**Probabilistic Decomposition in Machine Learning Problems**

by

Byoungwook Jang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in the University of Michigan
2022

Doctoral Committee:

Professor Alfred O. Hero III, Chair
Associate Professor Dawen Cai
Professor Long Nguyen
Professor Ambuj Tewari

Byoungwook Jang
bwjang@umich.edu
ORCID iD: 0000-0001-7783-2866

# DEDICATION

***To my family who was always there for me***

# ACKNOWLEDGMENTS

I would not have been able to complete my journey at the University of Michigan without the support of my family, mentors, collaborators, and friends. I want to first express my gratitude to my advisor Al Hero. Without his technical guidance and mentorship, I would not have been able to complete my studies at Michigan. He introduced to me a wide range of machine learning problems and guided me through my research journey presented in this thesis. He continued to challenge me to ask the right questions and mature as a researcher. I will continue to learn from his positive attitude to tackle difficult problems even after my departure from Ann Arbor. I also want to thank Shuheng Zhou for all her mentorship in my early years of Ph.D. She provided a tremendous amount of patience and attention to expose me to the world of statistical research, and I would have not been able to continue to do research without her mentorship during my first three years at Michigan.

I would like to thank my committee members. Long Nguyen introduced me to the world of theoretical machine learning problems in my first year, and I appreciated all the discussion of topic modeling ideas in my early years of Ph.D. studies. Ambuj Tewari exposed me to the world of online learning in my first year, and he provided me with numerous opportunities to discuss machine learning problems with my colleagues in the department in his discussion groups. Dawen Cai provided me with a wonderful opportunity to work on the exciting application of Brainbow images for interdisciplinary collaboration. His expertise in neuroimaging and image processing was invaluable in my collaboration with his group.

I would like to thank Roger Fan and Wayne Wang for their friendship and collaboration throughout my research life at the University of Michigan. Chapter 4 was a joint work with Wayne, and my collaboration with Roger under Shuheng eventually led to my work in Chapter 5. Additionally, I would like to thank my friends who helped me through my time at the University of Michigan: Mark, Tim, Zack, Spooner, Roger, Jack, Caleb, Sanjana, Young, and Baekjin. Without their loving support, I would not have been able to overcome difficult times in Ann Arbor and in the army.

And, most importantly, I would like to thank my family for their unconditional love and support. I am forever grateful to my parents for supporting my journey in the states for the past 15 years.

I would not have been possible to complete this journey without their encouragement and mental support. Lastly, I would like to thank my siblings for always being my best friends and the biggest supporters in my life.

# TABLE OF CONTENTS

# LIST OF FIGURES

FIGURE

# LIST OF TABLES

TABLE

# ABSTRACT

Decomposition models for understanding mean and covariance structures from high-dimensional data have attracted a lot of attention in recent years. This thesis visits selected machine learning problems with applications in topic modeling, neuroimaging, and experimental designs and tackles challenges in these applications by incorporating decomposable structures.

The first part of the thesis looks into the statistical learning problems for the applications with decomposable mean structures, namely topic modeling and multi-spectral imaging. The goal of topic modeling and multi-spectral unmixing is to decompose the spectrum for each document (or pixel) in the corpus (or the image of a scene) to find latent topics (or spectra of materials present in multi-spectral images). In topic modeling applications, the number of latent variables is a lot less than the ambient dimension. This allows us to estimate the topic simplex with the geometric approach by minimizing the volume of the topic polytope. In our second application, we aim to trace neurons present in multi-spectral images, called Brainbow images, which capture individual neurons in the brain and allow researchers to distinguish different neurons based on unique combinations of fluorescent colors. Brainbow images, however, have an over-defined problem as the number of unique neuron color combinations is greater than the number of spectral channels. Thus, we reformulate the neuron tracing problem as a hidden Markov model with underlying neuronal processes as latent variables to decompose the observed Brainbow images into individual neurons.

The second part of the thesis studies the decomposition of covariance models for tensor-variate data to introduce a scalable and interpretable structure. In the tensor-variate analysis, the observed data often exhibit spatio-temporal structure, and it is desirable to simultaneously learn partial correlation for each mode of the tensor data. However, estimating the unstructured covariance model for tensor-variate data scales quadratically in terms of the product of all dimensions of the tensor. Instead, we introduce a Kronecker sum model for the square root factor of the precision matrix. This model assumption results in a decomposable covariance matrix motivated by a well-known Sylvester equation.

For the last part of the thesis, we visit the linear contextual bandit problem with missing values to understand the effect of missing probabilities on the cumulative regret, showing that the regret degrades due to missingness by at most the square of minimum sampling probability. By separating the missing values from the context vectors in the covariance model, we can estimate

the linear parameter over time without explicitly imputing the missing values. Our method is applied to the experimental design for collecting gene expression data by sequentially selecting class discriminating DNA probes.

# CHAPTER 1

# Introduction

With the recent technical and scientific advancements, the abundance of data provided a window of opportunity for researchers to tackle new challenges in statistical machine learning problems. As part of this effort, researchers introduced algorithms based on decomposable structures that are both interpretable and scalable for high-dimensional data. The main advantage of this approach is that decompositions in real-life applications provide a simpler and more intuitive view of an unknown system that can be hard to understand due to the large scale and complexity of the underlying population.

In this thesis, we look into methods that are developed with a probabilistic decomposition point of view in mind. Chapter 2 and 3 look into the application where the observed data can be viewed as a mixture of latent signals. We assume that the observed data has a decomposable mean structure and aim to recover the latent signals that generated the data in topic modeling and Brainbow images. In both of these applications, the linear mixing process is assumed to generate the observed data.

The methods in Chapter 4 and 5 are motivated by decomposing the covariance structure for tensor-variate data and sequential data. Chapter 4 presents a new tensor covariance models by imposing decomposable structures based on the Sylvester equation for interpretable and scalable estimation. Chapter 5 studies the linear contextual bandit problem with missing values by separating out the missing signals from the true context variables in the covariance matrix.

## 1.1 Decomposition of Mean

Topic modeling and multi-spectral images have a common underlying structure that allows researchers to gain intuition about the given data despite their high dimensionality. Both of these data have a linear mixing structure with non-negative components. That is, we observe a set of points $x_j \in \mathbb{R}^p$ for $1 \leq j \leq n$ that are generated by $K$ latent variables. The linear mixture structure is generated through basis elements $\beta_k \in \mathbb{R}^p$ for $1 \leq k \leq K$ such that the observed data is

a noisy version of the linear combination of the basis $\beta_k$'s with mixing components $\theta_j \in \Delta^K$ for $1 \leq j \leq n$, where $\Delta^K$ is a $K$-dimensional simplex.

In other words, the data we consider is a convex combination of the basis vector $\beta_k$'s for $1 \leq k \leq K$:

$$x_j \approx \sum_{k=1}^{K} \beta_k \theta_{jk} \text{ for a mixing vector } \theta \in \Delta^K \tag{1.1}$$

In the matrix form, we can rewrite the data generation process as a familiar matrix decomposition form, i.e. $X \approx \Theta\beta \in \mathbb{R}^{n \times p}$, where $\Theta \in \mathbb{R}^{n \times k}$ is the abundance matrix and $\beta \in \mathbb{R}^{k \times p}$ is the basis matrix. As we will see in the following subsections, this linear structure appears in topic modeling and Brainbow images.

In topic modeling applications, each document $x_i$ for $i = 1, \ldots, n$ is observed as a distribution of words that can be seen as a linear combination of underlying topic distributions. The number of latent variables (or topics) is assumed to be less than the number of the ambient dimension ($K \ll p$). Similarly, multi-spectral images from Brainbow models each pixel $x_i$ for $i = 1, \ldots, n$ as a spectrum over different wavelengths (or color spectra), which is a mixture of neurons present in the image. In Brainbow application, however, the number of latent variables (or neuron spectra) is larger than the number of color spectra ($K \gg p$).

The main difference between these two applications is the degree of mixing that can be represented by the support of the corresponding mixing proportion $\theta_i \in \Delta^K$.The degree of mixing in these applications determines the difficulty of inference of the parameters $\beta$. Thus, while the method in Chapter 2 estimates the mixing parameters in a latent variable mixture model by unmixing signals based on $\beta$, such unmixing approach is not necessary for the application in Chapter 3. In fact, there is virtually no overlap between the color spectra in Brainbow images as the underlying neurons can be represented with a unique combination of fluorescent dyes.

### 1.1.1 Topic Modeling

Topic modeling was first introduced in Pritchard et al. [2000] and Blei et al. [2003] to understand text and genetic data. In topic modeling, we have an additional constraint that $\beta_k \in \Delta^p$ for $1 \leq k \leq K$ as the observed text data is on a $p$-dimensional simplex. Mathematically, we observe a document $x_j \in \Delta^p$ such that

$$x_j \approx \sum_{k=1}^{K} \beta_k \theta_{jk} \tag{1.2}$$

where $\beta_k \in \Delta^p$ is a distribution over words and $\theta_j \in \Delta^K$ is the distribution of topics in the $j$-th document with $K \ll p$. The goal of topic modeling algorithms is to estimate $\beta$ and $\Theta$ given a set of documents, each of which is a distribution of words.

One of the popular approaches to recovering the topic vectors relies on the low-rank matrix approximation [Anandkumar et al., 2012, Arora et al., 2013, Fu et al., 2018]. While topic modeling was initially motivated by a generative model in Blei et al. [2003], the connection with matrix factorization literature helped researchers to understand the geometry of the proposed models. Out of many low-rank models, we focus on the non-negative matrix factorization (NMF) with text and image applications. In Chapter 2, we motivate our method by showing that the likelihood of the topic modeling problem is asymptotically equivalent to the log determinant topic simplex. Therefore, we introduce a new method to estimate the topic structures by minimizing the volume of the topic simplex in the latent space.

### 1.1.2 Brainbow Image

Brainbow is an imaging process that captures neurons in the brain, each of which can be distinguished from adjacent neurons based on unique combinations of fluorescent proteins. Specifically, by randomly expressing different ratios of available fluorescent dyes in the given organism, each neuron is observed with a distinctive color. In this application, we observe pixels $x_j \in \mathbb{R}^p$ for $1 \leq j \leq n$ such that

$$x_j = \sum_{k=1}^{K} \beta_k \theta_{jk} \tag{1.3}$$

where $\beta_k \in \mathbb{R}^p_+$ is a spectrum of a $k$-th dye over $p$ spectrum and $\theta_j \in \Delta^K$ is the mixture component of the $j$-th pixel. Therefore, the support of $\theta_j$ represents the presence of latent neuron signals, expressed in the $j$-th pixel. However, as the number of distinct colors in the Brainbow images is a lot higher than the number of dyes ($K > p$), the traditional matrix factorization methods from topic modeling are not applicable in tracing individual neurons.

While the tracing problem can be formulated as an unmixing problem of the pixels $x_j$, Brainbow images have an over-defined problem ($K > p$) that leads to the identifiability problem. Furthermore, due to the process of mixed expressions of fluorescent dyes, each color in the observed image does not necessarily map to unique neurons present in the image. Instead, there have been numerous efforts in bioinformatics literature where they trace the neighboring fragments of the neuronal process to trace individual neurons [Athey et al., 2021]. We adopt the simple formulation of the hidden Markov model with the neuronal process as a latent feature to trace neurons in Chapter 3. While tracing and segmentation approaches for Brainbow images is at an early stage of development relative to hyperspectral remote sensing, our method in Chapter 3 provides a proof of concept for future works to perform volumetric segmentation of the whole brain.

## 1.2 Decomposition of Covariance

In addition to understanding the mean structures of high-dimensional data, estimating the precision matrices for high-dimensional data has attracted a lot of attention in the past couple of decades as well. While the sample covariance matrix for the low dimensional data has attractive properties, the consistency results can be violated in the high-dimensional data where the number of covariates is a lot higher than the number of observations ($n \ll p$). In this regime, it can be shown that the sample covariance matrix contains zero eigenvalues with high probability.

In order to estimate the precision matrix with provable guarantees and provide the interpretability of the final precision matrix, researchers adopted the $\ell_1$-penalized likelihood approach, known as *graphical lasso* [Meinshausen and Buhlmann, 2006, Yuan and Lin, 2007, Rothman et al., 2008, Friedman et al., 2008a, Banerjee et al., 2008]. The sparsity-inducing penalties have been additionally introduced with a minimum concave penalty (MCP) and smoothly clipped absolute deviation penalty (SCAD) in regression problems [Fan and Li, 2001, Zhang et al., 2010, Zhang and Zhang, 2012, Breheny and Huang, 2011]. MCP and SCAD resemble the $\ell_1$-penalty function around the centers, but they penalize the large coefficients of the regression parameters equally to avoid the bias that is present in $\ell_1$-penalized methods. The statistical and optimization behaviors of these non-convex penalties were analyzed in Loh and Wainwright [2015b, 2017] to understand the general M-estimators from the additive and/or multiplicative noise, including the variants of the sparse covariance estimation problem.

Following this line of work for the sparse precision matrix estimation for the multi-variate data, there have been interests in precision matrix estimation for matrix- and tensor-variate data. Such methods have been on-demand as spatio-temporal data are often collected in matrix- or tensor-values.

### 1.2.1 Kronecker Models

The covariance models for tensor-variate data can be naively estimated by vectorizing the data and calculating the sample covariance. Dawid [1981] introduced the first generalization of multivariate analysis for tensor-variate data by vectorizing the matrix variate data (two-dimensional tensor) to model the dependency among both rows and columns. In this model, the tensor samples $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{m_1 \times \cdots m_k}$ are described with the covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}(vec(\boldsymbol{\mathcal{X}})vec(\boldsymbol{\mathcal{X}})^T) \in \mathbb{R}^{m \times m}$, where $m = \prod_{i=1}^{k} m_i$. Even for a matrix-variate case ($k = 2$), the computational complexity and the sample complexity are prohibitive as the number of parameters grows quadratically with $m^2$. In order to circumvent this problem, recent work imposed the Kronecker structure on the covariance matrix instead of directly estimating the unstructured covariance matrix $\boldsymbol{\Sigma}$.

As a part of the first attempt to introduce the Kronecker structure, Tsiligkaridis et al. [2013]

and Zhou [2014] imposed a sparse Kronecker product (KP) structure on the matrix-variate model by modeling $\mathbf{\Sigma} = \mathbf{\Psi}_1 \otimes \cdots \otimes \mathbf{\Psi}_k$. Due to the properties of the Kronecker product, the models introduced by Tsiligkaridis et al. [2013] and Zhou [2014] preserve the Kronecker product structure in both $\Sigma$ and $\Sigma^{-1}$. As an alternative, Kalaitzis et al. [2013] introduced Bigraphical Lasso that imposes the precision matrix to have a Kronecker sum (KS) structure, i.e. $\mathbf{\Sigma}^{-1} = \mathbf{\Psi}_1 \oplus \mathbf{\Psi}_2 = (\mathbf{\Psi_1} \otimes \mathbf{I}_{m_2}) + (\mathbf{I}_{m_1} \otimes \mathbf{\Psi}_2)$. Greenewald et al. [2017] extended the Bigraphical lasso for tensor-variate data. Neither of these structures, however, has a generative model that gives an understanding of how these models can be used. In Chapter 4, we formally introduce our new model (Sylvester graphical lasso) that is motivated by the well-known Sylvester equation.

### 1.2.2 High-dimensional bandit and Experimental design

Multi-armed bandit (MAB) problems have attracted a lot of attention in recent years, as they found applications in clinical trials, recommendation systems, and empirical designs. In the simplest form, the learner is faced with $k$-arms to pull in each round, and a noisy reward for the pulled arm is given to the learner. The goal of the learner is to develop a policy for pulling arms at each round to maximize the cumulative rewards by appropriately balancing between exploration and exploitation.

The addition of contextual vectors for each arm in MAB problems lead researchers to study the theoretical behavior of contextual bandit problems, where each arm $a$ is associated with the corresponding context vector $x_a \in \mathbb{R}^p$. In this problem, the reward is modeled as a function of the contextual vector of the chosen arm. Linear bandits are one of the first functional approach to the contextual bandit problems [Abe et al., 2003, Dani et al., 2008, Auer, 2002, Chu et al., 2011]. As a natural extension, researchers started to explore the high-dimensional feature space in the sequential setting to learn $\beta$ over time by imposing a sparsity constraint on the linear parameter [Abbasi-Yadkori et al., 2012, Gilton and Willett, 2017, Bastani and Bayati, 2020, Wang et al., 2018, Kim and Paik, 2019, Oh et al., 2021]. Similar to the sparse regression problem, only a small subset of the context features are correlated with the reward. However, current research in the linear contextual bandit problems assumes noiseless context variables, which are not necessarily true in experimental design and mobile health.

In Chapter 5, we introduce the modification to the sparse-agnostic lasso bandit [Oh et al., 2021] to incorporate the missing values of the covariates. We decompose the covariance matrix into a missing structure and the underlying covariance matrix for the context variables. This approach allows us to successfully recover the linear parameter without imputing the missing values in the context variables. We demonstrate the performance of our model in a DNA probe selection problem, where the learner is faced with $k$-probes with gene expression measurements. The goal of this

empirical design problem is for the learner to choose the DNA probes that lead to discriminative gene expressions to classify different types of microbiomes.

## 1.3 Main problems and publications

The thesis contains four main chapters. The first two chapters (Chapter 2 and Chapter 3) focus on the decomposition of mean structures in topic modeling and Brainbow images. The last two chapters (Chapter 4 and Chapter 5) focus on the theoretical properties of the convergence of covariances in tensor-variate data and sequential data.

- Chapter 2 studies the geometric approach to topic modeling [Jang and Hero, 2019], which was published in the International Conference on Artificial Intelligence and Statistics. In this chapter, the topic modeling problem is formulated as finding the high-dimensional simplex by minimizing the volume contained by the topic vectors.

- Chapter 3 analyzes the Brainbow data to trace individual neurons in multi-spectral images. While the data generation process can be viewed as an unmixing problem, the tracing problem is tackled with a simple hidden Markov model.

- Chapter 4 introduces a new covariance model for tensor-variate data, called *Sylvester Graphical Lasso*. The generative model is motivated by the well-known Sylvester equations. We present the consistency results for the proposed model and demonstrate the flexibility of the model via empirical data and EEG data. This is based on a joint work with Yu Wang [Wang et al., 2020] and was published in the International Conference on Artificial Intelligence and Statistics.

- Chapter 5 introduces a new modification to the high-dimensional linear bandit problems to cope with missing values in the context variables. The model is applied to the experimental design problem to select the most discriminative DNA probes in microbiology studies. This chapter is submitted to the International Workshop on Machine Learning for Signal Processing.

# CHAPTER 2

# Minimum Volume Topic Modeling

We propose a new topic modeling procedure that takes advantage of the fact that the Latent Dirichlet Allocation (LDA) log-likelihood function is asymptotically equivalent to the logarithm of the volume of the topic simplex. This allows topic modeling to be reformulated as finding the probability simplex that minimizes its volume and encloses the documents that are represented as distributions over words. A convex relaxation of the minimum volume topic model optimization is proposed, and it is shown that the relaxed problem has the same global minimum as the original problem under the separability assumption and the sufficiently scattered assumption introduced by Arora et al. [2013] and Huang et al. [2016]. A locally convergent alternating direction method of multipliers (ADMM) approach is introduced for solving the relaxed minimum volume problem. Numerical experiments illustrate the benefits of our approach in terms of computation time and topic recovery performance.

## 2.1 Introduction

Since the introduction by Blei et al. [2003] and Pritchard et al. [2000], the Latent Dirichlet Allocation (LDA) model has remained an important tool to explore and organize large corpora of texts and images. The goal of topic modeling can be summarized as finding a set of topics that summarizes the observed corpora, where each document is a combination of topics lying on the topic simplex.

There are many extensions of LDA, including a nonparametric extension based on the Dirichlet process called Hierarchical Dirichlet Process [Teh et al., 2005], a correlated topic extension based on the logistic normal prior on the topic proportions [Lafferty and Blei, 2006], and a time-varying topic modeling extension [Blei and Lafferty, 2006]. There are two main approaches for estimation of the parameters of probabilistic topic models: the variational approximation popularized by Blei et al. [2003] and the sampling-based approach studied by Pritchard et al. [2000]. These inference algorithms either approximate or sample from the posterior distributions of the latent variable rep-

resenting the topic labels. Therefore, the estimates do not necessarily have a meaningful geometric interpretation in terms of the topic simplex - complicating the assessment of goodness of fit to the model. In order to address this problem, Yurochkin and Nguyen [2016] introduced Geometric Dirichlet Mean (GDM), a novel geometric approach to topic modeling. It is based on a geometric loss function that is surrogate to the LDA's likelihood and builds upon a weighted k-means clustering algorithm, introducing a bias correction. It avoids excessive redundancy of the latent topic label variables and thus improves computation speed and learning accuracy. This geometric viewpoint was extended to a nonparametric setting [Yurochkin et al., 2017].

LDA-type models also arise in the hyperspectral unmixing problem. Similar to the documents in topic modeling, hyperspectral image pixels are assumed to be mixtures of a few spectral signatures, called endmembers (equivalent to topics). Unmixing procedures aim to identify the number of endmembers, their spectral signatures, and their abundances at each pixel (equivalent to topic proportions). One difference between topic modeling and unmixing is that hyperspectral spectra are not normalized. Nonetheless, algorithms for hyperspectral unmixing are similar to topic model algorithms, and similar models have been applied to both problems. Geometric approaches in the hyperspectral unmixing literature take advantage of the fact that linearly mixed vectors also lie in a simplex set or a positive cone. One of the early geometric approaches to unmixing was introduced in Nascimento and Dias [2005] and Bioucas-Dias [2009], which aim to first identify the K-dimensional subspace of the data and then estimate the endmembers that minimize the volume of the simplex spanned by these endmembers. Bioucas-Dias [2009] estimates the endmembers by minimizing the log determinant of the endmember matrix, as the log-determinant is proportional to the volume of the simplex defined by the endmembers. This idea of minimizing the simplex volume motivated the algorithm proposed in this paper for topic modeling. In Bioucas-Dias [2009], however, the authors experience an optimization issue as their formulation is highly non-convex. It was found that the local minima of the objective in Bioucas-Dias [2009] may be unstable.

The topic modeling problem also has similarities to matrix factorization. In particular, nonnegative matrix factorization, while it does not enforce a sum-to-one constraint, is directly applicable to topic modeling [Deerwester et al., 1990, Xu et al., 2003, Anandkumar et al., 2012, Arora et al., 2013, Fu et al., 2018]. Recover KL, recently introduced by Arora et al. [2013], provides a fast algorithm that identifies the model under a separability assumption, which is the assumption that the sample set includes the vertices of the true topic model (pure endmembers). As the separability assumption is often not satisfied in practice, Fu et al. [2018] introduced a weaker assumption called the sufficiently scattered assumption. We provide a theoretical justification of our geometric minimum value method under this weaker assumption.

### 2.1.1 Contribution

We propose a new geometric inference method for LDA that is formulated as minimizing the volume of the topic simplex. The estimator is shown to be identifiable under the separability assumption and the sufficiently scattered assumption. Compared to Bioucas-Dias [2009], our geometric objective involves $\log\det \beta\beta^T$ instead of $\log|\det \beta|$, making our objective function convex. At the same time, the $\log\det \beta\beta^T$ term remains proportional to the volume enclosed by the topic matrix $\beta$ and simplifies the optimization. In particular, we propose a convex relaxation of the minimization problem whose global minimization is equivalent to the original problem. This relaxed objective function is minimized using an iterative augmented Lagrangian approach, implemented using the alternating direction method of multipliers (ADMM), which is shown to be locally convergent.

### 2.1.2 Notation

We use the following notations. We are given a corpus $W \in \mathbb{D}^{M\times V}$ with $M$ documents, $K$ topics, vocabulary size $V$ and $N_m$ words in document $m$ for $m = 1, \cdots, M$. Let $\mathbb{D}^{n\times p}$ be the space of $n \times p$ row-stochastic matrices. Then, our goal is to decompose $W$ as $W = \theta\beta$, where $\theta \in \mathbb{D}^{M\times K}$ is the matrix of topic proportions, and $\beta \in \mathbb{D}^{K\times V}$ is the topic-term matrix. Finally, $\Delta^d$ represents the $d$-dimensional simplex. It is assumed that the documents in the corpus obey the following generative LDA model.

