211(0.01088)	1.5852(0.24596)
$n = 10, d = 2, p = 0.5$	0.0675(0.00798)	0.1576(0.02215)	0.0796(0.03060)	0.2569(0.07604)

Table C.1: Comparison of two tensor nuclear norm approximations

C.5 More details on UCLAF dataset

As we mentioned, UCLAF is a highly skewed distributed dataset. Summary statistics for entries in UCLAF is reported in Table C.2. Below are the results for principle component analysis we conduct for user and activity similarity matrix. As we can see from Table C.3,

Min.	25% quantile	Median	Mean	75% quantile	Max.	Number of Missings
0	0	0	0.05503	0	249	8575

Table C.2: UCLAF data distribution

the cumulative proportion of variance explained by the first four components has achieved more than 98% and after doing some further experiments on nested double ADMM algorithm, there is no big difference in the testing error between selecting the top 4 principal components and more than 4 (such as 5, 6, ..., 10) principle components. And similar analysis can be conducted for activity similarity matrix as well.

	1	2	3	4	5
Standard deviation	6.4711	2.7522	1.58289	0.79832	0.60533
Proportion of variance	0.7806	0.1417	0.04671	0.01188	0.00683
Cumulative proportion	0.7806	0.9223	0.96903	0.98091	0.98774

Table C.3: Principal component analysis of user-user similarity matrix

	1	2	3	4	5
Standard deviation	0.4396	0.2157	0.03928	0.01472	1.273e-18
Proportion of variance	0.8	0.1927	0.00639	0.00090	0
Cumulative proportion	0.8000	0.9927	0.99910	1.0000	1.0000

Table C.4: Principal component analysis of activity-activity similarity matrix

Recall that we perform centering strategy to remove the mean of each fiber from the entries in the tensor. Even through there is no theoretical support to guarantee it will converge in a general case, Figure C.1 shows that when centering strategy is applied to UCLAF dataset, the following residuals converges to zero pretty quickly.

$$\text{residual} = (\sum_{i=1}^{n_1} \sum_{i \in \Omega_i} \tilde{\mathcal{R}}_{i,j,k})^2 + (\sum_{j=1}^{n_2} \sum_{j \in \Omega_j} \tilde{\mathcal{R}}_{i,j,k})^2 + (\sum_{k=1}^{n_3} \sum_{k \in \Omega_k} \tilde{\mathcal{R}}_{i,j,k})^2$$

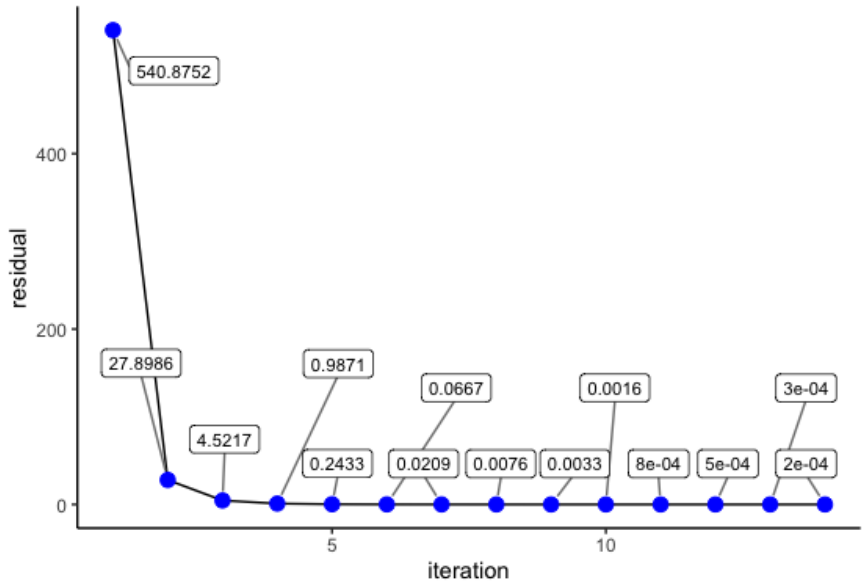


Figure C.1: Convergence analysis of centering on UCLAF

REFERENCES

Alekh Agarwal, Animashree Anandkumar, Prateek Jain, and Praneeth Netrapalli. Learning sparsely used overcomplete dictionaries via alternating minimization. *SIAM Journal on Optimization*, 26(4):2775–2799, 2016.