1. For each topic $\beta_i$ for $i = 1, \cdots, K$
   (a) Draw a topic distribution $\beta_i$
2. For $j$-th document $w^{(j)} \in \mathbb{R}^V$ in the corpus $W$ for $j = 1, \cdots, M$
   (a) Choose the topic proportion $\theta_w \sim Dir(\alpha)$
   (b) For each word $\delta_n$ in the document $w^{(j)}$
      i. Choose a topic $z_n \sim Mult(\theta)$
      ii. Choose a word $\delta_n \sim \beta_{z_n}$

## 2.2 Proposed Approach

We assume that the number $K$ of topics is known in advance and is much smaller than the size of the vocabulary, i.e. $K \ll V$. Furthermore, since LDA models the document as being inside the topic simplex, it is advantageous to represent the documents on a $K$-dimensional subspace basis.

Let $E_K = [e_1, \cdots, e_K]$ be a matrix of dimension $V \times K$ with $K$ orthogonal directions spanning the document subspace. Specifically, we define $E_K$ as the set of $K$ eigenvectors of the sample covariance matrix of the documents $w^{(i)}$, $i = 1, \ldots, M$.

Most of the paper focuses on working with $\widetilde{w}^{(i)} = w^{(i)}E_K \in \mathbb{R}^K$, which corresponds to the coordinates of $w^{(i)}$ in colspan$(E_K)$. Note that we can recover the projected documents in the original $V$-dimensional space by

$$\begin{aligned}\widehat{w}^{(i)} &= \overline{w} + (w^{(i)} - \overline{w})E_K E_K^T \\ &= \overline{w} + (\widetilde{w}^{(i)} - \overline{w}E_K)E_K^T \in \mathbb{R}^V\end{aligned}$$

where $\overline{w}$ is the sample average of the observed documents. Therefore,

$$W = \theta\beta \Rightarrow (W - \theta\beta)E_K = 0$$

where $\Theta$ belongs to the simplex $\Delta^K$. This $K$-dimensional probability simplex is defined by the topic distributions, which are the rows of $\beta E_K \in \mathbb{R}^{K\times K}$. For the rest of the paper, given $\omega \in \mathbb{R}^V$, we denote $\widetilde{\omega}$ as the corresponding coordinates in the projected subspace and $\widehat{\omega}$ as the projected vector in the original $V$-dimensional space.

### 2.2.1 Topic Estimation

Let $\gamma = (\beta E_K)^{-1}$. Then, it follows that $\theta = (WE_K)\gamma$. We know that $\beta E_K$ is invertible as we assume that there are $K$ distinct topics, and the rank of the topic matrix $\beta$ is $K$. Then, as noted in Nascimento and Bioucas-Dias [2012], the likelihood w.r.t. $\Theta$ can be written as

$$\begin{aligned}l(\theta, \beta|W) &= \sum_{i=1}^{M} p(w^{(i)}|\beta, \alpha) \\ &= \sum_{i=1}^{M} \log\left(p(\theta^{(i)} = (w^{(i)}E_K)\gamma|\beta, \alpha) \cdot |\det(\gamma)|\right) \\ &= \sum_{i=1}^{M} \log\left(p(\theta^{(i)} = (w^{(i)}E_K)\gamma|\beta, \alpha)\right) \\ &\qquad + M\log|\det(\gamma)|\end{aligned} \tag{2.1}$$

This formulation gives a nice geometric interpretation.

**Geometric Interpretation of log likelihood:** As we increase the number of documents $M \rightarrow$

$\infty$, the dominant term is $\log|\det(\beta)|$. That is,

$$\begin{aligned}
&\lim_{M\to\infty} \arg\max_{\beta} l(\theta,\beta|W) \\
&= \lim_{M\to\infty} \arg\max_{\widetilde{\beta}} \sum_{i=1}^{M} \log p(W_i|\theta,\beta) \\
&\approx \arg\min_{\beta} \log|\det(\beta E_K)| \\
&= \arg\min_{\gamma} -\log|\det\gamma|
\end{aligned} \tag{2.2}$$

Note that $\log|\det(\beta E_K)|$ is proportional to the volume enclosed by the row vectors of $\beta E_K$, i.e. the topic simplex in the projected subspace. In other words, the estimated topic matrix $\beta$ that minimizes its intrinsic volume is asymptotically equivalent to the asymptotic form of the log-likelihood (2.1). This is the main motivation for our proposal to minimize the volume of the topic simplex.

## 2.3 Minimum Volume Topic Modeling

In the remote sensing literature, Nascimento and Bioucas-Dias [2012] proposed to work with the likelihood (2.1) by modeling $\theta$ as a Dirichlet mixture. However, their endmembers are spectra and do not necessarily satisfy the sum-to-one constraints on the endmember matrix; constraints which are fundamental to topic modeling. These additional constraints on the endmember complicate the minimization of (2.1). The first difficulty arises from the $\log|\det\beta|$ term, as $\beta$ is not a symmetric matrix, which makes the log-likelihood (2.1) non-convex. Due to this non-convexity issue, Nascimento and Bioucas-Dias [2012] propose using a second-order approximation to the $\log\det\beta$ term. Yet, no rigorous justification has been provided for their approach. In contrast, we propose using $\log\det\beta\beta^T$ instead of $\log|\det\beta|$, prove identifiability under the sufficiently scattered assumption, and derive an ADMM update.

As we are optimizing $(\beta E_K)^{-1}$ directly, we use the notation $\gamma = (\beta E_K)^{-1}$ in the sequel. We can then rewrite the objective (2.2) as follows

$$\begin{aligned}
&\widehat{\gamma} = \arg\min_{\gamma\in\mathbb{R}^{K\times K}} -\log|\det(\gamma\gamma^T)| \\
&\text{s.t.} \quad \theta > 0 \quad \theta\mathbf{1} = \mathbf{1} \quad \theta = (WE_K)\gamma \\
&\qquad \beta > 0 \quad \beta\mathbf{1} = \mathbf{1}
\end{aligned} \tag{2.3}$$

where $\beta = \overline{w} + (\gamma^{-1} - \overline{w}E_K)E_K^T$. The first set of constraints corresponds to the sum-to-one and

non-negative constraint on the topic proportions $\theta = (WE_K)\gamma$, and the second constraint imposes the same conditions on $\beta$. Thus, the problem (2.3) provides an exact solution to the asymptotic estimation of (2.1). However, this is not a convenient formulation of the optimization problem, as it involves the constraint on the inverse of $\gamma$. Note that as we assume $\beta \in \mathbb{R}^V$ intrinsically lives in a $K$-dimensional subspace, there is a one-to-one mapping between $\beta$ and $\gamma = (\beta E_K)^{-1}$. Throughout this paper, we will make use of this relationship between $\beta$ and $\gamma$. Here, working with a geometric interpretation of the second set of constraints we propose a relaxed version of (2.3).

**Sum-to-one constraint on $\beta$:** Combined with the non-negativity constraint, the sum-to-one constraint $\beta\mathbf{1} = \mathbf{1}$ forces the rows of $\beta$ to lie in the $K$-dimensional topic simplex within the word simplex. To be specific, $\beta\mathbf{1} = 1$ narrows our search space to be in an affine subspace, which is accomplished with a projection of the documents onto this $K$-dimensional affine subspace. This projection takes care of the sum-to-one constraint in the objective (2.3).

**Non-negativity constraint on $\beta$:** We propose relaxing the non-negativity constraint to the following

$$\sigma_{min}(\gamma) \geq R^{-1}$$

where $\sigma_{min}(\gamma)$ is the minimum singular value of $\gamma$. As illustrated in Figure 1, this is interpreted as replacing the non-negativity constraint on the elements of the matrix $\beta$ with a radius $R$ ball constraint on the rows of the matrix $\beta$. As noted before, there is a mapping between $\gamma$ and $\beta$ through $\beta = \overline{w} + (\gamma^{-1} - \overline{w}E_K)E_K^T$. Thus, if $\gamma^t$ is the current iterate of an iterative optimization algorithm, to be specified below, then we can represent the corresponding $i$-th topic vector in the projected space as $b_i^t = (\gamma^t)^{-1}[i,:] = (\beta^t E_K)[i,:]$. It follows that

$$\begin{aligned} \|b_i^t\|^2 &= \frac{\operatorname{tr}((b_i^t)^T b_i^t)\lambda_{min}(\gamma\gamma^T)}{\lambda_{min}(\gamma\gamma^T)} \leq \frac{\operatorname{tr}((b_i^t)^T b_i^t \gamma\gamma^T)}{\lambda_{min}(\gamma\gamma^T)} \\ &= \frac{\operatorname{tr}(b_i^t\gamma(b_i^t\gamma)^T)}{\lambda_{min}(\gamma\gamma^T)} = \frac{\operatorname{tr}(e_i e_i^T)}{\sigma_{min}(\gamma)^2} \leq R^2 \end{aligned} \tag{2.4}$$

Then, imposing $\sigma_{min}(\gamma) \geq R^{-1}$ results in $\|b_i\|^2 \leq R$. The first inequality in (2.4) comes from the fact that $\lambda_{min}(A)\operatorname{tr}(B) \leq \operatorname{tr}(AB) \leq \lambda_{max}(A)\operatorname{tr}(B)$ for positive semidefinite matrices $A$ and $B$. The second equality in (2.4) comes from the definition that $b_i$ is the $i$-th row of $(\gamma^t)^{-1}$

With this spectral relaxation of the non-negativity constraint, the relaxed version of the problem (2.3) becomes

$$\begin{aligned} \widehat{\gamma} &= \arg\min_{\gamma} -\log|\det(\gamma\gamma^T)| \\ \text{s.t.} \quad &\theta > 0 \quad \theta\mathbf{1} = \mathbf{1} \quad \theta = (WE_K)\gamma \\ &\sigma_{min}(\gamma) \geq R^{-1} \end{aligned} \tag{2.5}$$

Intuitively, as shown in Figure 2.1, the optimization problem (2.3) and (2.5) are equivalent to each

other except that the ball relaxation has expanded the solution space beyond the feasible space.

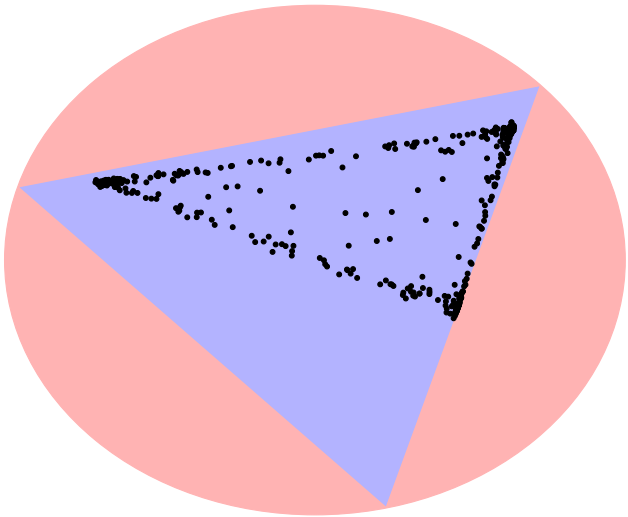


Figure 2.1: Visualization of the difference between the feasible space of (2.3) (blue triangle) and that of (2.5) (pink circle).

The blue triangle represents the set of feasible points for problem (2.3), and the red circle corresponds to the solution space in (2.5).

### 2.3.1 Identifiability

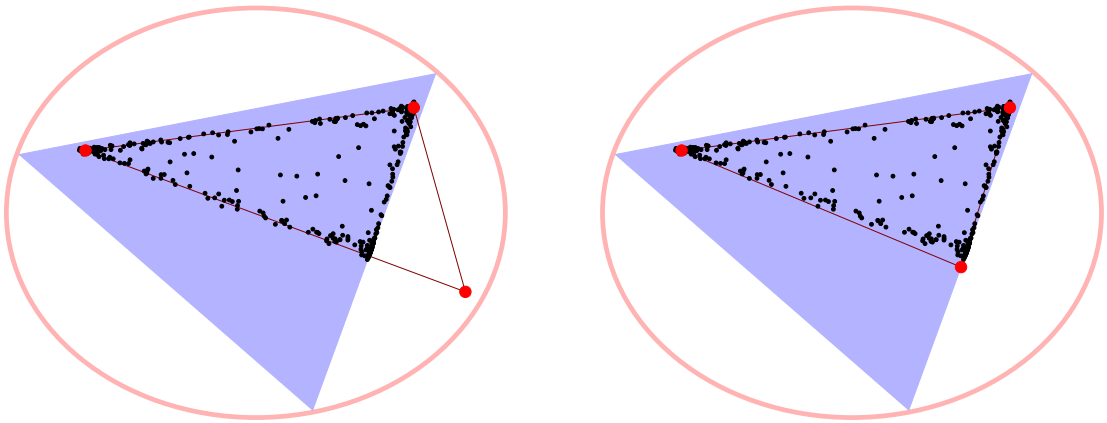
Here we establish the identifiability of the model obtained by solving problem (2.5). Identifiability gained interests in the topic modeling literature (Arora et al. [2013] and Fu et al. [2018]). We show the identifiability under the sufficiently scattered condition. We first state the following lemma.

**Lemma 2.3.1.** *Let $\widehat{\gamma}$ be a solution to the problem* (2.5)*. If $rank(W) = K$, we have that $\widehat{\gamma} \in \Gamma$, where*

$$\Gamma = \{\gamma \in \mathbb{R}^{K\times K} : \beta = \overline{w} + (\gamma^{-1} - \overline{w}E_K)E_K^T \in \mathbb{D}^{K\times V} \\ \text{and } \exists\theta \in \mathbb{D}^{M\times K} \text{ s.t. } \theta = (WE_K)\gamma\}$$

Intuitively, Lemma 2.3.1 tells us that we cannot have the solution outside of the blue triangle in Figure 2.2. If there was a solution outside of the triangle (Figure 2.2a), we could find the projection (Figure 2.2b) onto the word simplex (blue triangle) that still satisfies the constraint yet has a smaller volume, which is a contradiction.

*Proof.* We prove this statement by contradiction. Suppose $\widehat{\gamma} \notin \Gamma$. Then, as $\widehat{\gamma}$ is an optimal solution to the problem (2.5), we have that $\theta = (WE_K)\widehat{\gamma} \in \mathbb{D}^{M\times K}$. Furthermore, since $W \in \Delta^V$ and $E_K$



(a) Simplex $\hat{\beta}$ (b) Simplex $\text{Proj}_{\Delta^V}(\hat{\beta})$

Figure 2.2: Visualization of the proof for Lemma 2.3.1. The set $\Gamma$ corresponds to the blue triangle, which represents the feasible set of the problem (2.3). The red circle represents the feasible set of the relaxed problem (2.5). Given a potential solution $\widehat{\beta}$ for (2.5), we can always argue that the projection of $\widehat{\beta}$, namely $\text{Proj}_{\Delta^V}(\widehat{\beta})$, is a better solution to (2.5) as illustrated in Panel 2.2b.

is obtained from PCA, we have that $E_K^T \mathbf{1}_V = 0$. Thus, it follows that

$$\begin{aligned}\widehat{\beta}\mathbf{1}_V &= \overline{w}\mathbf{1}_V + (\widehat{\gamma}^{-1} - \overline{w}E_K)E_K^T\mathbf{1}_V \\ &= \mathbf{1}_K - (\widehat{\gamma}^{-1} - \overline{w}E_K)\mathbf{0}_V = \mathbf{1}_K\end{aligned}$$

Therefore, the only constraint that $\widehat{\gamma}$ could possibly violate is non-negativity of $\widehat{\beta}$. Let $\text{Proj}_{\Delta^V}(\widehat{\beta})$ be the projection of $\widehat{\beta}$ onto the simplex $\Delta^V$ and let $\widehat{\gamma}_{\text{proj}} = \text{Proj}_{\Delta^V}(\widehat{\beta})E_K$. Then, $\widehat{\gamma}_{\text{proj}} \in \Gamma$ satisfies all the constraints in the optimization problem (2.5), but we also have that

$$-\log(\det \widehat{\gamma}_{\text{proj}}\widehat{\gamma}_{\text{proj}}^T) < -\log(\det \widehat{\gamma}\widehat{\gamma}^T)$$

since the volume of $\text{Proj}_{\Delta^V}(\widehat{\beta})$ is smaller than that of $\widehat{\beta}$. This is a contradiction as $\widehat{\gamma}$ is the optimal solution to the problem (2.5). Thus, it follows that $\widehat{\gamma} \in \Gamma$. □

We now state the sufficiently scattered assumption from Huang et al. [2016].

**Assumption 1:** (*sufficiently scattered condition* (Huang et al. [2016])) Let $\text{cone}(\beta)^* = \{x : \beta x \geq 0\}$ be the polyhedral cone of $\beta$ and $S = \{x : \|x\|_2 \leq 1^T x\}$ be the second order cone. Matrix $\beta$ is called sufficiently scattered if it satisfies:

1) $\text{cone}(\beta)^* \subset S$

2) $\text{cone}(\beta)^* \cap bd(S) = \{ae_k : a \geq 0, k = 1, \cdots, K\}$, where $bdS$ denotes the boundary of $S$.

The sufficiently scattered assumption can be interpreted as an assumption that we observe a sufficient number of documents on the faces of the topic simplex. In real-world topic model appli-

cations, such an assumption is not unreasonable since there are usually documents in the corpora having sparse representations.

**Proposition 2.3.1.** *Let $\gamma_*$ be the optimal solution to the problem* (2.5) *and $\beta_* = \overline{w} + (\gamma_*^{-1} - \overline{w}E_K)E_K^T$ be the corresponding topic matrix. If the true topic matrix $\beta$ is sufficiently scattered and rank$(\widetilde{W}) = K$, then $\beta_* = \beta\Pi$, where $\Pi$ is a permutation matrix.*

The proof structure is similar to the one in Huang et al. [2016], and we include it here for completeness.

*Proof.* Given a corpus $W \in \mathbb{D}^{M\times K}$, let $\beta \in \mathbb{D}^{K\times V}$ be the true topic-word matrix. Suppose rank$(W) = K$ and $\beta$ is sufficiently scattered. Let $\gamma_*$ be the solution to the problem (2.5). Then, by Lemma 2.3.1, we have that $\gamma_* \in \Gamma$. Furthermore, since rank$(W) = K$, we have that rank$(\beta_*) = K$ as $W = \theta\beta_*$ where $\theta = (WE_K)\gamma_*$. It also follows that rank$(\beta) = K$ due to the constraint $W = \theta\beta$. Therefore, $|\det\beta|$ and $|\det\beta_*|$ are strictly positive. In other words, we cannot have a trivial solution to (2.5) as the objective is bounded. As $\beta$ and $\beta_*$ are full row rank, there exists an invertible matrix $Z \in \mathbb{R}^{K\times K}$ such that $\beta_* = Z\beta$. Also, as $\gamma_* \in \Gamma$, it follows that $\beta_* = Z\beta \geq 0$ and

$$\begin{aligned}\beta_*\mathbf{1}_V = Z\beta\mathbf{1}_V = \mathbf{1}_K\\ \Rightarrow Z\mathbf{1}_K = \mathbf{1}_K\end{aligned}$$

The inequality constraint $Z\beta \geq 0$ tells us that rows of $Z$ are contained in cone$(\beta)^*$. As $\beta$ is sufficiently scattered, it follows that

$$Z\big[k,:\big] \in \text{cone}(\beta)^* \subset S \tag{2.6}$$

by the first condition of (A1). Then, by the definition of the second order cone $S$, it follows that

$$\begin{aligned}|\det Z| = |\det Z^T| &\leq \prod_{k=1}^{K}\|Z\big[k,:\big]\|_2\\ &\leq \prod_{k=1}^{K} Z\big[k,:\big]\mathbf{1} = 1\end{aligned} \tag{2.7}$$

The first inequality comes from the Hadamard inequality, which states that the equality holds if and only if the vectors $Z\big[k,:\big]$'s are orthogonal to each other. The second inequality holds when $\|Z[k,:]\|_2 = Z[k,:]\mathbf{1}_K\ \forall k = 1\cdots K$. In other words, when $Z[k,:] \in \text{bd}(S)\ \forall k$. Then, together

with (2.6), it follows that

$$\begin{aligned} Z[k,:] \in\ & \text{cone}(\beta)^* \cap \text{bd}S \\ & = \{\lambda \mathbf{e}_k | \lambda \geq 0, k = 1, \cdots, K\} \end{aligned} \tag{2.8}$$

Thus, it follows that the $|\det Z|$ achieves its maximum at 1, when $Z \in \text{cone}(\beta)^* \cap \text{bd}S$ sums to one and is an orthogonal matrix, i.e. when $Z$ is a permutation matrix.

Furthermore, since $\gamma_* = (\beta_* E_K)^{-1} = Z^{-1}(\beta E_K)^{-1} = Z^{-1}\gamma$, we have that

$$\begin{aligned} -\det(\gamma_* \gamma_*^T) &= -\det(Z^{-1}\gamma\gamma^T(Z^{-1})^T) \\ &= -|\det Z^{-1}| \det(\gamma\gamma^T) |\det Z^{-1}| \\ &= -|\det Z|^{-2} \det(\gamma\gamma^T) \\ &\geq -|\det(\gamma\gamma^T)| \end{aligned}$$

where the equality holds when $|\det Z| = 1$. In other words, the minimum is achieved when $Z$ is a permutation matrix. Therefore, our solution $\gamma_*$ to the problem (2.5) and the corresponding $\beta_*$ are equal to the true topic-word matrix up to permutation. □

**Assumption 2** (*Separability assumption* from Arora et al. [2013]) There exists a set of indices $\Lambda = \{i_1, \cdots, i_K\}$ such that $\beta(\Lambda, :) = \text{Diag(c)}$, where $c \in \mathbb{R}_+^K$.

The separability assumption, also known as the anchor-word assumption, states that every topic $k$ has a unique word $w_k$ that only shows up in topic $k$. These words are also referred to as the anchor words as introduced in Arora et al. [2013].

**Remark:** The identifiability statement in Proposition 2.3.1 holds true under the separability assumption as well, as the sufficiently scattered assumption is a weaker version of the separability assumption.

### 2.3.2 Augmented Lagrangian Formulation

With $\mu \geq 0$, we work with the following augmented Lagrangian version of the constrained optimization problem (2.5)

$$\begin{aligned} \widehat{\gamma} &= \arg\min_{\gamma} -\log|\det(\gamma\gamma^T)| + \mu\|\widetilde{W}\gamma\|_h \\ &\text{s.t.} \quad \gamma \mathbf{1}_K = (\widetilde{W}^T\widetilde{W})^{-1}\widetilde{W}^T\mathbf{1}_V \quad \sigma_{min}(\gamma) \geq \zeta \end{aligned} \tag{2.9}$$

where $\widetilde{W} = WE_K$, $\zeta = R^{-1}$, and $\|X\|_h = \sum_{i,j} \max(-X_{i,j}, 0)$ is a hinge loss that captures the non-negativity constraint on $\theta$. Furthermore, the linear constraint is converted to $\gamma\mathbf{1}_K = (\widetilde{W}^T\widetilde{W})^{-1}\widetilde{W}^T\mathbf{1}_V$, which is the same constraint as $\widetilde{W}\gamma\mathbf{1}_K = \mathbf{1}_V$. For simplicity, we

define $\mathbf{a} = (\widetilde{W}^T\widetilde{W})^{-1}\widetilde{W}^T\mathbf{1}_V$.