Genevera Allen. Sparse higher-order principal components analysis. In *Artificial Intelligence and Statistics*, pages 27–36, 2012.

Orly Alter, Patrick O Brown, and David Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18):10101–10106, 2000.

Animashree Anandkumar, Rong Ge, and Majid Janzamin. Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates. *arXiv preprint arXiv:1402.5180*, 2014.

Taylor B Arnold and Ryan J Tibshirani. Efficient implementations of the generalized lasso dual path algorithm. *Journal of Computational and Graphical Statistics*, 25(1):1–27, 2016.

Arnab Auddy and Ming Yuan. On estimating rank-one spiked tensors in the presence of heavy tailed errors. *arXiv preprint arXiv:2107.09660*, 2021.

Brett W Bader, Tamara G Kolda, et al. Matlab tensor toolbox version 2.5. *Available online, January*, 7, 2012.

Linas Baltrunas, Marius Kaminskas, Bernd Ludwig, Omar Moling, Francesco Ricci, Aykan Aydin, Karl-Heinz Lüke, and Roland Schwaiger. Incarmusic: Context-aware music recommendations in a car. In *International conference on electronic commerce and web technologies*, pages 89–100. Springer, 2011.

Boaz Barak and Ankur Moitra. Noisy tensor completion via the sum-of-squares hierarchy. In *Conference on Learning Theory*, pages 417–445. PMLR, 2016.

Dimitris Bertsimas and Colin Pawlowski. Tensor completion with noisy side information. 2020.

Xuan Bi, Annie Qu, Junhui Wang, and Xiaotong Shen. A group-specific recommender system. *Journal of the American Statistical Association*, 112(519):1344–1353, 2017.

Xuan Bi, Annie Qu, and Xiaotong Shen. Multilayer tensor factorization with applications to recommender systems. *The Annals of Statistics*, 46(6B):3308–3333, 2018.

Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. Concentration inequalities. In *Summer school on machine learning*, pages 208–240. Springer, 2003.

Rasmus Bro. Parafac. tutorial and applications. *Chemometrics and intelligent laboratory systems*, 38(2):149–171, 1997.

Norris I Bruce, BPS Murthi, and Ram C Rao. A dynamic model for digital advertising: The effects of creative format, message content, and targeting on engagement. *Journal of marketing research*, 54(2):202–218, 2017.

Tianxi Cai, T Tony Cai, and Anru Zhang. Structured matrix completion with applications to genomic data integration. *Journal of the American Statistical Association*, 111(514): 621–633, 2016.

Tony Cai and Wen-Xin Zhou. A max-norm constrained minimization approach to 1-bit matrix completion. *J. Mach. Learn. Res.*, 14(1):3619–3647, 2013.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

J Douglas Carroll and Jih-Jie Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. *Psychometrika*, 35 (3):283–319, 1970.

O. Celma. *Music Recommendation and Discovery in the Long Tail*. Springer, 2010.

Gary K. Chen, Eric C. Chi, John M.O. Ranola, and Kenneth Lange. Convex clustering: An attractive alternative to hierarchical clustering. *PLoS Computational Biology*, 11(5): e1004228, 2015. doi: 10.1371/journal.pcbi.1004228.

Jiahua Chen and Zehua Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

Jiahua Chen and Zehua Chen. Extended bic for small-n-large-p sparse glm. *Statistica Sinica*, pages 555–574, 2012.

Yudong Chen, Srinadh Bhojanapalli, Sujay Sanghavi, and Rachel Ward. Coherent matrix completion. In *International Conference on Machine Learning*, pages 674–682. PMLR, 2014.

Eric C. Chi and Kenneth Lange. Splitting methods for convex clustering. *Journal of Computational and Graphical Statistics*, 24(4):994–1013, 2015. doi: 10.1080/10618600.2014.948181.

Eric C. Chi, Genevera I. Allen, and Richard G. Baraniuk. Convex biclustering. *Biometrics*, 73(1):10–19, 2017. doi: 10.1111/biom.12540. URL `http://dx.doi.org/10.1111/biom.12540`.

Eric C. Chi, Brian R. Gaines, Will Wei Sun, Hua Zhou, and Jian Yang. Provable convex co-clustering of tensors. arXiv:1803.06518 [stat.ME], 2018.

Kai-Yang Chiang, Cho-Jui Hsieh, and Inderjit S Dhillon. Matrix completion with noisy side information. *Advances in Neural Information Processing Systems*, 28, 2015.

Nadav Cohen, Or Sharir, Yoav Levine, Ronen Tamari, David Yakira, and Amnon Shashua. Analysis and design of convolutional networks via hierarchical tensor decompositions. *arXiv preprint arXiv:1705.02302*, 2017.

Pierre Comon, Xavier Luciani, and André LF De Almeida. Tensor decompositions, alternating least squares and other tales. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 23(7-8):393–405, 2009.

Mark A Davenport, Yaniv Plan, Ewout Van Den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3(3):189–223, 2014.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

Yixin Fang and Junhui Wang. Selection of the number of clusters via the bootstrap method. *Computational Statistics & Data Analysis*, 56(3):468–477, 2012.

Soheil Feizi, Hamid Javadi, and David Tse. Tensor biclustering. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1311–1320, 2017.

Shmuel Friedland and Lek-Heng Lim. Nuclear norm of higher-order tensors. *Mathematics of Computation*, 87(311):1255–1281, 2018.

Evgeny Frolov and Ivan Oseledets. Tensor methods and recommender systems. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(3):e1201, 2017.

Silvia Gandy, Benjamin Recht, and Isao Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse problems*, 27(2):025010, 2011.

Navid Ghadermarzy, Yaniv Plan, and Özgür Yilmaz. Near-optimal sample complexity for convex tensor completion. *Information and Inference: A Journal of the IMA*, 8(3):577–619, 2019.

David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

Trevor Hastie, Rahul Mazumder, Jason D Lee, and Reza Zadeh. Matrix completion and low-rank svd via fast alternating least squares. *The Journal of Machine Learning Research*, 16(1):3367–3402, 2015.

Victoria Hore, Ana Vinuela, Alfonso Buil, Julian Knight, Mark I McCarthy, Kerrin Small, and Jonathan Marchini. Tensor decomposition for multiple-tissue gene expression experiments. *Nature genetics*, 48(9):1094–1100, 2016.

Furong Huang, Sergiy Matusevych, Anima Anandkumar, Nikos Karampatziakis, and Paul Mineiro. Distributed latent dirichlet allocation via tensor factorization. In *NIPS Optimization Workshop*, volume 1, 2014.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1): 193–218, 1985.

Hilda S Ibriga and Will Wei Sun. Covariate-assisted sparse tensor completion. *arXiv preprint arXiv:2103.06428*, 2021.

Prateek Jain and Inderjit S Dhillon. Provable inductive matrix completion. *arXiv preprint arXiv:1306.0626*, 2013.

Prateek Jain and Sewoong Oh. Provable tensor factorization with missing data. *Advances in Neural Information Processing Systems*, 27, 2014.

Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems*, volume 21, 2009.

U Kang, Evangelos Papalexakis, Abhay Harpale, and Christos Faloutsos. Gigatensor: scaling tensor analysis up by 100 times-algorithms and discoveries. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 316–324, 2012.

Charles Kemp, Joshua B Tenenbaum, Thomas L Griffiths, Takeshi Yamada, and Naonori Ueda. Learning systems of concepts with an infinite relational model. In *AAAI*, volume 3, page 5, 2006.