The Lagrangian objective function in (2.9) can be written as

$$\begin{aligned} f(\gamma) = &- \log|\det(\gamma\gamma^T)| + \mu\|\widetilde{W}\gamma\|_h \\ &+ \mathbb{1}(\sigma_{min}(\gamma) > \zeta) \text{ s.t. } \gamma\mathbf{1}_K = \mathbf{a} \end{aligned} \tag{2.10}$$

Introducing the auxiliary optimization variables $V_1 \in \mathbb{R}^{n\times k}$ and $V_2 \in \mathbb{R}^{k\times k}$, we reformulate (2.5)

$$\begin{aligned} \hat{\gamma} = \underset{\gamma,V_1,V_2}{\arg\min} \Big\{ &- \log|\det\gamma\gamma^T| + \mu\|V_1\|_h + \\ &+ \mathbb{1}(\sigma_{min}(V_2) > \zeta)\Big\} \\ \text{s.t.} \quad V_1 = \widetilde{W}\gamma \quad &\gamma = V_2 \quad \gamma\mathbf{1}_K = \mathbf{a} \end{aligned} \tag{2.11}$$

For a penalty parameter $\rho > 0$ and Lagrange multiplier matrix $\Lambda \in \mathbb{R}^{n\times k}$, we consider the augmented Lagrangian of this problem

$$\begin{aligned} &\mathcal{L}(\gamma, V_1, V_2, \Lambda_1, \Lambda_2) \\ &= - \log|\det\gamma\gamma^T| + \mu\|V_1\|_h + \mathbb{1}(\sigma_{min}(V_2) > \zeta) \\ &\quad + \frac{\rho}{2}\|\widetilde{W}\gamma - V_1\|_F^2 + \langle\Lambda_1, \widetilde{W}\gamma - V_1\rangle \\ &\quad + \frac{\rho}{2}\|\gamma - V_2\|_F^2 + \langle\Lambda_2, \gamma - V_2\rangle \quad \text{ s.t. } \gamma\mathbf{1}_K = \mathbf{a} \end{aligned} \tag{2.12}$$

This function can be minimized using an iterative ADMM update scheme on the arguments $\gamma$, $V_1$, $V_2$, $\Lambda_1$, and $\Lambda_2$. The update for $V_1$ and $V_2$ can be accomplished by standard proximal operators that implement soft-thresholding and a projection. Furthermore, the $\gamma$-update can be derived in a closed-form by solving a quadratic equation in its singular values. The details of the ADMM updates are included in the supplement. First, consider the $\gamma$-subproblem without the linear constraint $\gamma\mathbf{1} = \mathbf{a}$. Then, as derived in the supplement, the resulting update equation for $\gamma$ is

$$\begin{aligned} \gamma^+ &= \underset{\gamma\in\mathbb{R}^{k\times k}}{\arg\min} \Big\{ - \log|\det\gamma^T\gamma| + \frac{\rho}{2}\|C^{1/2}(\gamma - A)\|_F^2\Big\} \\ &= U\widehat{D}W^T \end{aligned} \tag{2.13}$$

where $\widehat{D}$ is defined in the supplement. Using $\gamma_+$, we obtain a closed-form solution to the $\gamma$ subproblem in (2.12) as follows

$$\gamma^{t+1} = \gamma_+ - (\gamma_+\mathbf{1} - \mathbf{a})(\mathbf{1}^TC^{-1}\mathbf{1})^{-1}\mathbf{1}^TC^{-1}$$

This solution to the linear constrained problem can be easily derived as a stationary point of the convex function that is minimized in (2.13). Note that, by construction, $\gamma^{t+1}\mathbf{1} = \mathbf{a}$.

**Algorithm 1:** Minimum volume topic modeling

**Input:** W, $E_K$, $\gamma^0$, $\rho > 0$, $\mu > 0$
**Output:** $\hat{\beta}$
Initialize $V_2^0 = \gamma^0$, $V_1 = \widetilde{W}\gamma^0$, $\Lambda_1^0 = \mathbf{0}$, $\Lambda_2^0 = \mathbf{0}$ ;
Calculate $C = I + \widetilde{W}^T\widetilde{W}$ ;
Calculate the projected documents $\widetilde{W}$ ;
**while** *not converged* **do**
$V_1^{t+1} = \text{Prox}_{\|\cdot\|_{h,\mu}/\rho}\left(\frac{\rho\widetilde{W}\gamma^t + \Lambda_1^t}{\rho}\right)$
$V_2^{t+1} = \text{Proj}_{G_R}\left(\frac{\rho\gamma^t + \Lambda_2^t}{\rho}\right)$
$\gamma^{t+1} = \gamma_+ - (\gamma_+\mathbf{1} - \mathbf{a})(\mathbf{1}^T C^{-1}\mathbf{1})^{-1}\mathbf{1}^T C^{-1}$
where $\gamma_+$ is defined in (2.13)
$\Lambda_1^{t+1} = \Lambda_1^k + \rho(\widetilde{W}\gamma^{t+1} - V_1^{t+1})$
$\Lambda_2^{t+1} = \Lambda_2^k + \rho(\gamma^{t+1} - V_2^{t+1})$
**end**

In the non-negative matrix factorization literature, Liu et al. [2017] used a large-cone penalty that constrains either the volume or the pairwise angles of the simplex vertices. However, this does not impose a sum-to-one constraint on the topics, and the optimization is performed over $\beta$. Furthermore, our formulation has an advantage over the problem in Liu et al. [2017] as we directly work with the latent topic proportions $\theta$. This is possible in our formulation as we decoupled $\beta$ from $\theta$ using the ADMM mechanism.

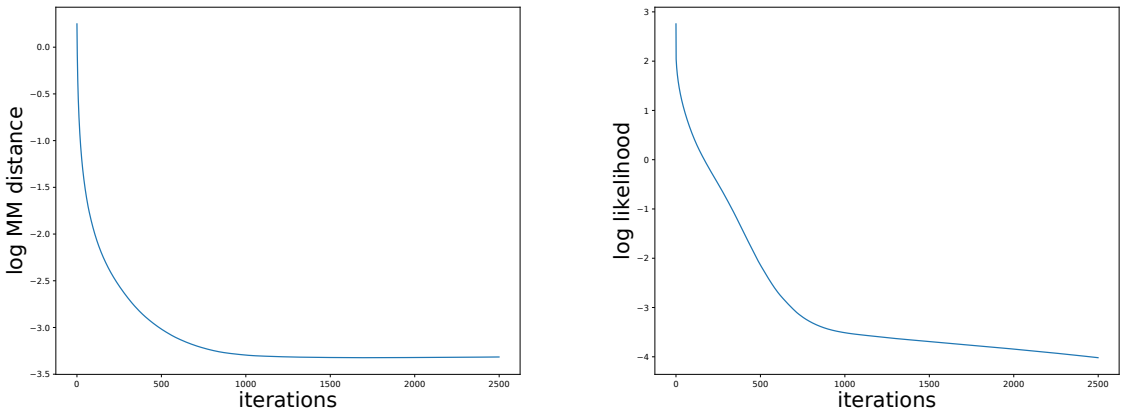
### 2.3.3 Convergence

The following proposition shows that Algorithm 1 converges to a stationary point of (2.10).

**Proposition 2.3.2.** *For any limit point* $(\gamma^*, V_1^*, V_2^*, \Lambda_1^*, \Lambda_2^*)$ *of Algorithm 1,* $\gamma^*$ *is also a stationary point of* (2.10)*.*

This follows by applying a standard convergence proof of the ADMM algorithm (Algorithm 1) based on the KKT condition. The proposition states that our ADMM formulation converges to a stationary point. However, while the unconstrained objective function in (2.10) is convex, the constraint on the minimum singular value makes the constrained optimization function non-convex. Thus, our algorithm is only guaranteed to converge to a stationary point of (2.10).

Figure 2.3 demonstrates the convergence of our algorithm with synthetic data generated from an LDA model with parameters $\alpha = 0.1, \eta = 0.1, V = 1200, K = 3, M = 1000$, and $N_m = 1000$.




(a) Frobenius Norm

(b) Objective

Figure 2.3: Experimental runs using Algorithm 1. The data was simulated from an LDA model with $\alpha = 0.1, \eta = 0.1, V = 1200, K = 3, M = 1000, N_m = 1000$. The algorithm was initialized with $\gamma$ equal to the identity matrix. The left panel shows the relative Frobenius error between the iterates $\gamma^t$ and the true $\gamma$. The right panel shows the convergence in terms of the objective values.

## 2.4 Performance Comparison

To demonstrate the performance of the proposed minimum volume topic model (MVTM) estimation algorithm (Algorithm 1), we generate the LDA data with the parameters $\eta = 0.1, V = 1200, K = 3, M = 1000, N_m = 1000$ with varying $\alpha$, which is Dirichlet hyperparameter for the topic proportion $\theta$. For ease of visualization, the first two dimensions of the projected documents and the estimated topics are used. The first scenario ($\alpha = 0.1$) in Figure 2.4 shows the performance of our algorithm is comparable to the vertex-based method GDM (Yurochkin and Nguyen [2016]), when there are plenty of observed documents around the vertices. While there is no anchor word in the generated dataset, we observe enough documents around the vertices. In other words, the separability assumption is slightly violated. With higher values of $\alpha$, however, Figure 2.5 shows the advantages of our method, denoted as MVTM. Note that the higher values of $\alpha$ correspond to the situation where the sufficiently scattered condition is satisfied, but the separability condition is violated. Thus, we can see the vertex-based method (GDM) starts to suffer in the oracle performance. In contrast, with an appropriate choice of $\mu$ for the hinge loss, our method recovers the correct topics even for the well-mixed scenario where $\alpha = 5$. Figure 2.5b shows that there is a kink in the optimization path, where MVTM is finding the right orientation of the true simplex. Furthermore, there is a lack of loops in the optimization path, illustrating the identifiability of MVTM.

Lastly, we explore the asymptotic behavior by varying document lengths $N_m$ with $M = 1000$, $K = 5$, $V = 1200$, $\eta = 0.1$, $\alpha = 0.1$ and 100 held-out documents. MVTM is all initialized at the

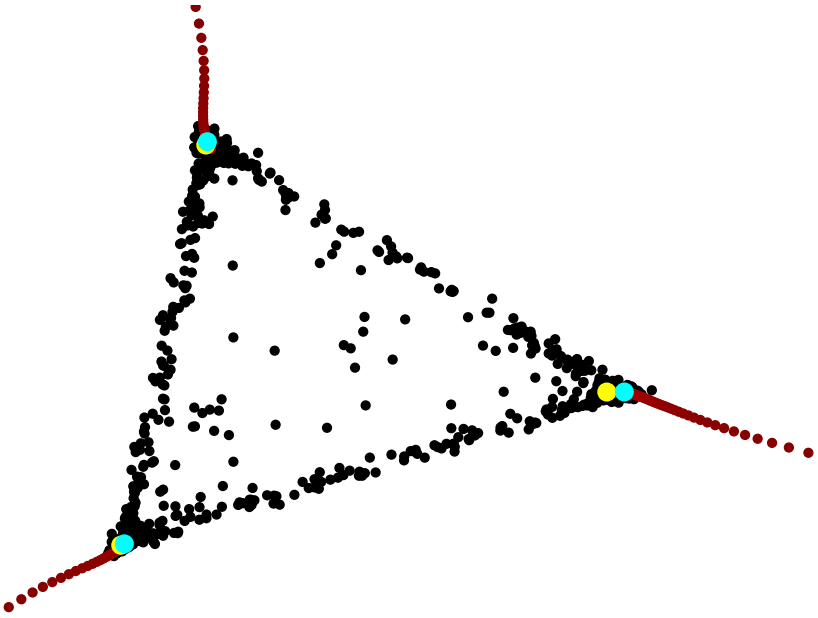


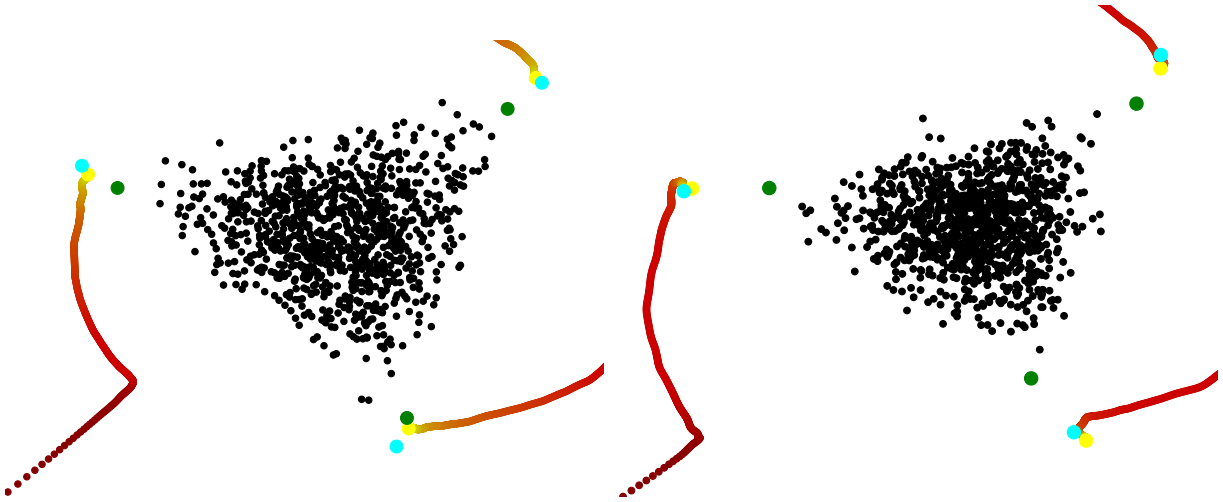
Figure 2.4: Visualization of Minimum Volume Topic Modeling (MVTM) with the observed documents in black, optimization path of MVTM in the gradient of red (dark red = beginning, light red = end), and the final estimate in yellow. The ground-truth topic vertices are plotted in cyan. The Dirichlet parameter for the topic proportion was set at $\alpha = 0.1$, and MVTM was initialized at the identity matrix.

identity matrix, and VEM had 10 restarts as the objective for the variational method is non-convex.

Figure 2.6 tells us that 1) Gibbs sampling and MVTM have comparable performance in terms of perplexity, 2) MVTM and VEM both show the computational advantages over the Gibbs sampling method, and 3) VEM suffers from the statistical performance due to the nature of the non-convex objective function of VEM. Additional simulation results can be found in the supplement.

### 2.4.1 NIPS dataset

To illustrate the performance of MVTM on real-world data, we apply our algorithm to NeurIPS dataset. We preprocess the raw data using a standard stop word list and filter the resulting data through a stemmer. After preprocessing, words that appeared more than 25 times across the whole corpus are retained. Then, we further remove the documents that have less than 10 words. The final dataset contained 4492 unique words and 1491 documents with a mean document length of 1187. We compare our algorithm's performance to GDM and Gibbs sampling at K=5, 10, 15, and 20. The perplexity score is used to perform the comparison in Table 2.1. The additional time comparison and top 10 words of top 10 learned topics for MVTM, GDM, and Gibbs sampling are provided in the supplement.



(a) $\alpha = 3$ (b) $\alpha = 5$

Figure 2.5: Visualization of the proposed MVTM algorithm with the observed documents in black, optimization path of MVTM in the gradient of red (dark red = beginning, light red = end), and the final estimate in yellow under different values of $\alpha$. The ground-truth topic vertices are plotted in cyan, and the final estimate of GDM is plotted in green for comparison. MVTM was initialized at the identity matrix.

| | MVTM | GDM | RecoverKL | Gibbs |
|---|---|---|---|---|
| K=5 | 1483 | 1602 | 1569 | 1336 |
| K=10 | 1387 | 1441 | 1507 | 1192 |
| K=15 | 1293 | 1344 | 1438 | 1109 |
| K=20 | 1273 | 1294 | 1574 | 1068 |

Table 2.1: Perplexity score of the geometric algorithms and the Gibbs sampling for analyzing the NIPS dataset. The proposed algorithm MVTM is performing better than the vertex methods (GDM and RecoverKL) in terms of perplexity as it only requires the documents lie on the face of the topic simplex. GDM provides a similar performance to MVTM.

## 2.5 Discussion

This paper presents a new estimation procedure for LDA topic modeling based on the minimization of the volume of the topic simplex $\beta$. Such formulation can be thought of as an asymptotic estimation of the LDA model. The proposed minimum volume topic model (MVTM) algorithm differs from moment-based methods including RecoverKL and the vertex-based method such as the GDM. We proved the identifiability of MVTM under the sufficiently scattered assumption introduced in Huang et al. [2016]. When the sufficiently scattered assumption is satisfied and the separability assumption is violated, MVTM continues to perform well with an appropriate choice of the hinge loss parameter.

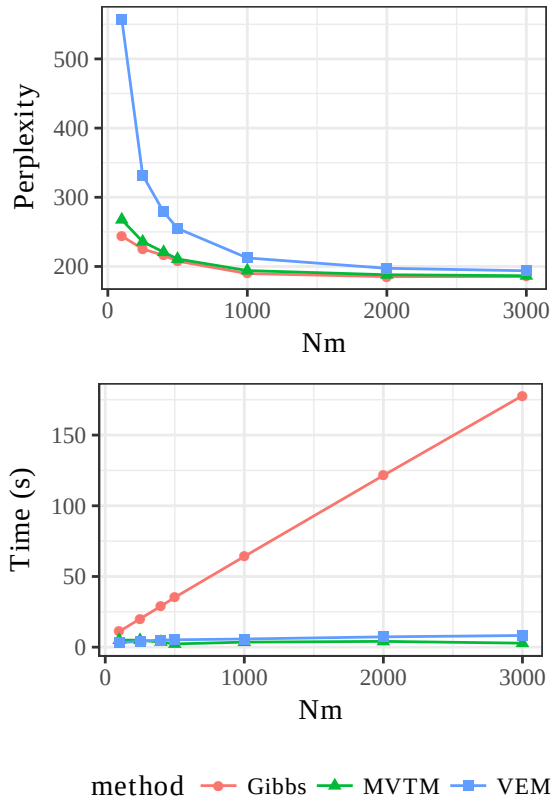



Figure 2.6: Perplexity of the held-out data and the corresponding time complexity of each method at varying values of the number words per document $N_m$ with $M = 1000$, $K = 5$, $V = 1200$, $\eta = 0.1$ and $\alpha = 0.1$

There are open questions on the statistical convergence of our estimator in terms of the document length and the number of documents. Such relationships have been explored in the work of Tang et al. [2014], and it would be interesting to see if these could be applied to the proposed MVTM. The understanding of the statistical behavior of MVTM will provide us with theoretical guidance on the choice of the hinge loss parameter. Besides the theoretical questions, MVTM also has some potential modeling extensions. The immediate extension includes the nonparametric setting, where one would also estimate the number of topics $K$.

## 2.6 Appendix

### 2.6.1 Algorithm Analysis

#### 2.6.1.1 ADMM update derivation

For completeness, we derive the ADMM steps of the problem in (2.12). Given current iterates $V_1^t, \gamma^t$, and $\Lambda^t$,

$$\begin{aligned} V_1^{t+1} &= \underset{V_1 \in \mathbb{R}^{n\times k}}{\arg\min} \Big\{ \mu \|V_1\|_h + \frac{\rho}{2}\|\widetilde{W}\gamma^t - V_1\|_F^2 \\ &\qquad\qquad + \langle \Lambda_1, \widetilde{W}\gamma^t - V_1\rangle \Big\} \\ &= \underset{V\in\mathbb{R}^{n\times K}}{\arg\min} \left\{ \frac{\mu}{\rho}\|V_1\|_h + \frac{1}{2}\|V_1 - \frac{\rho\widetilde{W}\gamma^t + \Lambda_1^t}{\rho}\|_F^2 \right\} \\ &= \text{Soft-Threshold}_{\mu/\rho}\left( \frac{\rho\widetilde{W}\gamma^t + \Lambda_1^t}{\rho} \right) \end{aligned} \tag{2.14}$$

where we soft-threshold the matrix with the regularization parameter $\frac{\lambda}{\rho}$.

$$\begin{aligned} V_2^{t+1} &= \underset{V_2\in\mathbb{R}^{k\times k}}{\arg\min} \Big\{ \frac{\rho}{2}\|\gamma^t - V_2\|_F^2 + \langle \Lambda_2, \gamma^t - V_2\rangle \\ &\qquad\qquad + \mathbb{1}(\lambda_{min}(V_2V_2^T) \geq \frac{1}{R^2}) \Big\} \\ &= \underset{V_2\in\mathbb{R}^{k\times k}}{\arg\min} \Big\{ \frac{1}{2}\|V_2 - \frac{\rho\gamma^t + \Lambda_2^t}{\rho}\|_F^2 \\ &\qquad\qquad + \mathbb{1}(\sigma_{min}(V_2) \geq \frac{1}{R}) \Big\} \\ &= \text{Proj}_{G_R}\left( \frac{\rho\gamma^t + \Lambda_2^t}{\rho} \right) \end{aligned} \tag{2.15}$$

where $G_R = \{X \in \mathbb{R}^{n\times K} | \sigma_{min}(X) \geq \frac{1}{R}\}$ and $\text{Proj}_{G_R}$ is the projection onto the set $G_R$.

$$\begin{aligned} \gamma^{t+1} &= \underset{\gamma\in\mathbb{R}^{k\times k}}{\arg\min} \Big\{ -\log|\det\gamma\gamma^T| + \frac{\rho}{2}\|\widetilde{W}\gamma - V_1\|_F^2 \\ &\qquad\qquad + \langle \Lambda, \widetilde{W}\gamma - V_1\rangle + \frac{\rho}{2}\|\gamma - V_2\|_F^2 \\ &\qquad\qquad + \langle \Lambda_2, \gamma - V_2\rangle \Big\} \quad \text{s.t.} \quad \gamma\mathbf{1}_K = \mathbf{a} \\ &= \underset{\gamma\in\mathbb{R}^{k\times k}}{\arg\min} \Big\{ -\log|\det\gamma\gamma^T| + \frac{\rho}{2}\|C^{1/2}(\gamma - A)\|_F^2 \Big\} \\ &\qquad \text{s.t.} \quad \gamma\mathbf{1}_K = \mathbf{a} \end{aligned} \tag{2.16}$$

where we have that

$$
\begin{aligned}
A &= C^{-1}B^T = UD_AV^T \\
B &= (V_1^{t+1})^T\widetilde{W} + (V_2^{t+1})^T - \frac{(\Lambda_2)^t)^T}{\rho} - \frac{(\Lambda_1)^t)^T\widetilde{W}}{\rho} \\
C &= I + \widetilde{W}^T\widetilde{W}
\end{aligned}
$$

We can derive the update for $\gamma^{t+1}$, as it is a convex problem with a linear constraint. First, consider the (2.16) without the linear constraint $\gamma\mathbf{1} = \mathbf{a}$. Then, we can rewrite the unconstrained $\gamma$-subproblem as

$$
\begin{aligned}
\gamma_+ &= \underset{\gamma\in\mathbb{R}^{k\times k}}{\arg\min}\Big\{ -\log(\det\gamma\gamma^T) + \frac{\rho}{2}\|C^{1/2}(\gamma - A)\|_F^2\Big\} \\
&= \underset{\gamma\in\mathbb{R}^{k\times k}}{\arg\min}\Big\{ -\log(\det\gamma\gamma^T) + \frac{\rho}{2}\operatorname{tr}(\gamma^TC\gamma) \\
&\qquad\qquad - \rho\operatorname{tr}(\gamma^TCA)\Big\} \\
&= \underset{\gamma=UDV^T}{\arg\min}\Big\{ -\log(\det\gamma\gamma^T) + \frac{\rho}{2}\operatorname{tr}(\gamma^TC\gamma) \\
&\qquad\qquad - \rho\operatorname{tr}(\gamma^TCA)\Big\} \\
&= \underset{\gamma=UDV^T}{\arg\min}\Big\{ -\log(\det D^2) + \frac{\rho}{2}\operatorname{tr}(UD^2U^TC) \\
&\qquad\qquad - \rho\operatorname{tr}(UD_ADU^TC)\Big\} \\
&= \underset{\gamma=UDV^T}{\arg\min}\Big\{ -\sum_{i=1}^K 2\log|D_{ii}| + \frac{\rho}{2}\operatorname{tr}(ED^2) - \rho\operatorname{tr}(FD)\Big\} \\
&= \underset{\gamma=UDV^T}{\arg\min}\Big\{ -\sum_{i=1}^K 2\log|D_{ii}| + \frac{\rho}{2}E_{ii}D_{ii}^2 - \rho F_{ii}D_{ii}\Big\}
\end{aligned}
$$

where $E = U^TCU$ and $F = U^TCUD_A$. Then we can solve the above problem element by element. Looking at the $i$-th entry, we can take the derivative and set it to zero. That is

$$
\frac{\partial}{\partial D_{ii}}\Big(\log|D_{ii}| + \frac{\rho}{2}E_{ii}D_{ii}^2 - \rho F_{ii}D_{ii}\Big) = 0
$$

leading to the following quadratic formula

$$D_{ii}^2 - \frac{F_{ii}}{E_{ii}} D_{ii} - \frac{2}{\rho E_{ii}} = 0$$

which has the solution

$$\widehat{D}_{ii} = \frac{\frac{F_{ii}}{E_{ii}} + \sqrt{\frac{F_{ii}^2}{E_{ii}^2} + \frac{8}{\rho E_{ii}}}}{2}$$

Then, using these diagonal elements $\hat{D}_{ii}$, it follows that

$$\begin{aligned}\gamma_+ &= \underset{\gamma \in \mathbb{R}^{k \times k}}{\arg\min} \Big\{ -\log(\det \gamma\gamma^T) + \frac{\rho}{2} \|C^{1/2}(\gamma - A)\|_F^2 \Big\} \\ &= U \widehat{D} V^T\end{aligned}$$

We make the final adjustment to satisfy the linear constraint. Thus, the $\gamma$ update is

$$\gamma^{(t+1)} = \gamma_+ - (\gamma_+ \mathbf{1} - \mathbf{a})(\mathbf{1}^T C^{-1} \mathbf{1})^{-1} \mathbf{1}^T C^{-1}$$

#### 2.6.1.2 Proof of Proposition 2.3.2

*Proof.* The first order conditions of the updates in Algorithm 1 give us

$$\begin{aligned}0 &\in \partial \|\cdot\|_{h,\mu}(V_1^{t+1}) - \rho(\widetilde{W}\gamma^t - V_1^{t+1}) - \Lambda_1^t \\ 0 &\in \mathbb{1}_{G_R}(V_2^{t+1}) - \rho(\gamma^t - V_2^{t+1}) - \Lambda_2^t \\ 0 &\in -2(\gamma^{t+1})^{-T} + \rho \widetilde{W}^T(\widetilde{W}\gamma^{t+1} - V_1^{t+1}) + \widetilde{W}^T \Lambda_1^t + \\ &\quad \rho(\gamma^{t+1} - V_2^{t+1}) + \Lambda_2^t + \mathbf{1}_K(\nu^{t+1})^T \text{ s.t. } \gamma^{t+1}\mathbf{1} = \mathbf{a}\end{aligned} \tag{2.17}$$

Note that the first order condition for $\gamma^{t+1}$ is different as it is a equality constrained convex problem. Also, by the definitions of $\Lambda_1^{t+1}$ and $\Lambda_2^{t+1}$

$$\begin{aligned}\Lambda_1^{t+1} &= \Lambda_1^t + \rho(\widetilde{W}\gamma^{t+1} - V_1^{t+1}) \\ \Lambda_2^{t+1} &= \Lambda_2^t + \rho(\gamma^{t+1} - V_2^{t+1})\end{aligned} \tag{2.18}$$

Then, combining these two sets of equations, we have that

$$\begin{aligned}
&\Lambda_1^{t+1} + \rho\widetilde{W}(\gamma^t - \gamma^{t+1}) \in \partial\|\cdot\|_{h,\mu}(V_1^{t+1}) \\
&\Lambda_2^{t+1} + \rho(\gamma^t - \gamma^{t+1}) \in \partial\mathbb{1}_{G_R}(V_2^{t+1}) \\
&2(\gamma^{t+1})^{-T} - \mathbf{1}_K(\nu^{t+1})^T = \widetilde{W}^T\Lambda_1^{t+1} + \Lambda_2^{t+1} \\
&\frac{1}{\rho}(\Lambda_1^{t+1} - \Lambda_1^t) = \widetilde{W}\gamma^{t+1} - V_1^{t+1} \\
&\frac{1}{\rho}(\Lambda_2^{t+1} - \Lambda_2^t) = \gamma^{t+1} - V_2^{t+1}
\end{aligned} \tag{2.19}$$

Then, let us define $(\gamma^t, V_1^t, V_2^t, \Lambda_1^t, \Lambda_2^t)_{t=1}^{\infty}$ be a sequence of iterates with a limit point $(\gamma^*, V_1^*, V_2^*, \Lambda_1^*, \Lambda_2^*)$. Then, by the last two equations of (2.19), we have that $\widetilde{W}\gamma^* = \widetilde{W}V_2^* = V_1^*$. Therefore, the first two equations give us that

$$\begin{aligned}
\Lambda_1^* &\in \partial\|\cdot\|_{h,\mu}(V_1^*) = \partial\|\cdot\|_{h,\mu}(\widetilde{W}\gamma^*) \\
\Lambda_2^* &\in \partial\mathbb{1}_{G_R}(V_2^*) = \partial\mathbb{1}_{G_R}(\gamma^*)
\end{aligned}$$

Lastly, using the third equation in (2.19), it follows that

$$\begin{aligned}
&2(\gamma^*)^{-T} - \mathbf{1}_K(\nu^*)^T = \\
&\quad \widetilde{W}^T\Lambda_1^* + \Lambda_2^* \in \partial\|\cdot\|_{h,\mu}(\widetilde{W}\gamma^*) + \partial\mathbb{1}_{G_R}(\gamma^*)
\end{aligned}$$

Noting that the optimality condition for $\arg\min_\gamma -\log|\det(\gamma\gamma^T)| \quad \text{s.t.} \quad \gamma\mathbf{1} = \mathbf{a}$ is

$$-2(\gamma^*)^{-T} + \mathbf{1}_K(\nu^*)^T = 0 \quad \text{and } \gamma^*\mathbf{1} = \mathbf{a}$$

We have that

$$\begin{aligned}
\mathbf{0} &= -2(\gamma^*)^{-T} + \mathbf{1}_K(\nu^*)^T + 2(\gamma^*)^{-T} - \mathbf{1}_K(\nu^*)^T \\
&= -2(\gamma^*)^{-T} + \mathbf{1}_K(\nu^*)^T + \widetilde{W}^T\Lambda_1^* + \Lambda_2^* \in \partial f(\gamma^*)
\end{aligned}$$

and we have that $\gamma^*\mathbf{1} = \mathbf{a}$ by the formulation of our update for $\gamma^t$. This shows that $\gamma^*$ satisfies the optimality condition of (2.10) and thus a stationary point for $f$. □

### 2.6.2 Simulations

We demonstrate the computational benefit as well as the accuracy of our model in terms of perplexity. The experiments are based on the simulated data from the LDA model, and we focus

on the comparison to the variational EM (VEM) and Gibbs sampling to illustrate the advantages of our method. As part of the future work, we plan to compare the stochastic implementation of MVTM with GDM [Yurochkin and Nguyen, 2016] and the improved implementations of the Gibbs sampling presented in Li et al. [2014] and Yuan et al. [2015] at a much larger scale.

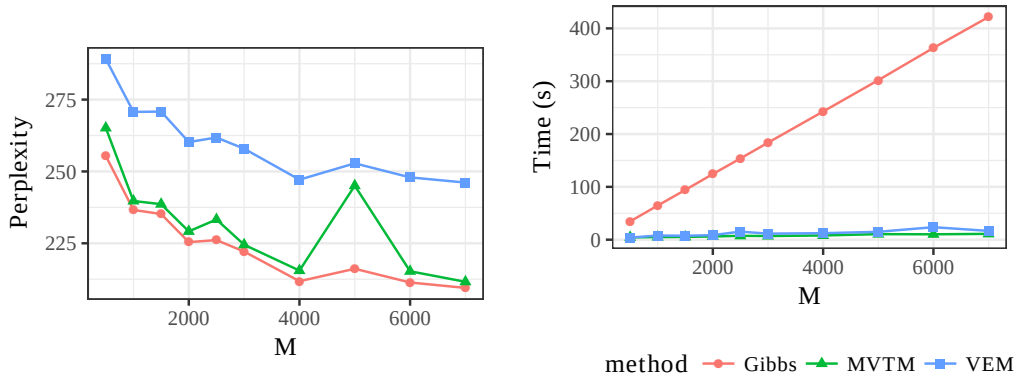



Figure 2.7: Perplexity of the held-out data and the corresponding time complexity of each method at varying values of the number of documents $M$ with $N_m = 1000$, $K = 5$, $V = 1200$, $\eta = 0.1$ and $\alpha = 0.1$

We first look at the behavior of the algorithms as $M$ increases when $N_m = 1000$ (Figure 2.8). At $N_m = 1000$, we are working with the setting that is close to the asymptotic regime, and MVTM has the computational speed comparable to VEM and the statistical performance similar to the Gibbs sampling.

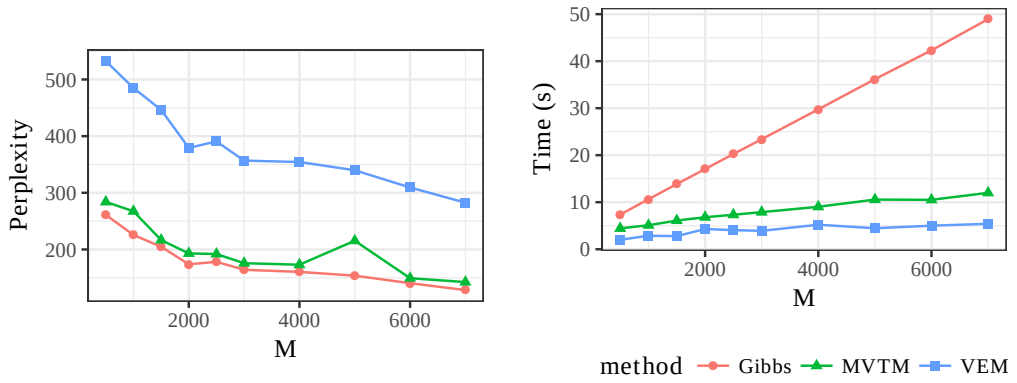



Figure 2.8: Perplexity of the held-out data and the corresponding time complexity of each method at varying values of the number of documents $M$ with $N_m = 100$, $K = 5$, $V = 1200$, $\eta = 0.1$ and $\alpha = 0.1$

In a more challenging case with the shorter documents at $N_m = 100$, MVTM continues to perform as well as the Gibbs sampling with a little additional computational cost. This performance

comparison would be of interest to the researchers who are working with shorter documents present in the modern application. As discussed in Tang et al. [2014] and Nguyen [2015], the limitation of LDA comes from the document lengths. Our results show that MVTM does not suffer from the short documents in terms of statistical performance when the regularization parameter $\mu$ for the hinge loss is appropriately chosen. The current batch implementation, however, suffers from the number of documents present in the dataset, as it has to soft-threshold every document. This computational limitation, however, can be alleviated by the stochastic implementation as demonstrated in the stochastic implementation of the variational method in Hoffman et al. [2013].

### 2.6.3 NIPS dataset Topics

#### 2.6.3.1 Computational Time

Figure 2.9 shows the time complexities of different algorithms on the NIPS dataset as we increase the number of topics. Compared to GDM, the proposed MVTM improvement on performance comes at a little computational cost. RecoverKL could achieve a similar computational speed if the anchor words are provided. However, when we include the computational cost of finding the anchor words, GDM and MVTM show computational advantages over RecoverKL.

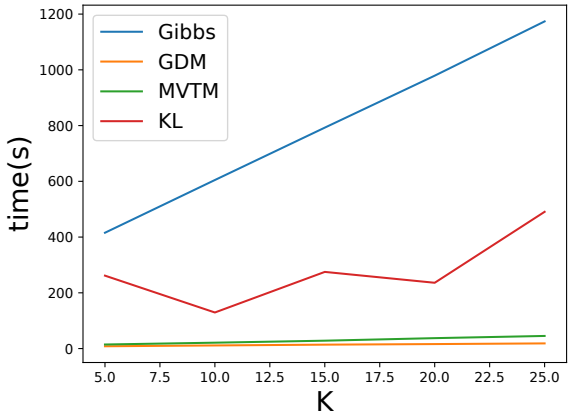



Figure 2.9: The computational performance of different algorithms as a function of the number of topics. NIPS dataset includes 1491 documents and 4492 unique words.

#### 2.6.3.2 Top 10 topics

| Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 | Topic 6 | Topic 7 | Topic 8 | Topic 9 | Topic 10 |
|---|---|---|---|---|---|---|---|---|---|
| neuron | input | training | training | algorithm | unit | model | network | function | learning |
| network | output | set | error | learning | network | data | neural | set | system |
| input | system | network | set | data | input | parameter | system | approximation | control |
| model | circuit | recognition | data | problem | hidden | distribution | problem | result | function |
| pattern | signal | data | cell | weight | weight | system | training | linear | action |
| neural | neural | algorithm | input | method | output | object | control | bound | algorithm |
| synaptic | network | vector | network | function | layer | gaussian | dynamic | number | task |
| learning | chip | learning | classifier | distribution | learning | likelihood | unit | point | reinforcement |
| cell | weight | classifier | weight | vector | pattern | cell | result | network | error |
| spike | analog | word | test | parameter | training | mixture | point | threshold | model |

Table 2.2: Top 10 MVTM topic for NIPS dataset

| Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 | Topic 6 | Topic 7 | Topic 8 | Topic 9 | Topic 10 |
|---|---|---|---|---|---|---|---|---|---|
| neuron | circuit | recognition | set | model | network | function | image | model | learning |
| cell | signal | speech | training | memory | input | algorithm | object | data | control |
| model | system | word | data | representation | unit | learning | images | distribution | system |
| input | neural | system | algorithm | node | weight | point | field | gaussian | action |
| activity | analog | training | error | rules | neural | vector | map | parameter | model |
| synaptic | chip | hmm | performance | tree | output | result | visual | mean | dynamic |
| pattern | output | character | classifier | structure | learning | case | motion | algorithm | policy |
| response | current | model | classification | level | training | problem | feature | probability | algorithm |
| firing | input | network | number | graph | layer | parameter | direction | method | reinforcement |
| cortex | neuron | context | learning | rule | hidden | equation | features | component | problem |

Table 2.3: Top 10 Gibbs topic for NIPS dataset

| Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 | Topic 6 | Topic 7 | Topic 8 | Topic 9 | Topic 10 |
|---|---|---|---|---|---|---|---|---|---|
| neuron | input | word | data | image | network | model | cell | learning | learning |
| network | output | speech | set | images | unit | data | visual | algorithm | control |
| spike | weight | recognition | training | object | neural | parameter | motion | function | model |
| synaptic | neural | system | error | point | weight | likelihood | direction | problem | system |
| input | network | training | function | features | hidden | mixture | response | action | task |
| pattern | net | character | vector | graph | training | distribution | orientation | policy | movement |
| firing | chip | hmm | method | representation | output | algorithm | neuron | optimal | controller |
| model | layer | speaker | classifier | feature | input | set | model | gradient | motor |
| activity | analog | context | kernel | information | error | gaussian | frequency | convergence | dynamic |
| neural | bit | network | gaussian | recognition | function | variables | field | step | reinforcement |

Table 2.4: Top 10 GDM topic for NIPS dataset

# CHAPTER 3

# Probabilistic Neuron Reconstruction for Brainbow Images

The automatic segmentation and reconstruction (tracing) of neurons in microscopic images pose many computational challenges. Traditional tracing algorithms rely on a set of seed points manually selected by researchers and/or adopt graph-based algorithms. As a result, these semi-supervised methods demand significant efforts from users and quickly become infeasible for large-scale neuroimages such as transgenic multicolor Brainbow images. Instead, the proposed method takes advantage of the fact that neurons have elongated curvilinear structures and thus strong geometric characteristics. We formulate these geometric properties by lifting the 2D images to $SE(2)$, i.e. the space of positions and orientations motivated by the geometry of the primal visual cortex $(V1)$ and contextual connections. We then impose a probabilistic model on the geometric image and reformulate the neuron tracing problem as a hidden Markov model to connect superpixels of neuronal processes.

## 3.1 Introduction

Developments of a transgenic, multicolor labeling strategy called *Brainbow* [Livet et al., 2007, Cai et al., 2013] have led to fast advancement in exploring the anatomy of individual neurons. Prior to the Brainbow technique, the classical approach to understanding how nerve systems work was restricted by the lack of tools to effectively reconstruct large number of neurons, as mapping neuronal structures requires a large imaging volume with high spatial resolution and the ability to differentiate the intermingled neuronal processes. Recent advancements, such as Gouwens et al. [2019], focus on a single cell characterization strategy to record morphological properties of neurons in the mouse visual cortex. These methods, however, need a vast amount of time and resources, as only one neuron can be imaged per subject. In order to overcome this disadvantage, the Brainbow technique uses stochastically expressed fluorescent protein mixtures to provide differential spectral

labeling in neighboring cells. Therefore, densely labeled neurons can be distinguished in the same field of view (Figure 3.1).

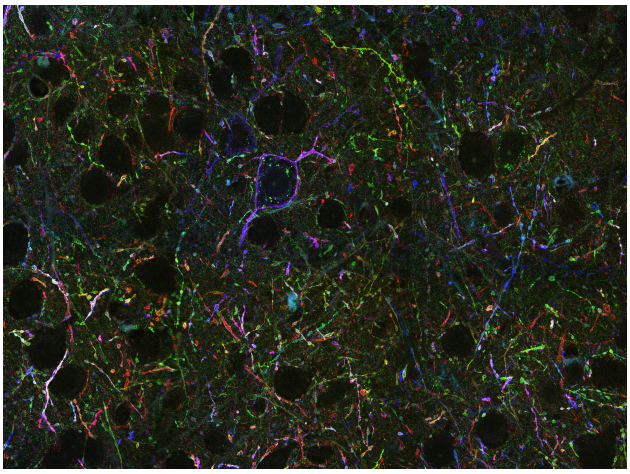


Figure 3.1: An example 3600 x 2400 x 500 3D Brainbow image from a section of a transgenic mouse brain, where several PV-expressing neurons and their respective synapses have been labeled fluorescently. Samples were imaged using the miriEx expansion microscopy method Shen et al. [2020].

In contrast to the technical developments in acquiring Brainbow images, the advancements in quantitative analysis of Brainbow images have been relatively slow due to the computational complexity of extracting the neuronal trees from large-scale Brainbow images. The available algorithms for segmentation of neurons in Brainbow images often require manual annotation and analysis of the neuronal tree by human experts [Wang et al., 2019, Roossien et al., 2019]. However, these algorithms become quickly infeasible due to the labor-intensive labeling of such images. Furthermore, in order to accommodate terabytes of Brainbow images; a suitable algorithm has to also be scalable while preserving the global structures of the neurons. For example, Figure 3.1 shows a fraction of a transgenic mouse brain. This $3600 \times 2400 \times 500$ microscopy image is about 20GB in size, thus recent advancements in neural networks cannot be applied to the whole dataset directly.

In order to implement our method for neuron tracing, we revisit the idea developed in Athey et al. [2021] that adopts a probabilistic reconstruction method by incorporating a hidden Markov state process with a random field appearance model based on neuron geometry. The method in Athey et al. [2021], ViterBrain, inputs the start and endpoint of the neuron of interest and builds the most probable neuron path. The geometric features in Viterbrain are estimated by fitting B-spline to each fragment, or superpixels, to enhance the probability of connecting different neuronal fragments. However, as Viterbrain by Athey et al. [2021] focuses on fluorescent images with a single spectrum, it is not able to simultaneously track different neurites of a given neuronal process.

Our method adopts the probabilistic viewpoint of Viterbrain but aims to integrate the information from the theory of orientation scores [Duits et al., 2007b] and multi-spectral information from Brainbow images. This additional information allows us to only require a single input from the user to trace a neuronal process.

The theory of orientation scores was developed in Duits et al. [2007a] that is motivated by the local and global laws to describe the properties of visual stimuli. The theory of orientation scores serves as a bridge to mimicking the properties in the visual cortex by characterizing image features in terms of location and orientation. In order to mathematically model this information, Bekkers et al. [2014] proposed to effectively calculate the orientations/orientation scores of each pixel by lifting the 2D images to the space of positions and orientations $SE(2) = \mathbb{R}^2 \times S^1$, which is referred to as the *Special Euclidean group* or the *Euclidean motion group*.

More recently, Favali et al. [2016] proposed a framework to analyze vessel connectivities in retinal images by incorporating orientation scores into spectral clustering. Their method applies spectral clustering on data in $SE(2)$, by constructing the similarity matrix and performing eigendecomposition to solve the clustering problem. As well established in the spectral clustering literature, such eigendecomposition-based methods experience significant disadvantages in terms of computation. In fact, Abbasi-Sureshjani et al. [2017] and Favali et al. [2016] have limited demonstration of their methods due to the construction of the affinity matrix.

### 3.1.1 Related Works

**Theory of orientation scores:** Based on the cortical orientation columns in the primary visual cortex introduced in Hubel and Wiesel [1959], Duits et al. [2007a] developed a mathematical framework to lift 2D images into the Euclidean motion group $SE(2)$. The space of positions and orientations mathematically formulate the perceptual organization of orientation in the visual cortex. By adding a third dimension with the orientation score transformation, curvilinear structures in 2D images are disentangled into different orientation planes according to their local orientations. A practical approach to efficiently calculating orientations for medical images was developed in Bekkers et al. [2014], while a framework to enhance elongated structures in the domain of an orientation score by developing a rotating frame on $SE(2)$ was developed in Zhang et al. [2016]. For a detailed description of these methods, refer to Zhang et al. [2016] and references therein.

**Neuron Tracing**: As a part of an effort to automatically reconstruct neuron morphology, researchers in various fields developed algorithms for tracing neurons in fluorescent images. Previous works focused on single-spectral images where the algorithm is faced with spatial and intensity information. Early works focused on graph-based methods using shortest path computation based on pre-defined seed points [Peng et al., 2010, Wang et al., 2011, Turetken et al., 2013]. More recent

work used Bayesian estimation [Radojević and Meijering, 2017] and deep learning methods [Li et al., 2017, Zhou et al., 2018, Friedmann et al., 2020].

### 3.1.2 Notation

We define the function $f : \mathbb{R}^2 \to \mathbb{R}^d$ as a 2D image with $d$ spectral information, and $U_f$ as a lifted image that has additional information about the orientation score constructed from image $f$. $SE(2) = \mathbb{R}^2 \times S^1$ refers to the Euclidean motion group of planar rotations and translations with $g = (\boldsymbol{x}, \theta)$ its group elements for $S^1$ the 1-sphere. Furthermore, $SE(2)$ can be identified as $SE(2) = \mathbb{R}^2 \times SO(2)$, for $SO(2)$ the special orthogonal group with elements $\mathbf{R}_\theta \in \mathbb{R}^{2\times 2}$. These elements are the counter-clockwise rotations matrix over the corresponding angle $\theta$. The tuple $\{\partial_x, \partial_y\}$ corresponds to a global horizontal and vertical frame, and $\{\partial_\zeta, \partial_\eta, \partial_\theta\}$ is a left-invariant rotating derivative frame of reference. We denote $\psi$ as an anisotropic wavelet kernel used to calculate $U_f$.

## 3.2 Preliminaries

Previous works in tracing curvilinear structures for images with light microscopy focused on working in Euclidean space with spectral information. However, as shown in Bekkers et al. [2014], spatial information is not sufficient to separate bifurcation and overlapping regions. Based on the theory of orientation scores, developed in Duits [2005], we lift the observed Brainbow image to a Euclidean motion group and perform tracing in the lifted space.

As first introduced in Duits et al. [2007a], we consider the orientation score as a square-integrable function; with is the Euclidean motion group as its domain, i.e. $U \in \mathbb{L}_2(SE(2))$. The Euclidean motion group $SE(2)$ is the group of all rotations and translations, whose elements $g = (\boldsymbol{x}, \theta)$ are composed of **1)** the position $\boldsymbol{x} = (x, y) \in \mathbb{R}^2$ in the domain of the image $f$ and **2)** the orientation angle $\theta \bmod 2\pi$ which captures the orientation of the structures in image $f$. For the rest of this paper, we use both the short notation $g$ and the explicit notation $(\boldsymbol{x}, \theta)$ for the group elements.

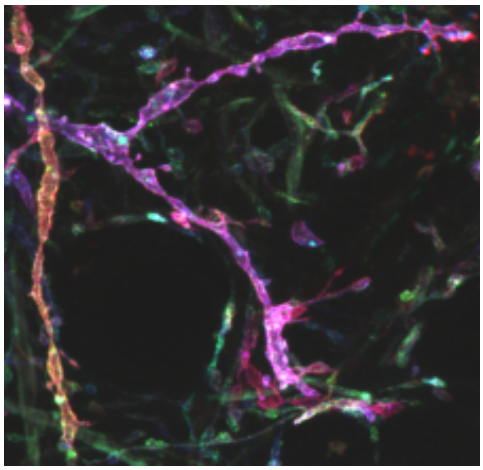
We obtain an orientation score $U_f$ from an image $f \in \mathbb{L}_2(\mathbb{R}^2)$ based on convolution with cake wavelets from Duits et al. [2007a], where $\mathbb{L}_2(\mathbb{R}^2)$ is the space of square integrable functions on $\mathbb{R}^2$. The additional orientation dimension encodes information on local orientations in the image. We define $U_f$ for the grey-scale image below.