Yong-Deok Kim, Eunhyeok Park, Sungjoo Yoo, Taelim Choi, Lu Yang, and Dongjun Shin. Compression of deep convolutional neural networks for fast and low power mobile applications. *arXiv preprint arXiv:1511.06530*, 2015.

Olga Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014.

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

Jean Kossaifi, Zachary C Lipton, Arinbjorn Kolbeinsson, Aran Khanna, Tommaso Furlanello, and Anima Anandkumar. Tensor regression networks. *arXiv preprint arXiv:1707.08308*, 2017.

Jean Kossaifi, Yannis Panagakis, Anima Anandkumar, and Maja Pantic. Tensorly: Tensor learning in python. *The Journal of Machine Learning Research*, 20(1):925–930, 2019.

Vadim Lebedev, Yaroslav Ganin, Maksim Rakhuba, Ivan Oseledets, and Victor Lempitsky. Speeding-up convolutional neural networks using fine-tuned cp-decomposition. *arXiv preprint arXiv:1412.6553*, 2014.

Chanwoo Lee and Miaoyan Wang. Tensor denoising and completion based on ordinal observations. In *International Conference on Machine Learning*, pages 5778–5788. PMLR, 2020.

Xiaoshan Li, Da Xu, Hua Zhou, and Lexin Li. Tucker tensor regression and neuroimaging analysis. *Statistics in Biosciences*, 10(3):520–545, 2018.

Lin Liao, Dieter Fox, and Henry A Kautz. Location-based activity recognition using relational markov networks. In *IJCAI*, volume 5, pages 773–778. Citeseer, 2005.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. Tensor completion for estimating missing values in visual data. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):208–220, 2012.

Li Liu, Douglas M Hawkins, Sujoy Ghosh, and S Stanley Young. Robust singular value decomposition analysis of microarray data. *Proceedings of the National Academy of Sciences*, 100(23):13167–13172, 2003.

Tianqi Liu, Ming Yuan, and Hongyu Zhao. Characterizing spatiotemporal transcriptome of human brain via low rank tensor decomposition. *arXiv preprint arXiv:1702.07449*, 2017.

Tianqi Liu, Ming Yuan, and Hongyu Zhao. Characterizing spatiotemporal transcriptome of the human brain via low-rank tensor decomposition. *Statistics in Biosciences*, pages 1–29, 2022.

Wei Ma and George H Chen. Missing not at random in matrix completion: The effectiveness of estimating missingness probabilities under a low nuclear norm assumption. *Advances in Neural Information Processing Systems*, 32, 2019.

Xiaojun Mao, Song Xi Chen, and Raymond KW Wong. Matrix completion with covariate information. *Journal of the American Statistical Association*, 114(525):198–210, 2019.

Cameron Martino, Liat Shenhav, Clarisse A Marotz, George Armstrong, Daniel McDonald, Yoshiki Vázquez-Baeza, James T Morton, Lingjing Jiang, Maria Gloria Dominguez-Bello, Austin D Swafford, et al. Context-aware dimensionality reduction deconvolutes gut microbial community dynamics. *Nature biotechnology*, 39(2):165–168, 2021.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11: 2287–2322, 2010.

Marina Meilă. Comparing clusterings—an information based distance. *Journal of multivariate analysis*, 98(5):873–895, 2007.

Philippos Mordohai and Gérard Medioni. Tensor voting: a perceptual organization approach to computer vision and machine learning. *Synthesis Lectures on Image, Video, and Multimedia Processing*, 2(1):1–136, 2006.

Mahdi Nasiri, M Rezghi, and B Minaei. Fuzzy dynamic tensor decomposition algorithm for recommender system. *UCT Journal of Research in Science, Engineering and Technology*, 2(2):52–55, 2014.

Nagarajan Natarajan and Inderjit S Dhillon. Inductive matrix completion for predicting gene–disease associations. *Bioinformatics*, 30(12):i60–i68, 2014.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *Icml*, volume 11, pages 809–816, 2011.