**Definition 3.2.1.** Consider a 2D image of $f$ as a function $f : \mathbb{R}^2 \to \mathbb{R}$ with compact support on the image domain $\Omega = [0, X] \times [0, Y]$, where $X, Y \in \mathbb{Z}$ are the image dimensions. For a square

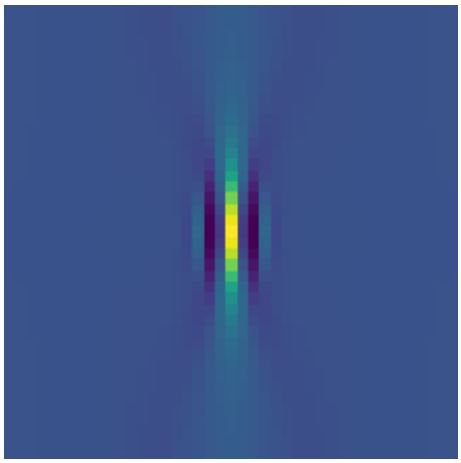
integrable $f$, the orientation score $U_f : SE(2) \to \mathbb{C}$

$$U_f(\boldsymbol{x}, \theta) = \int_{\mathbb{R}^2} \overline{\psi(\boldsymbol{R}_\theta^{-1}(\tilde{x} - \boldsymbol{x}))} \cdot f(\tilde{x})\, d\tilde{x}$$

where $\psi$ is an anisotropic wavelet with orientation $\theta = 0$ and $\boldsymbol{R}_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$.

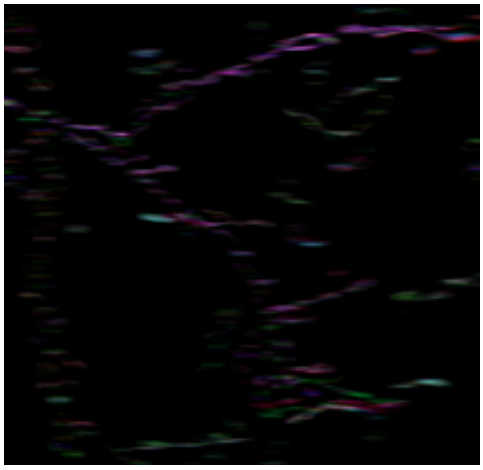


(a) Brainbow Image



(b) Cake Wavelet

(c) Orientation Scores with a vertical wavelet



(d) Orientation Scores with a horizontal wavelet

Figure 3.2: Example of the orientation score calculation for Brainbow. Given the observed Brainbow image (a), orientation scores based on cake wavelet is able to recover spectral information aligned with a vertical wavelet in (c) and a horizontal wavelet in (d).

In other words, $U_f$ is obtained by convolving the image with a rotating anisotropic convolution kernel $\psi$ and measures the alignment of the rotated images with the vertical axis in the 2D space.

For some choices of $\psi$, including the cake wavelets and Gabor wavelets, there exists a stable inverse transformation obtained by $f = \int_0^{2\pi} U(\boldsymbol{x}, \theta) d\theta$ [Duits, 2005].

Thus, given the input images, we assign the orientation for each pixel $\boldsymbol{x}_i$ by

$$\theta_i = \underset{\theta \in [0,\pi]}{\arg\max}\, Re(U_f(\boldsymbol{x}, \theta)). \tag{3.1}$$

## 3.3 Methods

Our method focuses on minimizing the input from users to segment and trace neurons present in Brainbow images. Based on the theory of orientation scores described in the previous section, we calculate the orientation information for each pixel and perform segmentation based on the derived geometric information. Then, we adopt the hidden Markov model framework from Athey et al. [2021] to trace neurons from a single input.

### 3.3.1 Geometric Segmentation based on Left Invariant Derivatives (LID)

Based on the OS calculation, the geometric features of 2D images can be extracted and further developed to enhance the curvilinear structures of the neuron. The enhancement of the neurites in Brainbow images allows us to segment the neurons from the background by thresholding the geometric features.

We adjust the horizontal and vertical axes by incorporating the orientation information. A rotating frame, called Left invariant derivatives (LID), defined by $\{\partial_\zeta, \partial_\eta, \partial_\theta\}$ allows us to work with translations over $x$ and rotations over $\theta$ [Zhang et al., 2016]. Intuitively, the adjusted axis $\partial\zeta$ is aligned with the direction of the image at $\theta$, and the magnitude $\partial_\zeta^2$ shows the magnitude of spectral expression in the direction of $\theta$.

Based on the LID-based filters developed in Zhang et al. [2016], the second-order operator $\Phi_\zeta^{\sigma_s,\sigma_o}(U_f) = \partial_\zeta^2 G_{\sigma_s,\sigma_o} * U_f$ is used as LID filter for neuron enhancement, where * is a convolution operator. Here, $\sigma_s$ and $\sigma_o$ correspond to the spatial and orientation standard deviation for the Gaussian derivative $G_{\sigma_s,\sigma_o}$. The final image enhancement from multi-scale filtered orientation scores is obtained through

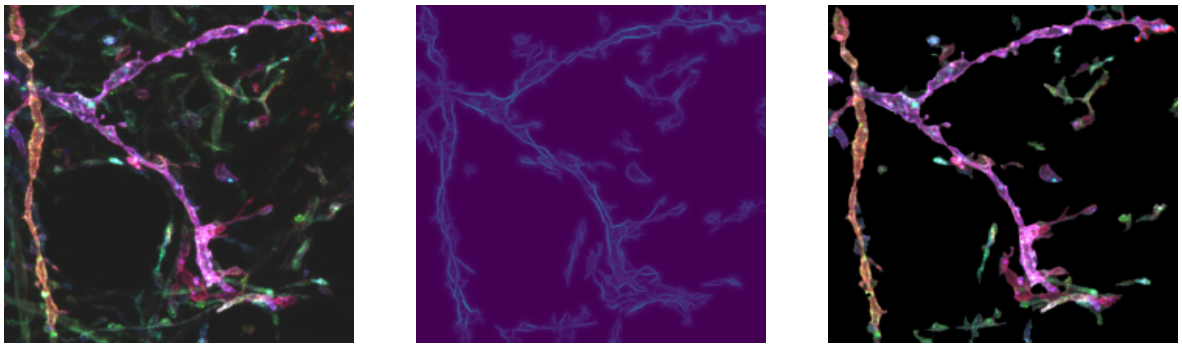
$$\Gamma(f)(\boldsymbol{x}) = \max_{\theta \in \frac{\pi}{N_o}\{1,\ldots,N_o\}} \left\{ \sum_{\sigma_s \in \mathcal{S}} \Phi_\eta^{\sigma_s,\sigma_o}(U_f)(\boldsymbol{x}, \theta_i) \right\} \tag{3.2}$$

where $N_o$ is the number of orientation and $S$ is the spatial scaling.

By applying the LID filter directly to orientation scores, the responses on neuronal processes are enhanced in each orientation layer. As the elongated structures (neurites) are disentangled

into different orientation layers separately, the LID filter from Zhang et al. [2016] is capable of enhancing and preserving complex crossings of different neurites.

After applying the LID filter to the orientation scores of a Brainbow image, the neuronal processes will be enhanced with high filter responses and the background will be suppressed with low responses. By rescaling the enhanced image between 0 and 1, the final segmented image is obtained with a suitable threshold value.



(a) Brainbow Image

(b) The enhancement of the Brainbow image based on LID filter.

(c) Segmented Brainbow image after the LID filter is applied

Figure 3.3: The segmented result based on the LID filter.

### 3.3.2 Neuron Tracing

The neurons are modeled as connected curves in $\mathbb{R}^2$ as a function of arclength as introduced in Athey et al. [2021]. That is, the neurons are denoted as $c(\cdot) := \{c(\ell), \ell \geq 0\}$ over the pixel lattice $D = \cup_{i \in \mathcal{Z}^{m^2}} \Delta y_i \subset \mathbb{R}^2$ where $y_i \in \Delta y_i$ is the center of the pixel. Following the notations of Athey et al. [2021], the image is modeled as a random field $\{I_{y_i}, \Delta y_i \in D\}$ with the probability

$$P(I_Y | c(\ell), \ell \geq 0) = \Pi_{y \in Y} p(I_y | c(\ell), \ell \geq 0). \tag{3.3}$$

In other words, the elements of the observed image are conditionally independent given the underlying neuronal process. As the number of pixels in Brainbow images is prohibitive for any computational methods, each pixel is clustered into superpixels based on the linear spectral clustering [Chen et al., 2017].

### 3.3.3 Hidden Markov Model

In order to describe the hidden Markov model, we need to define a state for each superpixel. The state of a given superpixel includes spatial, spectral, and orientation information. That is, for a superpixel $v_i$, the corresponding state is defined as $s_i = (\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\rho})$, where $\boldsymbol{x} \in \mathbb{R}^2$ is the spatial information, $\theta \in [0, \pi]^p$ is the orientation calculated from orientation scores, and $\boldsymbol{\rho} \in \mathbb{R}^p$ is the average spectral information of the superpixel $v_i$, where $p$ is the number of spectral channels for the given Brainbow image.

The collection of states $\mathcal{S}$ is the finite space of the HMM, and the tracing problem can be reformulated as estimating the state sequence $(s_1, \ldots, s_n)$ that corresponds to a given neuronal process. In order to simplify the probabilistic model on the observed image $I_D$, we **1)** assume the first order Markov property ($p(s_i|s_{i-1}, s_{1:i-2}) = p(s_i|s_{i-1})$) and **2)** model the transition probability as a Boltzmann distribution with energy $U$ is imposed

$$p(s_i|s_{i-1}) = \frac{e^{-U(s_{i-1},s_i)}}{Z(s_{i-1})}$$

where $Z(s_{i-1}) = \sum_{s_i \in \mathcal{S}} e^{-U(s_{i-1},s_i)}$ and

$$U(s_{i-1}, s_i) = \alpha_d \|\boldsymbol{x}_i - \boldsymbol{x}_{i-1}\|_2^2 + \alpha_\theta \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\|_1 + \alpha_\rho \|\boldsymbol{\rho}_i - \boldsymbol{\rho}_{i-1}\|_2^2$$

These are simple assumptions on the probabilistic model that was used in Athey et al. [2021] for analyzing neurons. Compared to the B-spline fitting for orientation calculations in Athey et al. [2021], we use orientation scores for calculating the orientation of the superpixel. More importantly, the multi-spectral information from the Brainbow images allows us to incorporate the spectral difference and trace out the entire neuron from a single input.

Based on this information, we have the following joint probability for the observed image

$$p(s_{1:n}, I_D) = \prod_{i=2}^{n} \left( \prod_{j \in N_e(s_i)} p(s_j|s_i) \right) p(s_1, I_D) \tag{3.4}$$

where $N_e(s_i)$ is the set of adjacent superpixel of $v_i$.

As shown in Athey et al. [2021], taking the log of the probability (3.4) leads to a sequentially additive cost function that can be solved with a breath-first search with edge weights given by $e(s_{i-1}, s_i) = -\log p(s_i|s_{i-1})$.

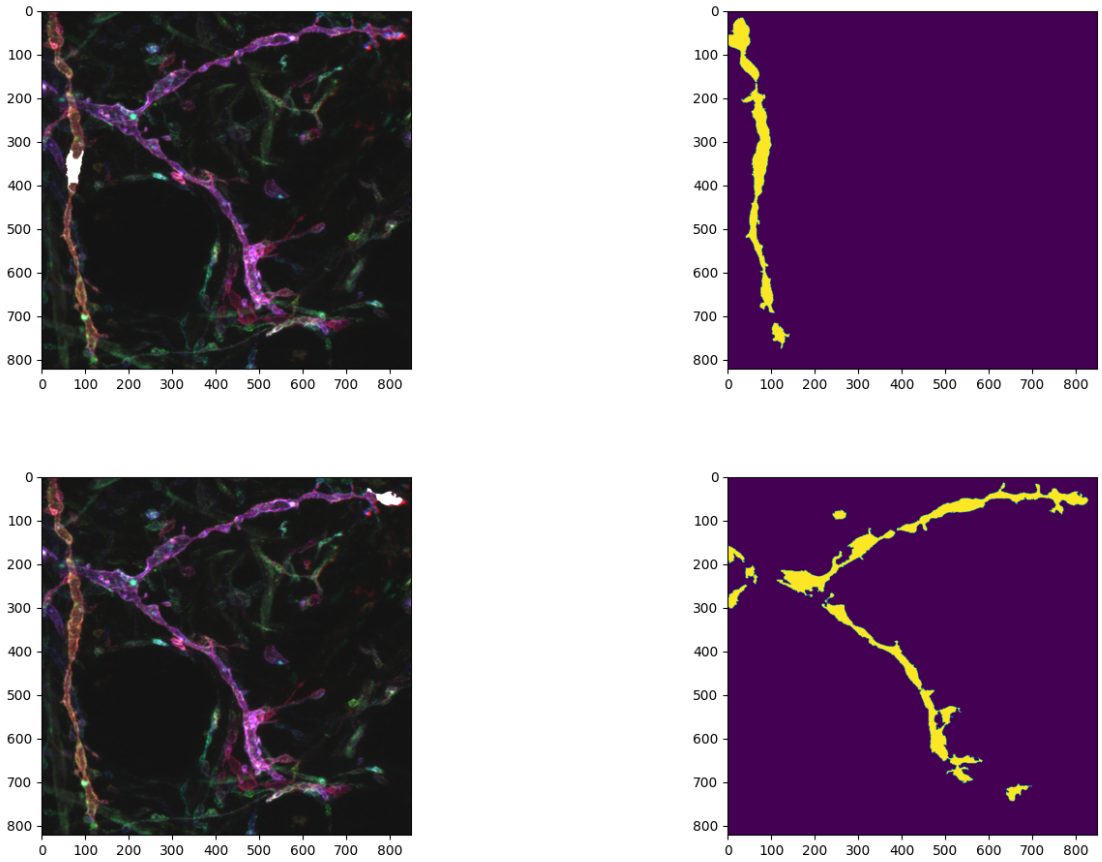



Figure 3.4: Tracing results of the Brainbow image. There are two neuronal processes of interest that have the strongest signals in the image. Each of the neuron is traced based on the region of interest (white labelled superpixel) from the user on shown on the left column.

## 3.4 Results

Figure 3.4 shows the traced neurons based on a single input with individual neurites from the initial input. Compared to the existing algorithm, where the traced neuron is based on tree structures, we are able to estimate the volumetric segmentation of the neuron of interest. Furthermore, we are able to automatically segment the neurons from the background without any pre-defined labels. Despite the missing superpixel in the segmented image, we can impute the missing superpixel based on the adjacency graph and the resulting sequencey by the convex combination of the adjacent states with $\hat{F}_i = \lambda F_i + (1 - \lambda) F_{i+1}$ for $\lambda \in [0, 1]$.

## 3.5 Discussion

Segmenting individual curvilinear structures in neuroimaging remains one of the biggest challenges for neuroscience. Existing methods either require a significant amount of inputs from the user or rely on prohibitive computation to trace neurons in the entire brain. The proposed method extended the tracing algorithms from Athey et al. [2021] by using the theory of orientation scores to borrow information from the underlying geometric structure of the latent neuronal process.

This work serves as a building block for parallelizing the tracing problem in Brainbow images. As we use the graphical relationships among superpixels in the tracing step, we can incorporate conditionally independent assumptions of Brainbow image sections and incorporate the decomposable adjacency matrix from Wiesel and Hero [2009]. While Wiesel and Hero [2009] focused on decomposing the adjacency graph for principal component analysis, we can use their decomposable structure to parallelize the breadth-first search algorithm to estimate the global sequence of superpixel that corresponds to a specific input. Additionally, one could calculate the curvature for each of the neuronal superpixels by using the locally adaptive frame (LAD) filter from Zhang et al. [2016].

Extension to 3D tracing can be easily done with the additional spatial information in our framework. However, the orientation score used in our method is limited to 2D orientations. It would be interesting to compare the tracing results with 3D orientation scores from Janssen et al. [2018] to understand the tradeoff between the computational burden and the additional orientational information when tracing 3D objects.

Lastly, multi-spectral Brainbow images have similar characteristics as the application in hyperspectral satellite images. However, as the number of latent variables (or neurons) is larger than the color spectra in Brainbow images, the spectral unmixing algorithms are not yet applicable in our problem. The need for spectral unmixing algorithms from hyperspectral imaging literature will come into play as the number of fluorescent dyes in Brainbow images increases in the future.

# CHAPTER 4

# Kronecker Sum Structures in Covariance and Precision Matrices

This paper introduces *the Sylvester graphical lasso* (SyGlasso) that captures multiway dependencies present in tensor-valued data. The model is based on the Sylvester equation that defines a generative model. The proposed model complements the tensor graphical lasso [Greenewald et al., 2019] that imposes a Kronecker sum model for the inverse covariance matrix by providing an alternative Kronecker sum model that is generative and interpretable. A nodewise regression approach is adopted for estimating the conditional independence relationships among variables. The statistical convergence of the method is established, and empirical studies are provided to demonstrate the recovery of meaningful conditional dependency graphs. We apply the SyGlasso to an electroencephalography (EEG) study to compare the brain connectivity of alcoholic and nonalcoholic subjects. We demonstrate that our model can simultaneously estimate both brain connectivity and its temporal dependencies.

## 4.1 Introduction

Estimating conditional independence patterns of multivariate data has long been a topic of interest for statisticians. In the past decade, researchers have focused on imposing sparsity on the precision matrix (inverse covariance matrix) to develop efficient estimators in the high-dimensional statistics regime where $n \ll p$. The success of the $\ell_1$-penalized method for estimating multivariate dependencies was demonstrated in Meinshausen and Bühlmann [2006] and Friedman et al. [2008b] for the multivariate setting. This has naturally led researchers to generalize these methods to multiway tensor-valued data. Such generalizations are of benefit for many applications. Examples of applications include the estimation of brain connectivity in neuroscience, reconstruction of molecular networks, and detecting anomalies in social networks over time.

The first generalizations of multivariate analysis to the tensor-variate settings were presented by Dawid [1981], where the matrix-variate (a.k.a. two-dimensional tensor) distribution was first

introduced to model the dependency structures among both rows and columns. Dawid [1981] extended the multivariate setting by rewriting the tensor-variate data as a vectorized (vec) representation of the tensor samples $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{m_1 \times \cdots \times m_k}$ and analyzing the overall precision matrix $\boldsymbol{\Omega} = \mathbb{E}\big(\text{vec}(\boldsymbol{\mathcal{X}})\text{vec}(\boldsymbol{\mathcal{X}})^T\big) \in \mathbb{R}^{m \times m}$, where $m = \prod_{k=1}^{K} m_k$. Even for a two-dimensional tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{m_1 \times m_2}$, the computation complexity and sample complexity is high since the number of parameters in the precision matrix grows quadratically as $\prod_{k=1}^{K} m_k^2$. Therefore, in the regime of tensor-variate data, unstructured precision matrix estimation has posed challenges due to the large number of samples needed for accurate structure recovery.

To address the sample complexity challenges, sparsity can be imposed on the precision matrix $\boldsymbol{\Omega}$ by using a sparse Kronecker product (KP) or Kronecker sum (KS) decompositions of $\boldsymbol{\Omega}$. The earliest and most popular form of sparse structured precision matrix estimation represents $\boldsymbol{\Omega}$ as the Kronecker product of smaller precision matrices. Tsiligkaridis et al. [2013] and Zhou [2014] proposed to model the precision matrix as a sparse Kronecker product of the covariance matrices along each mode of the tensor in the form $\boldsymbol{\Omega} = \boldsymbol{\Psi}_1 \otimes \cdots \otimes \boldsymbol{\Psi}_K$. The KP structure on the precision matrix has the nice property that the corresponding covariance matrix is also a KP. Zhou [2014] provides a theoretical framework for estimating the $\boldsymbol{\Omega}$ under KP structure and showed that the precision matrices can be estimated from a single instance under the matrix-variate normal distribution. Recently, Lyu et al. [2019] extended the KP structured model to tensor-valued data and studied its theoretical properties. An alternative, called the Bigraphical Lasso, was proposed by Kalaitzis et al. [2013] to model conditional dependency structures of precision matrices by using a Kronecker sum representation $\boldsymbol{\Omega} = \boldsymbol{\Psi}_1 \oplus \boldsymbol{\Psi}_2 = (\boldsymbol{\Psi}_1 \otimes \mathbf{I}) + (\mathbf{I} \otimes \boldsymbol{\Psi}_2)$. Recently, Greenewald et al. [2019] generalized the Kronecker sum (KS) structure to the multiway tensor valued data, called the TeraLasso. As shown in Greenewald et al. [2019], compared to the KP model, KS structure on the precision matrix leads to a non-separable covariance matrix that provides a richer model than the KP structure.

**KP vs KS:** The Kronecker structures can be characterized by the product graphs of the individual components [Greenewald et al., 2019]. The KP method admits a simple stochastic representation as $\text{vec}\,(\mathbf{X}) = (\boldsymbol{\Psi}_1 \otimes \boldsymbol{\Psi}_2)^{1/2} Z$, where $Z$ is a vector of i.i.d. Gaussian $\mathcal{N}(0,1)$ and $\mathbf{A}^{1/2}$ denotes the square root matrix of $\mathbf{A}$. Kalaitzis et al. [2013] first motivated the KS structure on the precision matrix by relating the Kronecker sum $(\boldsymbol{\Psi}_1 \oplus \cdots \oplus \boldsymbol{\Psi}_K)$ to the associated Cartesian product graph. Thus, the overall structure of $\boldsymbol{\Omega}$ naturally leads to an interpretable model that brings the individual components together. The Kronecker product, however, corresponds to the direct tensor product of the individual components $\boldsymbol{\Psi}_k$ and leads to a denser dependency structure in the precision matrix. In related work, Rudelson and Zhou [2017] and Park et al. [2017] studied the Kronecker sum structure on the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Omega}^{-1} = \mathbf{A} \oplus \mathbf{B}$ which corresponds to an errors-in-variables model. Unlike the KP model, the KS model does not have a simple stochastic

representation.

**The Sylvester Graphical Lasso (SyGlasso):** We propose a *Sylvester structured graphical model* to estimate precision matrices associated with tensor data. Similar to the KP- and KS-structured graphical models, we simultaneously learn $K$ graphs along each mode of the tensor data. However, instead of a Kronecker sum model for the precision matrix, as used in KS models, the Sylvester structured graphical model uses a Kronecker sum model for the square root factor of the precision matrix. The model is estimated by joint sparse regression models that impose sparsity of the individual sparse components $\mathbf{\Psi}_k$ for $k = 1, \ldots, K$. The Sylvester model reduces to a squared Kronecker sum representation for the precision matrix $\mathbf{\Omega} = (\mathbf{\Psi}_1 \oplus \cdots \oplus \mathbf{\Psi}_K)^2$, which is motivated by a stochastic representation of multivariate data with such a precision matrix.

## Notations

We adopt the notations used by Kolda and Bader [2009]. A $K$-th order tensor is denoted by boldface Euler script letters, e.g, $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{m_1 \times \cdots \times m_K}$. $\boldsymbol{\mathcal{X}}$ reduces down to a vector for $K = 1$ and to a matrix for $K = 2$. The $(i_1, \ldots, i_K)$-th element of $\boldsymbol{\mathcal{X}}$ is denoted by $\boldsymbol{\mathcal{X}}_{i_1,\ldots,i_K}$, and we define the vectorization of $\boldsymbol{\mathcal{X}}$ to be $\text{vec}(\boldsymbol{\mathcal{X}}) := (\boldsymbol{\mathcal{X}}_{1,1,\ldots,1}, \boldsymbol{\mathcal{X}}_{2,1,\ldots,1}, \ldots, \boldsymbol{\mathcal{X}}_{m_1,1,\ldots,1}, \boldsymbol{\mathcal{X}}_{1,2,\ldots,1}, \ldots, \boldsymbol{\mathcal{X}}_{m_1,m_2,\ldots,m_k})^T \in \mathbb{R}^m$ with $m = \prod_{k=1}^K m_k$.

There are several tensor algebra concepts that we recall. A fiber is the higher order analogue of the row and column of matrices. It is obtained by fixing all but one of the indices of the tensor, e.g., the mode-$k$ fiber of $\boldsymbol{\mathcal{X}}$ is $\boldsymbol{\mathcal{X}}_{i_1,\ldots,i_{k-1},:,i_{k+1},\ldots,i_K}$. Matricization, also known as unfolding, is the process of transforming a tensor into a matrix. The mode-$k$ matricization of a tensor $\boldsymbol{\mathcal{X}}$, denoted by $\boldsymbol{\mathcal{X}}_{(k)}$, arranges the mode-$k$ fibers to be the columns of the resulting matrix. It is possible to multiply a tensor by a matrix – the $k$-mode product of a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{m_1 \times \cdots \times m_K}$ and a matrix $\mathbf{A} \in \mathbb{R}^{J \times m_k}$, denoted as $\boldsymbol{\mathcal{X}} \times_k \mathbf{A}$, is of size $m_1 \times \cdots \times m_{k-1} \times J \times m_{k+1} \times \ldots m_k$. Its entry is defined as $(\boldsymbol{\mathcal{X}} \times_k \mathbf{A})_{i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K} := \sum_{i_k=1}^{m_k} \boldsymbol{\mathcal{X}}_{i_1,\ldots,i_K} A_{j,i_k}$. In addition, for a list of matrices $\{\mathbf{A}_1, \ldots, \mathbf{A}_K\}$ with $\mathbf{A}_k \in \mathbb{R}^{m_k \times m_k}$, $k = 1, \ldots, K$, we define $\boldsymbol{\mathcal{X}} \times \{\mathbf{A}_1, \ldots, \mathbf{A}_K\} := \boldsymbol{\mathcal{X}} \times_1 \mathbf{A}_1 \times_2 \cdots \times_K \mathbf{A}_K$. Lastly, we define the $K$-way Kronecker product as $\bigotimes_{k=1}^K \mathbf{\Psi}_k = \mathbf{\Psi}_1 \otimes \cdots \otimes \mathbf{\Psi}_K$, and the equivalent notation for the Kronecker sum as $\bigoplus_{k=1}^K \mathbf{\Psi}_k = \mathbf{\Psi}_1 \oplus \cdots \oplus \mathbf{\Psi}_K = \sum_{k=1}^K \mathbf{I}_{[d_{1:k-1}]} \otimes \mathbf{\Psi}_k \otimes \mathbf{I}_{[d_{k+1:K}]}$, where $\mathbf{I}_{[d_{k:\ell}]} = \mathbf{I}_{d_k} \otimes \cdots \otimes \mathbf{I}_{d_\ell}$.

# 4.2 Sylvester Graphical Lasso

Let a random tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{m_1 \times \cdots \times m_K}$ be generated by the following representation:

$$\boldsymbol{\mathcal{X}} \times_1 \mathbf{\Psi}_1 + \cdots + \boldsymbol{\mathcal{X}} \times_K \mathbf{\Psi}_K = \boldsymbol{\mathcal{T}}, \tag{4.1}$$

where $\mathbf{\Psi}_k \in \mathbb{R}^{m_k \times m_k}, k = 1, \ldots, K$ are sparse symmetric positive definite matrices and $\boldsymbol{\mathcal{T}}$ is a random tensor of the same order as $\boldsymbol{\mathcal{X}}$. Equation (4.1) is known as the Sylvester tensor equation. The equation often arises in finite difference discretization of linear partial equations in high dimension [Bai et al., 2003] and discretization of separable PDEs [Kressner and Tobler, 2010, Grasedyck, 2004]. When $K = 2$ it reduces to the Sylvester matrix equation $\mathbf{\Psi_1 X} + \mathbf{X}{\mathbf{\Psi_2}}^T = \mathbf{T}$ which has wide application in control theory, signal processing and system identification (see, for example Golub et al. [1979] and references therein).

It is not difficult to verify that the Sylvester representation (4.1) is equivalent to the following system of linear equations:

$$\left(\bigoplus_{k=1}^{K} \mathbf{\Psi}_k\right) \operatorname{vec}(\boldsymbol{\mathcal{X}}) = \operatorname{vec}(\boldsymbol{\mathcal{T}}), \tag{4.2}$$

If $\boldsymbol{\mathcal{T}}$ is a random tensor such that $\operatorname{vec}(\boldsymbol{\mathcal{T}})$ has zero mean and identity covariance, it follows from (4.2) that any $\boldsymbol{\mathcal{X}}$ generated from the stochastic relation (4.1) satisfies $\mathbb{E}\operatorname{vec}(\boldsymbol{\mathcal{X}}) = \mathbf{0}$ and $\mathbf{\Sigma} = \mathbf{\Omega}^{-1} := \mathbb{E}\operatorname{vec}(\boldsymbol{\mathcal{X}})\operatorname{vec}(\boldsymbol{\mathcal{X}})^T = \left(\bigoplus_{k=1}^{K} \mathbf{\Psi}_k\right)^{-2}$. In particular, when $\operatorname{vec}(\boldsymbol{\mathcal{T}}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$, we have that $\operatorname{vec}(\boldsymbol{\mathcal{X}}) \sim \mathcal{N}\left(\mathbf{0}, \left(\bigoplus_{k=1}^{K} \mathbf{\Psi}_k\right)^{-2}\right)$.

This paper proposes a procedure for estimating $\mathbf{\Omega}$ with $N$ independent copies of the tensor data $\{\boldsymbol{\mathcal{X}}^i\}_{i=1}^N$ that are generated from (4.1). For the rest of the paper, we assume that the last mode of the data tensor corresponds to the observations mode. For example, when $K = 2$, $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{m_1 \times m_2 \times N}$ is the matrix-variate data with $N$ observations. Our goal is to estimate the $K$ precision matrices $\{\mathbf{\Psi_k}\}_{k=1}^K$ each of which describes the conditional independence of $k$-th data dimension. The resulting precision matrix is $\mathbf{\Omega} = \left(\bigoplus_{k=1}^{K} \mathbf{\Psi}_k\right)^2$. By rewriting (4.2) element-wise, we first observe that

$$\begin{aligned} &\left(\sum_{k=1}^{K} (\mathbf{\Psi}_k)_{i_k, i_k}\right) \boldsymbol{\mathcal{X}}_{i_{[1:K]}} \\ &= -\sum_{k=1}^{K} \sum_{j_k \neq i_k} (\mathbf{\Psi}_k)_{i_k, j_k} \boldsymbol{\mathcal{X}}_{i_{[1:k]}, j_k, i_{[k+1:K]}} + \boldsymbol{\mathcal{T}}_{i_{[1:K]}}. \end{aligned} \tag{4.3}$$

Note that the left-hand side of (4.3) involves only the summation of the diagonals of the $\mathbf{\Psi}$'s and the right-hand side is composed of columns of $\mathbf{\Psi}$'s that exclude the diagonal terms. Equation (4.3) can be interpreted as an autogregressive model relating the $(i_1, \ldots, i_K)$-th element of the data tensor (scaled by the sum of diagonals) to other elements in the fibers of the data tensor. The columns of $\mathbf{\Psi}'s$ act as regression coefficients. The formulation in (4.3) naturally leads us to consider a pseudolikelihood-based estimation procedure [Besag, 1977] for estimating $\mathbf{\Omega}$. Specifically, we define the sparse estimate of the underlying precision matrices along each axis of the data as the

solution to the following convex optimization problem:

$$\begin{aligned}\min_{\substack{\boldsymbol{\Psi}_k \in \mathbb{R}^{m_k \times m_k} \\ k=1,\dots K}} & -N \sum_{i_1,\dots,i_K} \log \boldsymbol{\mathcal{W}}_{i_{[1:K]}} \\ & + \frac{1}{2} \sum_{i_1,\dots,i_K} \|(I) + (II)\|_2^2 + \sum_{k=1}^{K} P_{\lambda_k}(\boldsymbol{\Psi}_k).\end{aligned} \tag{4.4}$$

where $P_{\lambda_k}(\cdot)$ is a penalty function indexed by the tuning parameter $\lambda_k$ and

$$\begin{aligned}(I) &= \boldsymbol{\mathcal{W}}_{i_{[1:K]}} \boldsymbol{\mathcal{X}}_{i_{[1:K]}} \\ (II) &= \sum_{k=1}^{K} \sum_{j_k \neq i_k} (\boldsymbol{\Psi}_k)_{i_k, j_k} \boldsymbol{\mathcal{X}}_{i_{[1:k]}, j_k, i_{[k+1:K]}},\end{aligned}$$

with $\boldsymbol{\mathcal{W}}_{i_{[1:K]}} := \sum_{k=1}^{K} (\boldsymbol{\Psi}_k)_{i_k, i_k}$. Here we focus on the $\ell_1$-norm penalty, i.e., $P_{\lambda_k}(\boldsymbol{\Psi}_k) = \lambda_k \|\boldsymbol{\Psi}_k\|_{1,\text{off}}$.

The optimization problem (4.4) can be put into the following matrix form:

$$\begin{aligned}\min_{\substack{\boldsymbol{\Psi}_k \in \mathbb{R}^{m_k \times m_k} \\ k=1,\dots K}} & -\frac{N}{2} \log |(\text{diag}(\boldsymbol{\Psi}_1) \oplus \cdots \oplus \text{diag}(\boldsymbol{\Psi}_K))^2| \\ & + \frac{N}{2} \text{tr}(\mathbf{S}(\boldsymbol{\Psi}_1 \oplus \cdots \oplus \boldsymbol{\Psi}_K)^2) \\ & + \sum_{k=1}^{K} P_{\lambda_k}(\boldsymbol{\Psi}_k).\end{aligned}$$

where $\text{diag}(\boldsymbol{\Psi}_k) \in \mathbb{R}^{m_k \times m_k}$ is a matrix of the diagonal entries of $\boldsymbol{\Psi}_k$ and $\mathbf{S} \in \mathbb{R}^{m \times m}$ is the sample covariance matrix, i.e., $\mathbf{S} = \frac{1}{N} \text{vec}(\boldsymbol{\mathcal{X}})^T \text{vec}(\boldsymbol{\mathcal{X}})$. Note that the pseudolikelihood (4.5) approximates the $\ell_1$-penalized Gaussian negative loglikelihood in the log-determinant term by including only the Kronecker sum of the diagonal matrices instead of the Kronecker sum of the full matrices. Further discussion of pseudolikelihood- and likelihood-based approaches for (inverse) covariance estimations can be found in Khare et al. [2015].

We also note that when $K = 1$ the objective (4.4) reduces to the objective of the CONCORD estimator [Khare et al., 2015], and is similar to those of SPACE [Peng et al., 2009] and Symmetric lasso [Friedman et al., 2010]. Our framework is a generalization of these methods to higher-order tensor-valued data, when the Sylvester representation (4.1) holds.

**Remark:** In our formulation $\boldsymbol{\Omega} = (\bigoplus_{k=1}^{K} \boldsymbol{\Psi}_k)^2$ does not uniquely determine $\{\boldsymbol{\Psi}_k\}_{k=1}^{K}$ due to the trace ambiguity: scaled identity factors can be added to/subtracted from the $\boldsymbol{\Psi}_k's$ without changing

the matrix $\mathbf{\Omega}$. To address this non-identifiability, we rewrite the overall precision matrix $\mathbf{\Omega}$ as

$$\mathbf{\Omega} = \left( \bigoplus_{k=1}^{K} \mathbf{\Psi}_k \right)^2 = \left( \bigoplus_{k=1}^{K} \mathbf{\Psi}_k^{\text{off}} + \bigoplus_{k=1}^{K} \text{diag}(\mathbf{\Psi}_k) \right)^2, \tag{4.5}$$

where $\mathbf{\Psi}_k^{\text{off}} = \mathbf{\Psi}_k - \text{diag}(\mathbf{\Psi}_k)$, and estimate the off-diagonal entries $\mathbf{\Psi}_k^{\text{off}}$ and $\bigoplus_{k=1}^{K} \text{diag}(\mathbf{\Psi}_1)$. This allows us to reconstruct the overall precision matrix $\mathbf{\Omega}$ when $\mathbf{\Psi}_k^{\text{off}}$ is penalized with an $\ell_1$ penalty.

### 4.2.1 Estimation of the graphical model

Let $Q_N(\boldsymbol{\mathcal{W}}, \{\mathbf{\Psi}_k^{\text{off}}\}_{k=1}^K)$ denote the objective function in (4.4), where $\boldsymbol{\mathcal{W}} = \bigoplus_{k=1}^{K} \text{diag}(\mathbf{\Psi}_k)$. We adopt an alternating minimization approach that cycles between optimizing $\mathbf{\Psi}_k$ and $\boldsymbol{\mathcal{W}}$ while fixing other parameters. In particular, for $1 \leq k \leq K$, $1 \leq i_k < j_k \leq m_k$, define

$$\begin{aligned} T_{i_k j_k}(\mathbf{\Psi}_k^{\text{off}}) &= \underset{\substack{(\tilde{\mathbf{\Psi}}_l)_{m,n} = (\mathbf{\Psi}_l)_{m,n} \\ \forall (l,m,n) \neq (k,i_k,j_k)}}{\arg\min} \quad Q_N(\tilde{\boldsymbol{\mathcal{W}}}, \{\tilde{\mathbf{\Psi}}_k^{\text{off}}\}_{k=1}^K) \\ T(\boldsymbol{\mathcal{W}}) &= \underset{\substack{\tilde{\mathbf{\Psi}}_k^{\text{off}} = \mathbf{\Psi}_k^{\text{off}} \\ \forall k}}{\arg\min} \quad Q_N(\tilde{\boldsymbol{\mathcal{W}}}, \{\tilde{\mathbf{\Psi}}_k^{\text{off}}\}_{k=1}^K). \end{aligned} \tag{4.6}$$

For each $(k, i_k, j_k)$, $T_{i_k j_k}(\mathbf{\Psi}_k^{\text{off}})$ updates the $(i_k, j_k)$-th entry with the minimizer of $Q_N(\boldsymbol{\mathcal{W}}, \{\mathbf{\Psi}\}_{k=1}^K)$ with respect to $(\mathbf{\Psi}_k)_{i_k j_k}^{\text{off}}$ holding all other variables constant. Similarly, $T(\boldsymbol{\mathcal{W}})$ updates $\boldsymbol{\mathcal{W}}_{i_{[1:K]}}$ with the solution of $\min Q_N(\boldsymbol{\mathcal{W}}, \{\mathbf{\Psi}\}_{k=1}^K)$ with respect to $\boldsymbol{\mathcal{W}}_{i_{[1:K]}}$ holding all other variables constant. The closed form updates $T_{i_k j_k}(\mathbf{\Psi}_k^{\text{off}})$ and $T(\boldsymbol{\mathcal{W}})$ are detailed in Appendix 4.7.

**Tuning parameter:** The penalty parameters found in Meinshausen and Bühlmann [2006] and Friedman et al. [2008b] are equivalent to $\rho_k := \frac{\lambda_k}{N}$. For comparison, theoretical results in Meinshausen and Bühlmann [2006] calculate the regularization parameter to be $\rho(\alpha) = \frac{\Phi^{-1}(1-(\alpha/2p^2))}{\sqrt{N}}$ for standardized data, where $\Phi^{-1}$ is the standard Gaussian cumulative distribution function and $\alpha \in [0.01, 0.05]$ controls the false discovery rate. Peng et al. [2009] and Khare et al. [2015] have the same rate for the penalty parameter $\rho(\alpha) = C_1 \frac{\Phi^{-1}(1-(\alpha/2p^2))}{\sqrt{N}}$ for $C_1 > 0$ and further cross-validate $\rho(\alpha)$ based on BIC-type criteria. Extending these results, Algorithm 2 works with $\rho_k = \frac{\lambda_k}{N} = C_2 \frac{\Phi^{-1}(1-(\alpha/(2\prod_{j=1}^{K} m_j^2)))}{\sqrt{N \prod_{j \neq k} m_j}}$ for some constant $C_2 > 0$, which is equivalent to the results in Peng et al. [2009] and Khare et al. [2015] for multivariate setting.

## 4.3 Large Sample Properties

We show that under suitable conditions, the Sylvester graphical lasso (SyGlasso) estimator (Algorithm 2) achieves both model selection consistency and estimation consistency. As in other studies,

**Algorithm 2:** Nodewise SyGlasso

**Input:** Standardized data $\boldsymbol{\mathcal{X}}$, penalty parameter $\lambda_k$

**Output:** $\{\hat{\boldsymbol{\Psi}}_k\}_{k=1}^K$, $\hat{\boldsymbol{\Omega}} = \left(\bigoplus_{k=1}^K \hat{\boldsymbol{\Psi}}_k\right)^2$

Initialize $\{\hat{\boldsymbol{\Psi}}_k^{(0)}\}_{k=1}^K$, $\hat{\boldsymbol{\Omega}}^{(0)} = \left(\bigoplus_{k=1}^K \hat{\boldsymbol{\Psi}}_k^{(0)}\right)^2$

**while** *not converged* **do**

  # *Update off-diagonal elements*;

  **for** $k \leftarrow 1, \ldots, K$ **do**

    **for** $i_k \leftarrow 1, \ldots, m_k - 1$ **do**

      **for** $j_k \leftarrow i_k + 1, \ldots, m_k$ **do**

        $(\hat{\boldsymbol{\Psi}}_k^{(\text{t+1})})_{i_k,j_k} \leftarrow (T_{i_k,j_k}(\boldsymbol{\Psi}_k^{(\text{t})}))_{i_k,j_k}$;

        from (4.10) in Appendix 4.7.1

      **end**

    **end**

  **end**

  # *Update diagonal elements*;

  $\hat{\boldsymbol{\mathcal{W}}}^{(\text{t+1})} \leftarrow T(\boldsymbol{\mathcal{W}}^{(\text{t})})$ from (4.11) in Appendix 4.7.2

**end**

we make standard assumptions that the diagonal of $\boldsymbol{\Omega}$ is known. We analyze the theoretical properties of the SyGlasso under the assumption that $\boldsymbol{\mathcal{W}}$ is given. In practice, we can estimate $\boldsymbol{\mathcal{W}}$ using Algorithm 2, and if the diagonals of each individual $\boldsymbol{\Psi}_k$ are desired, we can incorporate any available prior knowledge of the variation along each data dimension.

We estimate $\{\boldsymbol{\Psi}_k^{\text{off}}\}_{k=1}^K$ by solving the following $\ell_1$ penalized problem:

$$\min_{\boldsymbol{\beta}} L_N\Big(\boldsymbol{\mathcal{W}}, \boldsymbol{\beta}, \boldsymbol{\mathcal{X}}\Big) + \sum_{k=1}^K \lambda_k \|\boldsymbol{\Psi}_k\|_{1,\text{off}}, \tag{4.7}$$

where $L_N\Big(\boldsymbol{\mathcal{W}}, \boldsymbol{\beta}, \boldsymbol{\mathcal{X}}\Big) := \frac{1}{N}\sum_{s=1}^N L\Big(\boldsymbol{\mathcal{W}}, \boldsymbol{\beta}, \boldsymbol{\mathcal{X}}^s\Big)$, with

$$\begin{aligned} L\Big(\boldsymbol{\mathcal{W}}, \boldsymbol{\beta}, \boldsymbol{\mathcal{X}}^s\Big) = -N \sum_{i_{[1:K]}} \log \boldsymbol{\mathcal{W}}_{i_{[1:K]}} \\ + \frac{1}{2} \sum_{i_1,\ldots,i_K} ((I) + (II))^2. \end{aligned} \tag{4.8}$$

where

$$(I) = \boldsymbol{\mathcal{W}}_{i_{[1:K]}} \boldsymbol{\mathcal{X}}_{i_{[1:K]}}$$
$$(II) = \sum_{k=1}^{K} \sum_{j_k \neq i_k} (\boldsymbol{\Psi}_k)_{i_k, j_k} \boldsymbol{\mathcal{X}}_{i_{[1:k-1]}, j_k, i_{[k+1:K]}}$$
$$\boldsymbol{\beta} = ((\boldsymbol{\Psi}_1)_{1,2}, (\boldsymbol{\Psi}_1)_{1,3}, \ldots, (\boldsymbol{\Psi}_1)_{1,m_1}, \ldots, (\boldsymbol{\Psi}_k)_{m_k-1, m_k})^T$$

and $\boldsymbol{\beta}$ denotes the off-diagonal entries of all $\boldsymbol{\Psi}_k's$.

We first state the regularity conditions needed for establishing convergence of the SyGlasso estimator. Let $\mathcal{A}_{N,k} := \{(i,j) : (\boldsymbol{\Psi}_k)_{i,j} \neq 0, i \neq j\}$ and $q_{N,k} := |\mathcal{A}_{N,k}|$ for $k = 1, \ldots, K$ be the true edge set and the number of edges, respectively. Let $\mathcal{A}_N = \cup_{k=1}^{K} \mathcal{A}_{N,k}$.

**(A1 - Subgaussianity)** The data $\boldsymbol{\mathcal{X}}^1, \ldots, \boldsymbol{\mathcal{X}}^N$ are i.i.d subgaussian random tensors, that is, $\text{vec}(\boldsymbol{\mathcal{X}}^i) \sim \mathbf{x}$, where $\mathbf{x}$ is a subgaussian random vector in $\mathbb{R}^p$, i.e., there exist a constant $c > 0$, such that for every $\mathbf{a} \in \mathbb{R}^p$, $\mathbb{E}e^{\mathbf{a}^T x} \leq e^{c\mathbf{a}^T \bar{\boldsymbol{\Sigma}} \mathbf{a}}$, and there exist $\rho_j > 0$ such that $\mathbb{E}e^{tx_j^2} \leq K$ whenever $|t| < \rho_j$, for $1 \leq j \leq p$.

**(A2 - Bounded eigenvalues)** There exist constants $0 < \Lambda_{\min} \leq \Lambda_{\max} < \infty$, such that the minimum and maximum eigenvalues of $\boldsymbol{\Omega}$ are bounded with $\lambda_{\min}(\bar{\boldsymbol{\Omega}}) = (\sum_{k=1}^{K} \lambda_{\max}(\boldsymbol{\Psi}_k))^{-2} \geq \Lambda_{\min}$ and $\lambda_{\max}(\bar{\boldsymbol{\Omega}}) = (\sum_{k=1}^{K} \lambda_{\min}(\boldsymbol{\Psi}_k))^{-2} \leq \Lambda_{\max}$.

**(A3 - Incoherence condition)** There exists a constant $\delta < 1$ such that for $k = 1, \ldots, K$ and all $(i,j) \in \mathcal{A}_{N,k}$

$$|\bar{L}''_{ij, \mathcal{A}_{N,k}}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}) [\bar{L}''_{\mathcal{A}_{N,k}, \mathcal{A}_{N,k}}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}})]^{-1} \text{sign}(\bar{\boldsymbol{\beta}}_{\mathcal{A}_{N,k}})| \leq \delta,$$

where for each $k$ and $1 \leq i < j \leq m_k$, $1 \leq k < l \leq m_k$,

$$\bar{L}''_{ij,kl}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}) := E_{\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}} \left( \frac{\partial^2 L(\boldsymbol{\mathcal{W}}, \boldsymbol{\beta}, \boldsymbol{\mathcal{X}})}{\partial (\boldsymbol{\Psi}_k)_{i,j} \partial (\boldsymbol{\Psi}_k)_{k,l}} |_{\boldsymbol{\mathcal{W}} = \bar{\boldsymbol{\mathcal{W}}}, \boldsymbol{\beta} = \bar{\boldsymbol{\beta}}} \right).$$

Note that conditions analogous to (A3) have been used in Meinshausen and Bühlmann [2006] and Peng et al. [2009] to establish high-dimensional model selection consistency of the nodewise graphical lasso in the case of $K = 1$. Zhao and Yu [2006] show that such a condition is almost necessary and sufficient for model selection consistency in lasso regression, and they provide some examples of when this condition is satisfied.