Madhav Nimishakavi, Bamdev Mishra, Manish Gupta, and Partha 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*, pages 307–316, 2018.

Richard A Olshen and Bala Rajaratnam. Successive normalization of rectangular arrays. *Annals of statistics*, 38(3):1638, 2010.

Piyush Rai, Yingjian Wang, Shengbo Guo, Gary Chen, David Dunson, and Lawrence Carin. Scalable bayesian low-rank decomposition of incomplete multiway tensors. In *International Conference on Machine Learning*, pages 1800–1808. PMLR, 2014.

Piyush Rai, Yingjian Wang, and Lawrence Carin. Leveraging features and networks for probabilistic tensor decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(12), 2011.

Steffen Rendle and Lars Schmidt-Thieme. Pairwise interaction tensor factorization for personalized tag recommendation. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 81–90, 2010.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *international conference on machine learning*, pages 1670–1679. PMLR, 2016.

Ohad Shamir and Shai Shalev-Shwartz. Matrix completion with the trace norm: Learning, bounding, and transducing. *The Journal of Machine Learning Research*, 15(1):3401–3423, 2014.

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

Vatsal Sharan and Gregory Valiant. Orthogonalized als: A theoretically principled tensor decomposition algorithm for practical use. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3095–3104. JMLR. org, 2017.

Yiyuan She et al. Sparse regression with exact clustering. *Electronic Journal of Statistics*, 4:1055–1096, 2010.

Marco Signoretto, Lieven De Lathauwer, and Johan AK Suykens. Nuclear norms for tensors and their use for convex multilinear estimation. *Submitted to Linear Algebra and Its Applications*, 43, 2010.

Nathan Srebro, Jason Rennie, and Tommi Jaakkola. Maximum-margin matrix factorization. *Advances in neural information processing systems*, 17, 2004.

Will Wei Sun and Lexin Li. Store: sparse tensor response regression and neuroimaging analysis. *The Journal of Machine Learning Research*, 18(1):4908–4944, 2017.

Will Wei Sun and Lexin Li. Dynamic tensor clustering. *Journal of the American Statistical Association*, pages 1–28, 2019.

Will Wei Sun, Junwei Lu, Han Liu, and Guang Cheng. Provable sparse tensor decomposition. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):899–916, 2017.

Kean Ming Tan and Daniela Witten. Statistical properties of convex clustering. *Electronic journal of statistics*, 9(2):2324, 2015.

Xiwei Tang, Xuan Bi, and Annie Qu. Individualized multilayer tensor learning with an application in imaging analysis. *Journal of the American Statistical Association*, 115 (530):836–851, 2020.

Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

Ryota Tomioka and Taiji Suzuki. Spectral norm of random tensors. *arXiv preprint arXiv:1407.1870*, 2014.

Ryota Tomioka, Taiji Suzuki, Kohei Hayashi, and Hisashi Kashima. Statistical performance of convex tensor decomposition. *Advances in neural information processing systems*, 24: 972–980, 2011.

Joel Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Ledyard R Tucker. Implications of factor analysis of three-way matrices for measurement of change. *Problems in measuring change*, 15(122-137):3, 1963.

David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

Jiebiao Wang, Eric R Gamazon, Brandon L Pierce, Barbara E Stranger, Hae Kyung Im, Robert D Gibbons, Nancy J Cox, Dan L Nicolae, and Lin S Chen. Imputing gene expression in uncollected tissues within and beyond gtex. *The American Journal of Human Genetics*, 98(4):697–708, 2016a.

Junhui Wang. Consistent selection of the number of clusters via crossvalidation. *Biometrika*, 97(4):893–904, 2010.

Miaoyan Wang and Lexin Li. Learning from binary multiway data: Probabilistic tensor decomposition and its statistical optimality. *Journal of machine learning research*, 21 (154), 2020.

Miaoyan Wang and Yuchen Zeng. Multiway clustering via tensor block models. In *Advances in Neural Information Processing Systems*, pages 713–723, 2019.