Inspired by Meinshausen and Bühlmann [2006] and Peng et al. [2009] we prove the following properties:

1. Theorem 3.1 establishes estimation consistency and sign consistency for the nodewise SyGlasso restricted to the true support, i.e., $\boldsymbol{\beta}_{\mathcal{A}_N^c} = 0$,

2. Theorem 3.2 shows that no wrong edge is selected with probability tending to one,

3. Theorem 3.3 establishes the consistency result of the nodewise SyGlasso.

**Theorem 4.3.1.** *Suppose that conditions (A1-A2) are satisfied. Suppose further that* $q_{N,k} = o(\sqrt{N/\log N})$, $\lambda_{N,k}\sqrt{N/\log N} \to \infty$, $\sqrt{q_{N,k}\log N/N} = o(\lambda_{N,k})$, *and* $\sqrt{q_{N,k}}\lambda_{N,k} = o(1)$ *as* $N \to \infty$, *for all* $k$. *Then there exists a constant* $C(\bar{\boldsymbol{\beta}})$, *such that for any* $\eta > 0$, *the following hold with probability at least* $1 - O(N^{-\eta})$*:*

- *There exists a global minimizer* $\hat{\boldsymbol{\beta}}_{\mathcal{A}_N}$ *of the restricted SyGlasso problem:*

$$\min_{\boldsymbol{\beta}:\boldsymbol{\beta}_{\mathcal{A}_N^c}=0} L_N(\bar{\boldsymbol{\mathcal{W}}}, \boldsymbol{\beta}, \boldsymbol{\mathcal{X}}) + \sum_{k=1}^{K} \lambda_k \|\boldsymbol{\Psi}_k\|_{1,off}. \tag{4.9}$$

- *(Estimation consistency) Any solution* $\hat{\boldsymbol{\beta}}_{\mathcal{A}_N}$ *of* (4.9) *satisfies:*

$$\|\hat{\boldsymbol{\beta}}_{\mathcal{A}_N} - \boldsymbol{\beta}_{\mathcal{A}_N}\|_2 \le C(\bar{\boldsymbol{\beta}})\sqrt{K}\max_k \sqrt{q_{N,k}}\lambda_{N,k}.$$

- *(Sign consistency) If further a minimal signal strength:* $\min_{(i,j)\in\mathcal{A}_{N,k}} |(\boldsymbol{\Psi}_k)_{i,j}| \ge 2C(\bar{\boldsymbol{\beta}})\sqrt{K}\max_k \sqrt{q_{N,k}}\lambda_{N,k}$ *is assumed for each* $k$, *then sign(*$\hat{\boldsymbol{\beta}}_{\mathcal{A}_{N,k}}$*)=sign(*$\bar{\boldsymbol{\beta}}_{\mathcal{A}_{N,k}}$*).*

**Theorem 4.3.2.** *Suppose that conditions (A1-A3) are satisfied. Suppose further that* $p = O(N^{\kappa})$ *for some* $\kappa \ge 0$, $q_{N,k} = o(\sqrt{N/\log N})$, $\lambda_{N,k}\sqrt{N/\log N} \to \infty$, $\sqrt{q_{N,k}\log N/N} = o(\lambda_{N,k})$, *and* $\sqrt{q_{N,k}}\lambda_{N,k} = o(1)$ *as* $N \to \infty$, *for all* $k$. *Then for* $\eta > 0$, *for* $N$ *sufficiently large, the solution of* (4.9) *satisfies:*

$$\begin{aligned} &P_{\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}}}\Big(\max_{(i,j)\in\mathcal{A}_{N,k}^c} |L'_{N,ij}(\bar{\boldsymbol{\mathcal{W}}}, \hat{\boldsymbol{\beta}}_{\mathcal{A}_{N,k}}, \boldsymbol{\mathcal{X}})| < \lambda_{N,k}\Big) \\ &\quad \ge 1 - O(N^{-\eta}) \end{aligned}$$

*for each* $k$, *where* $L'_{N,ij} := \partial L_N/\partial(\boldsymbol{\Psi}_k)_{ij}$.

**Theorem 4.3.3.** *Assume the conditions of Theorem 3.2. Then there exists a constant* $C(\bar{\boldsymbol{\beta}}) > 0$ *such that for any* $\eta > 0$ *the following events hold with probability at least* $1 - O(N^{-\eta})$*:*

- *There exists a global minimizer* $\hat{\boldsymbol{\beta}}$ *to problem* (4.4)*.*

- *(Estimation consistency) Any minimizer* $\hat{\boldsymbol{\beta}}$ *of* (4.4) *satisfies:*

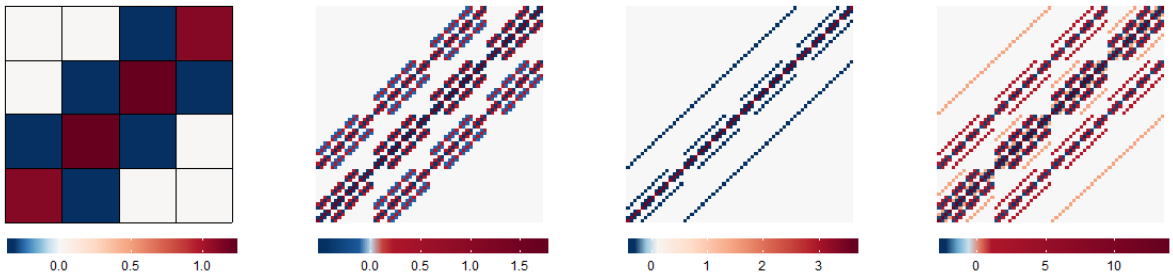
$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \le C(\bar{\boldsymbol{\beta}})\sqrt{K}\max_k \sqrt{q_{N,k}}\lambda_{N,k}.$$

- *(Sign consistency) If* $\min_{(i,j)\in\mathcal{A}_{N,k}} |(\boldsymbol{\Psi}_k)_{i,j}| \ge 2C(\bar{\boldsymbol{\beta}})\max_k \sqrt{q_{N,k}}\lambda_{N,k}$ *for each* $k$, *then sign(*$\hat{\boldsymbol{\beta}}$*)=sign(*$\bar{\boldsymbol{\beta}}$*).*

Proofs of the above theorems are given in Appendix 4.8.

## 4.4 Numerical Illustrations

We evaluate the proposed SyGlasso estimator (Algorithm 2) in terms of optimization and graph recovery accuracy. We also compare the graph recovery performance with other models recently proposed for matrix- and tensor-variate precision matrices. We first illustrate the differences among these models by investigating $\mathbf{\Omega}$ with $K = 3$ modes and $m_k = 4$. For simplicity, we generate $\mathbf{\Psi}_k$ for $k = 1, 2, 3$ as identical $4 \times 4$ precision matrices that follow a one dimensional autoregressive-1 (AR1) process. We recall the KP and KS models:




(a) $\mathbf{\Psi}_k$ (b) KP $\mathbf{\Omega}$ (c) KS $\mathbf{\Omega}$ (d) SyGlasso $\mathbf{\Omega}$

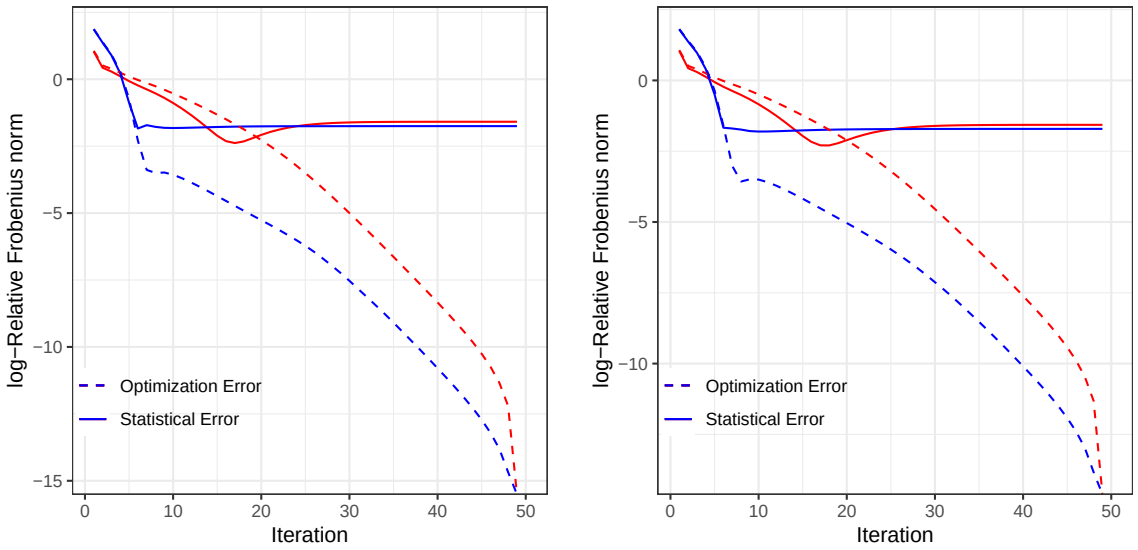
Figure 4.1: Comparison of SyGlasso to Kronecker sum (KS) and product (KP) structures. All models are composed of the same components $\mathbf{\Psi}_k$ for $k = 1, 2, 3$ generated as an AR(1) model with $m_k = 4$ as shown in (a). The AR(1) components are brought together to create the final $64 \times 64$ precision matrix $\mathbf{\Omega}$ following (b) the KP structure with $\mathbf{\Omega} = \bigotimes_{k=1}^{3} \mathbf{\Psi}_k$, (c) the KS structure with $\mathbf{\Omega} = \bigoplus_{k=1}^{3} \mathbf{\Psi}_k$, and (d) the proposed Sylvester model with $\mathbf{\Omega} = \left(\bigoplus_{k=1}^{3} \mathbf{\Psi}_k\right)^2$. The KP does not capture nested structures as it simply replicates the individual component with different multiplicative scales. The SyGlasso model admits a precision matrix structure that strikes a balance between KS and KP.

**Kronecker Product (KP):** The matrix-variate model under the sparse Kronecker product was introduced in Tsiligkaridis et al. [2013] and Zhou [2014]. The theoretical properties of KP models for the tensor-valued data were further analyzed in Lyu et al. [2019]. The precision matrices are decomposed as $\mathbf{\Omega} = \bigotimes_{k=1}^{K} \mathbf{\Psi}_k$. The KP model restricts the precision matrix to be separable across the $K$ data dimensions and suffers from a multiplicative explosion in the number of edges. As they are separable models and the constructed $\mathbf{\Omega}$ corresponds to the direct product of the $K$ graphs, KP is unable to capture more complex nested patterns captured by the KS and SyGlasso models as shown in Figure 4.1 (c) and (d).

**Kronecker Sum (KS):** Kalaitzis et al. [2013] first proposed to impose the Kronecker sum structure on the precision matrix for the matrix-variate data. Recently, Greenewald et al. [2019] introduced the TeraLasso that extended this KS structure to tensor-valued data. The TeraLasso was motivated by the relationship between the Kronecker sum of adjacency matrices and the Cartesian product of the associated graphs. Moreover, the covariance matrix under the Kronecker sum precision matrix assumption is nonseparable across $K$ data dimensions and has a maximum entropy motivation. Contrary to the KP structure, the number of edges in the Kronecker sum structure grows as the sum of the edges individual graphs. This growth allows the final precision matrix $\mathbf{\Omega}$ to remain sparse.

We compare these methods under different model assumptions to explore the flexibility of the proposed SyGlasso model under different model assumptions.

**Synthetic data:** To empirically assess the efficiency of the proposed model, we generate tensor-valued data based on three different precision matrices. The $\mathbf{\Psi}_k$'s are generated from one of 1) AR1($\rho$), 2) Star-Block (SB), or 3) Erdos-Renyi (ER) random graph models described in Appendix 4.9.



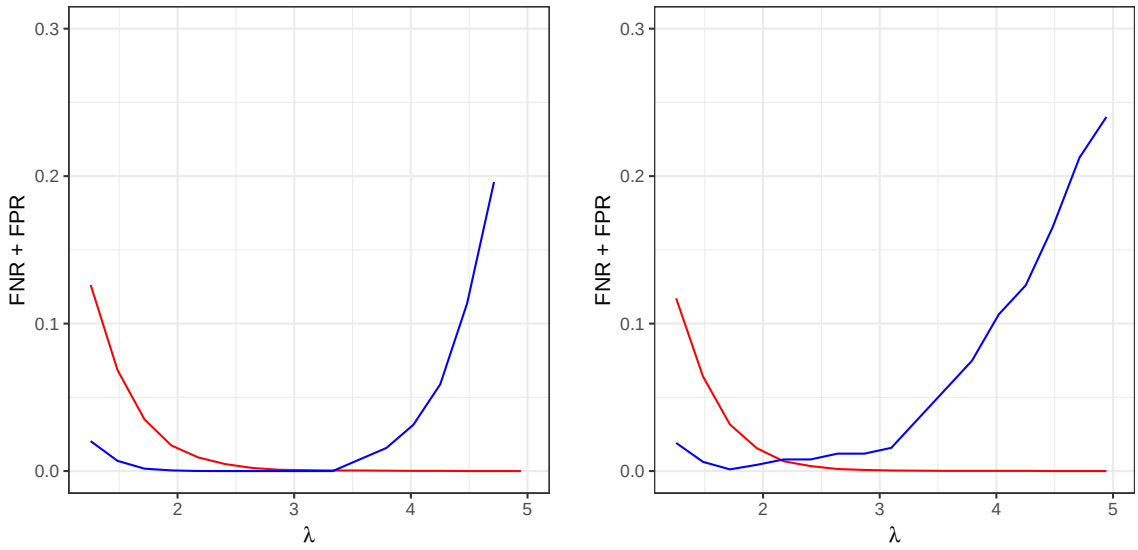

(a) SB and AR

(b) SB and ER

Figure 4.2: Performance of the SyGlasso estimator against the number of iterations under different topologies of $\mathbf{\Psi}_k$'s. The solid line shows the statistical error $\log(\|\hat{\mathbf{\Psi}}_k^{(t)} - \mathbf{\Psi}_k\|_F \backslash \|\mathbf{\Psi}_k\|_F)$, and the dotted line shows the optimization error $\log(\|\hat{\mathbf{\Psi}}_k^{(t)} - \hat{\mathbf{\Psi}}_k\|_F \backslash \|\hat{\mathbf{\Psi}}_k\|_F)$, where $\hat{\mathbf{\Psi}}_k$ is the final SyGlasso estimator. The performances of $\mathbf{\Psi}_1$ and $\mathbf{\Psi}_2$ are represented by red and blue lines, respectively.

We test SyGlasso for $K = 2$ under: 1) SB with $\rho = 0.6$ and sub-blocks of size 16 and

AR1($\rho = 0.6$); 2) SB with $\rho = 0.6$ and sub-blocks of size 16 and ER with 256 randomly selected edges. In both scenarios we set $m_1 = 128$ and $m_2 = 256$ with 10 samples. Figure 4.2 shows the iterative optimization performance of Algorithm 2. All the plots for the various scenarios exhibit iterative optimization approximation errors that quickly converge to below the statistical errors. Note that these plots also suggest that our algorithm can attain linear convergence rates. We also test our method for model selection accuracy over a range of penalty parameters (we set $\lambda_k = \lambda, \forall k$). Figure 4.3 displays the sum of false positive rate and false negative rate (FPR+FNR), which suggests that the nodewise SyGlasso estimator is able to fully recover the graph structures for each mode of the tensor data.




(a) SB and AR

(b) SB and ER

Figure 4.3: The performance of model selection measured by FPR + FNR. The performances of $\mathbf{\Psi}_1$ and $\mathbf{\Psi}_2$ are represented by red and blue lines, respectively. With an appropriate choice of $\lambda$, the SyGlasso recovers the dependency structures encoded in each $\mathbf{\Psi}_k$.

We compare the proposed SyGlasso to the TeraLasso estimator [Greenewald et al., 2019], and to the Tlasso estimator proposed by Lyu et al. [2019] for KP, on data generated using precision matrices $(\mathbf{\Psi}_1 \oplus \mathbf{\Psi}_2 \oplus \mathbf{\Psi}_3)^2$, $\mathbf{\Psi}_1 \oplus \mathbf{\Psi}_2 \oplus \mathbf{\Psi}_3$, and $\mathbf{\Psi}_1 \otimes \mathbf{\Psi}_2 \otimes \mathbf{\Psi}_3$, where $\mathbf{\Psi}$'s are each $16 \times 16$ ER graphs with 16 nonzero edges. We use the Matthews correlation coefficient (MCC) to compare model selection performances. The MCC is defined as [Matthews, 1975]

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(TP + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}},$$

where we follow Greenewald et al. [2019] to consider each nonzero off-diagonal element of $\mathbf{\Psi}_k$ as

a single edge.

The results shown in Figure 4.4 indicate that all three estimators perform well when $N = 5$, even under model misspecification. In the single sample scenario, the graph recovery performance of each estimator does well under each true underlying data generating process. Note that for data generated using KP, the SyGlasso performs surprisingly well and is comparable to Tlasso. These results seem to indicate that SyGlasso is very robust under model misspecification. The superior performance of SyGlasso under KP model, even with one sample, suggests again that SyGlasso structure has a flavor of both KS and KP structures, as seen in Figure 4.1. This follows from the observation that $(\mathbf{\Psi}_1 \oplus \mathbf{\Psi}_2)^2 = \mathbf{I}_{m_1} \otimes \mathbf{\Psi}_1^2 + \mathbf{\Psi}_2^2 \otimes \mathbf{I}_{m_2} + 2\mathbf{\Psi}_1 \otimes \mathbf{\Psi}_2 = \mathbf{\Psi}_1^2 \oplus \mathbf{\Psi}_2^2 + 2\mathbf{\Psi}_1 \otimes \mathbf{\Psi}_2$.

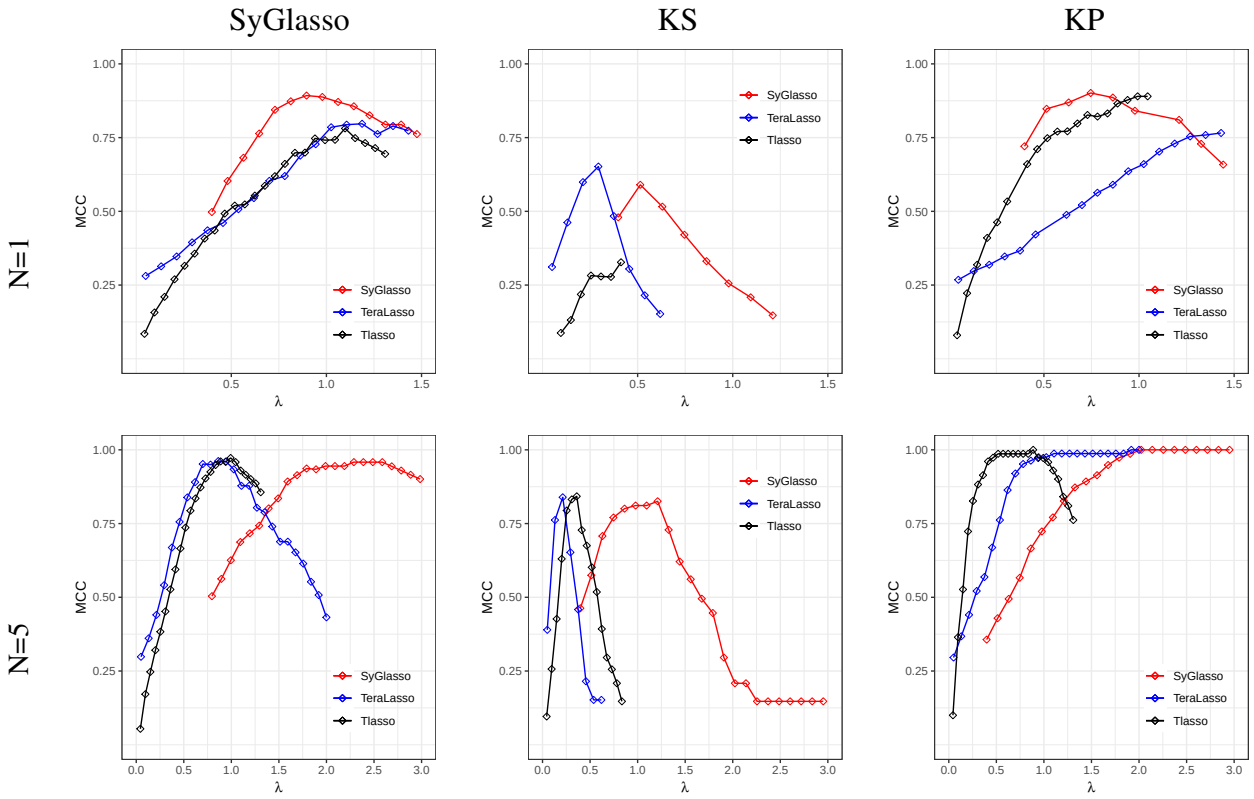



Figure 4.4: Performance of SyGlasso, TeraLasso (KS), and Tlasso (KP) measured by MCC under model misspecification. MCC of 1 represents a perfect recovery of the sparsity pattern in $\mathbf{\Omega}$, and MCC of 0 corresponds to a random guess. From top to bottom, the synthetic data were generated with the precision matrices from SyGlasso, KS, and KP models. The left column shows the results for a single sample ($N = 1$), and the right column shows the results for $N = 5$ observations.

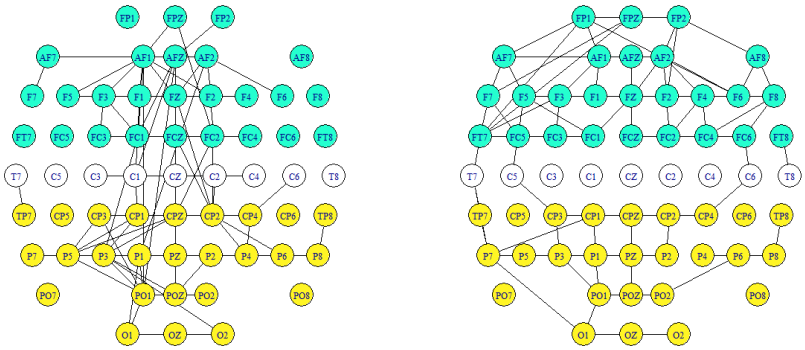
## 4.5 EEG Analysis

With the proposed SyGlasso, we revisit the alcoholism study conducted by Zhang et al. [1995] to explore multiway relationships in EEG measurements of alcoholic and control subjects. Each of 77 alcoholic subjects and 45 control subjects was visually stimulated by either a single picture or a pair of pictures on a computer monitor. Following the analyses of Zhu et al. [2016] and Qiao et al. [2019], we focus on the $\alpha$ frequency band (8 - 13 Hz) that is known to be responsible for the inhibitory control of the subjects (see Knyazev [2007] for more details). The EEG signals were bandpass filtered with the cosine-tapered window to extract $\alpha$-band signals. Previous Gaussian graphical models applied to such $\alpha$ frequency band filtered EEG data could only estimate the connectivity of the electrodes as they cannot be generalized to tensor valued data. The SyGlasso reveals a similar dependency structure as reported in Zhu et al. [2016] and Qiao et al. [2019] while recovering the chain structure of the temporal relationship. To show the benefit of multiway dependency, we estimate 2-way dependencies based on SyGlasso. After the band-pass filter was applied, we work with the tensor data $\boldsymbol{\mathcal{X}}_{alcoholic}, \boldsymbol{\mathcal{X}}_{control} \in \mathbb{R}^{n_{nodes} \times n_{time} \times n_{trial}}$ corresponding to an alcoholic subject and a control subject. We simultaneously estimate $\boldsymbol{\Psi}_{node} \in \mathbb{R}^{n_{node} \times n_{node}}$ that encodes the dependency structure among electrodes and $\boldsymbol{\Psi}_{time} \in \mathbb{R}^{n_{time} \times n_{time}}$ that shows the relationship among time points that span the duration of each trial.

Previous studies consider the average of all trials, for each subject and use the number of subjects as observations to estimate the dependency structures among p = 64 electrodes. Instead, we look at one subject at a time and consider different experimental trials as observations. Our analysis focuses on recovering the precision matrices of electrodes and time points, but it can be easily generalized to estimate the dependency structure among trials as well.