Miaoyan Wang, Jonathan Fischer, Yun S Song, et al. Three-way clustering of multi-tissue multi-individual gene expression data using semi-nonnegative tensor decomposition. *The Annals of Applied Statistics*, 13(2):1103–1127, 2019.

Yining Wang and Aarti Singh. Provably correct algorithms for matrix column subset selection with selectively sampled data. *arXiv preprint arXiv:1505.04343*, 2015.

Yu-Xiang Wang, James Sharpnack, Alexander J Smola, and Ryan J Tibshirani. Trend filtering on graphs. *The Journal of Machine Learning Research*, 17(1):3651–3691, 2016b.

Dong Xia and Ming Yuan. On polynomial time methods for exact low rank tensor completion. *arXiv preprint arXiv:1702.06980*, 2017.

Dong Xia, Ming Yuan, and Cun-Hui Zhang. Statistically optimal and computationally efficient low rank tensor completion from noisy entries. *arXiv preprint arXiv:1711.04934*, 2017.

Dong Xia, Ming Yuan, and Cun-Hui Zhang. Statistically optimal and computationally efficient low rank tensor completion from noisy entries. *The Annals of Statistics*, 49(1): 76–99, 2021.

Miao Xu, Rong Jin, and Zhi-Hua Zhou. Speedup matrix completion with side information: Application to multi-label learning. *Advances in neural information processing systems*, 26, 2013.

Jinmian Ye, Guangxi Li, Di Chen, Haiqin Yang, Shandian Zhe, and Zenglin Xu. Block-term tensor neural networks. *Neural Networks*, 130:11–21, 2020.

Ming Yuan and Cun-Hui Zhang. On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics*, 16(4):1031–1068, 2016.

Anru Zhang. Cross: Efficient low-rank tensor completion. *The Annals of Statistics*, 47(2): 936–964, 2019.

Muhan Zhang and Yixin Chen. Inductive matrix completion based on graph neural networks. *arXiv preprint arXiv:1904.12058*, 2019.

Xiang Zhang, Lexin Li, Hua Zhou, Dinggang Shen, et al. Tensor generalized estimating equations for longitudinal imaging analysis. *arXiv preprint arXiv:1412.6592*, 2014.

Yanqing Zhang, Xuan Bi, Niansheng Tang, and Annie Qu. Dynamic tensor recommender systems. *arXiv preprint arXiv:2003.05568*, 2020.

Yanqing Zhang, Xuan Bi, Niansheng Tang, and Annie Qu. Dynamic tensor recommender systems. *Journal of Machine Learning Research*, 22(65):1–35, 2021.

Zhengwu Zhang, Genevera I Allen, Hongtu Zhu, and David Dunson. Tensor network factorizations: Relationships between brain structural connectomes and traits. *Neuroimage*, 197:330–343, 2019.

Vincent Zheng, Bin Cao, Yu Zheng, Xing Xie, and Qiang Yang. Collaborative filtering meets mobile recommendation: A user-centered approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, 2010.

Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th international conference on World wide web*, pages 791–800, 2009.

Kai Zhong, Zhao Song, Prateek Jain, and Inderjit S Dhillon. Provable non-linear inductive matrix completion. *Advances in Neural Information Processing Systems*, 32, 2019.

Hua Zhou, Lexin Li, and Hongtu Zhu. Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552, 2013.

Tengfei Zhou, Hui Qian, Zebang Shen, Chao Zhang, and Congfu Xu. Tensor completion with side information: A riemannian manifold approach. In *IJCAI*, pages 3539–3545, 2017.

Yingying Zhu, Xiaofeng Zhu, Minjeong Kim, Jin Yan, and Guorong Wu. A tensor statistical model for quantifying dynamic functional connectivity. In *International Conference on Information Processing in Medical Imaging*, pages 398–410. Springer, 2017.

Yunzhang Zhu. An augmented admm algorithm with application to the generalized lasso problem. *Journal of Computational and Graphical Statistics*, 26(1):195–204, 2017.