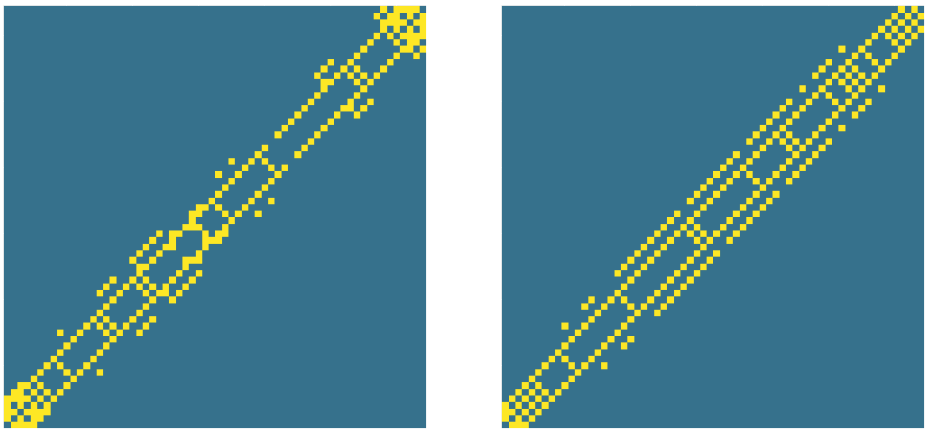
Figure 4.5 shows the result of the SyGlasso estimated network of electrodes. For comparison, both graphs were thresholded to match 5% sparsity level. Similar to the findings of Qiao et al. [2019], our estimated graph $\boldsymbol{\Psi}_{node}$ for the alcoholic group shows the asymmetry between the left and the right side of the brain compared to the more balanced control group. Our finding is consistent with the result in Hayden et al. [2006] and Zhu et al. [2016] that showed frontal asymmetry of the alcoholic subjects.

While previous analyses of this EEG data using graphical models only focused on the precision matrix of the electrodes, here we exhibit the second precision matrix factor that encodes temporal dependency. Figure 4.6 shows a comparison between the alcoholic subject and the control subject. Overall both of these graphs show a strong autoregressive structure. Both subjects exhibit banded dependency structures over time since adjacent timepoints are conditionally dependent.




(a) Alcoholic subject (b) Control subject

Figure 4.5: Estimated brain connectivity results from SyGlasso for (a) the alcoholic subject and (b) the control subject. The blue nodes correspond to the frontal region, and the yellow nodes correspond to the parietal and occipital regions. The alcoholic subject has asymmetric brain connections in the frontal region compared to the control subject.



(a) Alcoholic subject (b) Control subject

Figure 4.6: Estimated time conditional dependencies $\hat{\mathbf{\Psi}}_{time}$ from SyGlasso for (a) the alcoholic subject and (b) the control subject. Both subjects experience banded graph structures over time.

## 4.6 Discussion

This paper proposed Sylvester-structured graphical model and an inference algorithm, the SyGlasso, that can be applied to tensor-valued data. The current tools available for researchers are

limited to Kronecker product and Kronecker sum models on either the covariance or the precision matrix. Our model is motivated by a generative stochastic representation based on the Sylvester equation. We showed that the resulting precision matrix corresponds to the squared Kronecker sum of the precision matrices $\mathbf{\Psi}_k$ along each mode. The individual components $\mathbf{\Psi}_k$'s are estimated by the nodewise regression based approach.

Development of the proposed SyGlasso has promising future directions. First is to relax the assumption that the diagonals of the factors are fixed - an assumption that is standard in the Kronecker structured models literature. While SyGlasso can recover the off-diagonals of the individual components and $\bigoplus_{k=1}^{K} \text{diag}(\mathbf{\Psi}_k)$, it cannot recover the diagonal elements of the individual components separately. In practice, such restriction can be addressed when there is prior knowledge of the variation along each mode. However, we believe that analyzing the sparsity pattern of the squared Kronecker sum matrix would help us estimate the diagonal entries of the individual components $\mathbf{\Psi}_k$'s. In terms of improving the statistical convergence, our theoretical results guarantee variable selection of the individual components but do not guarantee the statistical convergence of individual $\mathbf{\Psi}_k$'s with respect to the operator norm. Similar to the solution in Zhou et al. [2011] for the multivariate case, we propose a two-step procedure using SyGlasso for variable selection followed by refitting the precision matrix $\mathbf{\Omega}$ using maximum likelihood estimation with the edge constraint.

Recently, Wang and Hero [2021] investigated the utility of Sylvester equation by relating to the generative model of the discretization of separable PDEs with finite elements. Specifically, Wang and Hero [2021] connected the Syglasso generative equation with the convection-diffusion type of PDEs and demonstrated the benefit of SyGlasso for forecasting of solar-flares.

While Syglasso has an interpretable connection to a family of PDE problems, there still remains an open problem of understanding the space of Kronecker covariance models. Without the domain knowledge, there is no theoretical guideline on selecting a covariance model from the family of Kronecker structured covariance models. A promising direction in understanding the space of Kronecker structures has been recently presented in Benzi and Simoncini [2017] based on the matrix exponential relationship of the Kronecker structures, i.e. $\exp(A \oplus B) = \exp(A) \otimes \exp(B)$. This exponential relationship is shown to be useful for evaulating the total communicability of a complex network in Estrada et al. [2012] and Benzi and Klymko [2013]. Further investigation towrads the matrix expoential relationship of Kronecker structures in covariance models would provide a meaningful direction towards understanding the model selection problem for Kronecker structured covariance models.

# Appendix

**A** provides the detailed derivation of the updates for Algorithm 2.

**B** provides the proofs of theorems stated in Section 4.3.

**C** provides details on the simulated data in Section 4.4.

## 4.7 Derivation of the Nodewise Tensor Lasso Estimator

### 4.7.1 Off-Diagonal updates

For $1 \leq i_k < j_k \leq m_k$, $T_{i_k j_k}(\mathbf{\Psi}_k^{\text{off}})$ can be computed in closed form:

$$(T_{i_k j_k}(\mathbf{\Psi}_k))^{\text{off}}_{i_k j_k} = \frac{S_{\frac{\lambda_k}{N}}\Big(F_{\boldsymbol{\mathcal{X}},\{\mathbf{\Psi}_k\}_{k=1}^K}\Big)}{(\frac{1}{N}\boldsymbol{\mathcal{X}}_{(k)}\boldsymbol{\mathcal{X}}_{(k)}^T)_{i_k i_k} + (\frac{1}{N}\boldsymbol{\mathcal{X}}_{(k)}\boldsymbol{\mathcal{X}}_{(k)}^T)_{j_k j_k}}, \tag{4.10}$$

where

$$\begin{aligned} F_{\boldsymbol{\mathcal{X}},\{\mathbf{\Psi}_k\}_{k=1}^K} = -\frac{1}{N}\Bigg( & \Big((\boldsymbol{\mathcal{W}}_{(k)} \circ \boldsymbol{\mathcal{X}}_{(k)})\boldsymbol{\mathcal{X}}_{(k)}^T\Big)_{i_k j_k} + \Big((\boldsymbol{\mathcal{W}}_{(k)} \circ \boldsymbol{\mathcal{X}}_{(k)})\boldsymbol{\mathcal{X}}_{(k)}^T\Big)_{j_k i_k} \\ & + \Big(\boldsymbol{\mathcal{X}}_{(k)}(\boldsymbol{\mathcal{X}} \times_k \mathbf{\Psi}_k^{\text{off},i_k j_k})_{(k)}^T\Big)_{j_k i_k} + \Big(\boldsymbol{\mathcal{X}}_{(k)}(\boldsymbol{\mathcal{X}} \times_k \mathbf{\Psi}_k^{\text{off},i_k j_k})_{(k)}^T\Big)_{i_k j_k} \\ & + \sum_{l \neq k}\Big(\boldsymbol{\mathcal{X}}_{(k)}(\boldsymbol{\mathcal{X}} \times_l \mathbf{\Psi}_l^{\text{off}})_{(k)}^T\Big)_{i_k j_k} + \sum_{l \neq k}\Big(\boldsymbol{\mathcal{X}}_{(k)}(\boldsymbol{\mathcal{X}} \times_l \mathbf{\Psi}_l^{\text{off}})_{(k)}^T\Big)_{j_k i_k}\Bigg). \end{aligned}$$

Here the $\circ$ operator denotes the Hadamard product between matrices; $\mathbf{\Psi}_k^{\text{off},i_k j_k}$ is $\mathbf{\Psi}_k^{\text{off}}$ with the $(i_k, j_k)$ entry being zero; and $S_\lambda(x) := \text{sign}(x)(|x| - \lambda)_+$ is the soft-thresholding operator.

### 4.7.2 Diagonal updates

For $\boldsymbol{\mathcal{W}}$,

$$(T(\boldsymbol{\mathcal{W}}))_{i_{[1:K]}} = \frac{-\Big(\boldsymbol{\mathcal{X}}_{(N)}^T\boldsymbol{\mathcal{Y}}_{(N)}\Big)_{i_{[1:K]}} + \sqrt{\Big(\boldsymbol{\mathcal{X}}_{(N)}^T\boldsymbol{\mathcal{Y}}_{(N)}\Big)^2_{i_{[1:K]}} + 4\Big(\boldsymbol{\mathcal{X}}_{(N)}\boldsymbol{\mathcal{X}}_{(N)}^T\Big)_{i_{[1:K]}}}}{2\Big(\boldsymbol{\mathcal{X}}_{(N)}\boldsymbol{\mathcal{X}}_{(N)}^T\Big)_{i_{[1:K]}}}. \tag{4.11}$$

Here we define $\boldsymbol{\mathcal{Y}} := \sum_{k=1}^K \Big(\boldsymbol{\mathcal{X}} \times_k \mathbf{\Psi}_k^{\text{off}}\Big)$. Equations (4.10) and (4.11) give necessary ingredients for designing a coordinate descent approach to minimizing the objective function in (4.4). The

optimization procedure is summarized in Algorithm 2.

### 4.7.3 Derivation of updates

Note that for $1 \leq i_k < j_k \leq m_k$, $1 \leq k \leq K$,

$$\begin{aligned}
&Q_N(\{\boldsymbol{\Psi}_k\}_{k=1}^K) \\
&= (N/2)\Big(\sum_{i_{[1:k-1,k+1:K]}} (\boldsymbol{\mathcal{X}}_{i_{[1:K]}}^{i_k}{}^2 + \boldsymbol{\mathcal{X}}_{i_{[1:K]}}^{j_k}{}^2)\Big)\Big((\boldsymbol{\Psi}_k)_{i_k j_k}\Big)^2 \\
&+ NF_{\boldsymbol{\mathcal{X}},\{\boldsymbol{\Psi}\}_{k=1}^K}(\boldsymbol{\Psi}_k)_{i_k j_k} + \lambda_k|(\boldsymbol{\Psi}_k)_{i_k j_k}| \\
&+ \text{terms independent of } (\boldsymbol{\Psi}_k)_{i_k j_k},
\end{aligned}$$

where

$$\begin{aligned}
F_{\boldsymbol{\mathcal{X}},\{\boldsymbol{\Psi}\}_{k=1}^K} = -\sum_{i_{[1:k-1,k+1:K]}} \Big(&\boldsymbol{\mathcal{W}}_{i_{[1:K]}}^{i_k} \boldsymbol{\mathcal{X}}_{i_{[1:K]}}^{i_k} \boldsymbol{\mathcal{X}}_{i_{[1:K]}}^{j_k} + \boldsymbol{\mathcal{W}}_{i_{[1:K]}}^{j_k} \boldsymbol{\mathcal{X}}_{i_{[1:K]}}^{j_k} \boldsymbol{\mathcal{X}}_{i_{[1:K]}}^{i_k} \\
&+ (\boldsymbol{\Psi}_k)_{i_k,\setminus\{i_k,j_k\}}^T \boldsymbol{\mathcal{X}}_{i_{[1:K]}}^{\setminus\{i_k,j_k\}} \boldsymbol{\mathcal{X}}_{i_{[1:K]}}^{j_k} \\
&+ (\boldsymbol{\Psi}_k)_{j_k,\setminus\{i_k,j_k\}}^T \boldsymbol{\mathcal{X}}_{i_{[1:K]}}^{\setminus\{i_k,j_k\}} \boldsymbol{\mathcal{X}}_{i_{[1:K]}}^{i_k} \\
&+ \sum_{l\in[1:k-1,k+1:K]} (\boldsymbol{\Psi}_l)_{i_l,\setminus i_l}^T \boldsymbol{\mathcal{X}}_{i_{[1:K]}}^{i_k,\setminus i_l} \boldsymbol{\mathcal{X}}_{i_{[1:K]}}^{j_k} \\
&+ \sum_{l\in[1:k-1,k+1:K]} (\boldsymbol{\Psi}_l)_{i_l,\setminus i_l}^T \boldsymbol{\mathcal{X}}_{i_{[1:K]}}^{j_k,\setminus i_l} \boldsymbol{\mathcal{X}}_{i_{[1:K]}}^{i_k} \Big).
\end{aligned}$$

Here $\boldsymbol{\mathcal{X}}_{i_{[1:K]}}^{i_k}$ denotes the element of $\boldsymbol{\mathcal{X}}$ indexed by $i_{[1:K]}$ except that the $k$th index is replaced by $i_k$ and $\boldsymbol{\mathcal{X}}_{i_{[1:K]}}^{i_k,j_l}$ denotes the element of $\boldsymbol{\mathcal{X}}$ indexed by $i_{[1:K]}$ except that the $k,l$th indices are replaced by $i_k, j_l$. Note the following equivalence:

$$\begin{aligned}
\sum_{i_{[1:k-1,k+1:K]}} \boldsymbol{\mathcal{W}}_{i_{[1:K]}}^{i_k} \boldsymbol{\mathcal{X}}_{i_{[1:K]}}^{i_k} \boldsymbol{\mathcal{X}}_{i_{[1:K]}}^{j_k} &= \Big((\boldsymbol{\mathcal{W}}_{(k)} \circ \boldsymbol{\mathcal{X}}_{(k)}) \boldsymbol{\mathcal{X}}_{(k)}^T\Big)_{i_k j_k} \\
\sum_{i_{[1:k-1,k+1:K]}} \boldsymbol{\mathcal{X}}_{i_{[1:K]}}^{i_k} \boldsymbol{\mathcal{X}}_{i_{[1:K]}}^{j_k} &= (\boldsymbol{\mathcal{X}}_{(k)} \boldsymbol{\mathcal{X}}_{(k)}^T)_{i_k j_k} \\
\sum_{i_{[1:k-1,k+1:K]}} (\boldsymbol{\Psi}_l)_{i_l,\cdot}^T \boldsymbol{\mathcal{X}}_{i_{[1:K]}}^{i_k,\cdot} \boldsymbol{\mathcal{X}}_{i_{[1:K]}}^{j_k} &= \Big(\boldsymbol{\mathcal{X}}_{(k)} (\boldsymbol{\mathcal{X}} \times_l \boldsymbol{\Psi}_l)_{(k)}^T\Big)_{j_k i_k},
\end{aligned}$$

where $\boldsymbol{\mathcal{W}}$ is a tensor of the same dimensions of $\boldsymbol{\mathcal{X}}$, formed by tensorize values in $\boldsymbol{\mathcal{W}}$, and in the case of $N > 1$ the last mode of $\boldsymbol{\mathcal{W}}$ is the observation mode similarly to $\boldsymbol{\mathcal{X}}$ but with exact replicates. Using the tensor notation and standard sub-differential method, Equation (4.10) then follows.

For $\boldsymbol{\mathcal{W}}_{i_{[1:K]}}$, using similar tensor operations,

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\mathcal{W}}_{i_{[1:K]}}} Q_N(\boldsymbol{\mathcal{W}}, \{\boldsymbol{\Psi}_k^{\text{off}}\}_{k=1}^K) = 0 \\
&\iff -\frac{1}{\boldsymbol{\mathcal{W}}_{i_{[1:K]}}} + \boldsymbol{\mathcal{W}}^2_{i_{[1:K]}} \boldsymbol{\mathcal{X}}^2_{i_{[1:K]}} + \boldsymbol{\mathcal{W}}_{i_{[1:K]}} \Big( \boldsymbol{\mathcal{X}}_{i_{[1:K]}} \sum_{k=1}^K (\boldsymbol{\mathcal{X}} \times_k \boldsymbol{\Psi}_k^{\text{off}})_{i_{[1:K]}} \Big) = 0 \\
&\iff \boldsymbol{\mathcal{W}}^2_{i_{[1:K]}} \Big( \boldsymbol{\mathcal{X}}^T_{(N)} \boldsymbol{\mathcal{X}}_{(N)} \Big)_{i_{[1:K]}} + \boldsymbol{\mathcal{W}}_{i_{[1:K]}} \Big( \boldsymbol{\mathcal{X}}^T_{(N)} \sum_{k=1}^K (\boldsymbol{\mathcal{X}} \times_k \boldsymbol{\Psi}_k^{\text{off}}) \Big)_{i_{[1:K]}} - 1 = 0
\end{aligned}$$

which is a quadratic equation in $\boldsymbol{\mathcal{W}}_{i_{[1:K]}}$ and since $\boldsymbol{\mathcal{W}}_{i_{[1:K]}} > 0$, so the positive root has been retained as the solution. Note that the estimation for one entry of $\boldsymbol{\mathcal{W}}$ is independent of the other entries. So during the estimation process we update all the entries at once by noting that $\text{diag}\Big( \boldsymbol{\mathcal{X}}^T_{(N)} \boldsymbol{\mathcal{X}}_{(N)} \Big) = \Big( \Big( \boldsymbol{\mathcal{X}}^T_{(N)} \boldsymbol{\mathcal{X}}_{(N)} \Big)_{i_{[1:K]}}, \forall i_{[1:K]} \Big)$.

## 4.8 Proofs of Main Theorems

We first list some properties of the loss function.

**Lemma 4.8.1.** *The following is true for the loss function:*

*(i) There exist constants* $0 < \Lambda^L_{\min} \leq \Lambda^L_{\max} < \infty$ *such that for* $\mathcal{S}_{N,k} := \{(i_k, j_k) : 1 \leq i_k < j_k \leq m_k\}, k = 1, \ldots, K$,

$$\Lambda^L_{\min} \leq \lambda_{\min}(\bar{L}''_{\mathcal{S}_{N,k}, \mathcal{S}_{N,k}}(\bar{\boldsymbol{\beta}})) \leq \lambda_{\max}(\bar{L}''_{\mathcal{S}_{N,k}, \mathcal{S}_{N,k}}(\bar{\boldsymbol{\beta}})) \leq \Lambda^L_{\max}$$

*(ii) There exists a constant* $K(\bar{\boldsymbol{\beta}}) < \infty$ *such that for all* $1 \leq i_k < j_k \leq m_k$, $\bar{L}''_{i_k j_k, i_k j_k}(\bar{\boldsymbol{\beta}}) \leq K(\bar{\boldsymbol{\beta}})$

*(iii) There exist constant* $M_1(\bar{\boldsymbol{\beta}}), M_2(\bar{\boldsymbol{\beta}}) < \infty$, *such that for any* $1 \leq i_k < j_k \leq m_k$

$$\text{Var}_{\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}}(L'_{i_k j_k}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}})) \leq M_1(\bar{\boldsymbol{\beta}}),\ \text{Var}_{\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}}(L''_{i_k j_k, i_k j_k}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}})) \leq M_2(\bar{\boldsymbol{\beta}})$$

*(iv) There exists a constant* $0 < g(\bar{\boldsymbol{\beta}}) < \infty$, *such that for all* $(i, j) \in \mathcal{A}_{N,k}$

$$\bar{L}''_{ij,ij}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}) - \bar{L}''_{ij, \mathcal{A}^{ij}_{N,k}}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}) [\bar{L}''_{\mathcal{A}^{ij}_{N,k}, \mathcal{A}^{ij}_{N,k}}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}})]^{-1} \bar{L}''_{\mathcal{A}^{ij}_{N,k}, ij}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}) \geq g(\bar{\boldsymbol{\beta}}),$$

*where* $\mathcal{A}^{ij}_{N,k} := \mathcal{A}_{N,k} / \{(i, j)\}$.

*(v) There exists a constant $M(\bar{\boldsymbol{\beta}}) < \infty$, such that for any $(i,j) \in \mathcal{A}^c_{N,k}$*

$$\|\bar{L}''_{ij,\mathcal{A}_{N,k}}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}})[\bar{L}''_{\mathcal{A}_{N,k},\mathcal{A}_{N,k}}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}})]^{-1}\|_2 \leq M(\bar{\boldsymbol{\beta}}).$$

*proof of Lemma B.1.* We prove $(i)$. $(ii - v)$ are then direct consequences, and the proofs follow from the proofs of B1.1-B1.4 in Peng et al. [2009], with the modifications being that the indexing is now with respect to each $k$ for $1 \leq k \leq K$.

Consider the loss function in matrix form as in (4.5). Then $\bar{L}''_{\mathcal{S}_{N,k},\mathcal{S}_{N,k}}(\bar{\boldsymbol{\beta}})$ is equivalent to $\frac{\partial^2}{\partial \boldsymbol{\Psi}_k^{\text{off}} \partial \boldsymbol{\Psi}_k^{\text{off}}} L(\boldsymbol{\mathcal{W}}, \{\boldsymbol{\Psi}_k^{\text{off}}\}_{k=1}^K)$, which is

$$\begin{aligned}
&\frac{\partial^2}{\partial \boldsymbol{\Psi}_k^{\text{off}} \partial \boldsymbol{\Psi}_k^{\text{off}}} \Bigg( \operatorname{tr}(\boldsymbol{\Psi}_k^T \mathbf{S} \boldsymbol{\Psi}_k) + \text{first order terms in } \boldsymbol{\Psi}_k + \text{terms independent of } \boldsymbol{\Psi}_k \Bigg) \\
&= \frac{\partial^2}{\partial \boldsymbol{\Psi}_k^{\text{off}} \partial \boldsymbol{\Psi}_k^{\text{off}}} \Bigg( \operatorname{tr}((\boldsymbol{\Psi}_k^{\text{off}} + \operatorname{diag}(\boldsymbol{\Psi}_k))^T \mathbf{S} (\boldsymbol{\Psi}_k^{\text{off}} + \operatorname{diag}(\boldsymbol{\Psi}_k))) + \text{first order terms in } \boldsymbol{\Psi}_k^{\text{off}} \\
&\qquad\qquad + \text{terms independent of } \boldsymbol{\Psi}_k^{\text{off}} \Bigg) \\
&= \frac{\partial^2}{\partial \boldsymbol{\Psi}_k^{\text{off}} \partial \boldsymbol{\Psi}_k^{\text{off}}} \Bigg( \operatorname{tr}((\boldsymbol{\Psi}_k^{\text{off}})^T \mathbf{S} \boldsymbol{\Psi}_k^{\text{off}}) + \text{first order terms in } \boldsymbol{\Psi}_k^{\text{off}} + \text{terms independent of } \boldsymbol{\Psi}_k^{\text{off}} \Bigg) \\
&= \mathbf{S} = \frac{1}{N} \operatorname{vec}(\boldsymbol{\mathcal{X}})^T \operatorname{vec}(\boldsymbol{\mathcal{X}}).
\end{aligned}$$

Thus $\bar{L}''_{\mathcal{S}_{N,k},\mathcal{S}_{N,k}}(\boldsymbol{\beta}) = E_{\boldsymbol{\mathcal{W}},\boldsymbol{\beta}}(\mathbf{S})$. Then for any non-zero $\mathbf{a} \in \mathbb{R}^p$, we have

$$\mathbf{a}^T \bar{L}''_{\mathcal{S}_{N,k},\mathcal{S}_{N,k}}(\bar{\boldsymbol{\beta}})\mathbf{a} = \mathbf{a}^T \bar{\boldsymbol{\Sigma}} \mathbf{a} \geq \|\mathbf{a}\|_2^2 \lambda_{\min}(\bar{\boldsymbol{\Sigma}}).$$

Similarly, $\mathbf{a}^T \bar{L}''_{\mathcal{S}_{N,k},\mathcal{S}_{N,k}}(\bar{\boldsymbol{\beta}})\mathbf{a} \leq \|\mathbf{a}\|_2^2 \lambda_{\max}(\bar{\boldsymbol{\Sigma}})$. By (A2), $\bar{\boldsymbol{\Sigma}}$ has bounded eigenvalues, thus the lemma is proved.

□

**Lemma 4.8.2.** *Suppose conditions (A1-A2) hold and if $q_{N,k} = o(\sqrt{N/\log N})$, then for any $\eta > 0$, there exist constant $c_{0,\eta}, c_{1,\eta}, c_{2,\eta}, c_{3,\eta}$, such that for any $u \in \mathbb{R}^{q_{N,k}}$ the following events hold with probability at least $1 - O(N^{-\eta})$ for sufficiently large $N$:*

*(i)* $\|L'_{N,\mathcal{A}_{N,k}}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}})\|_2 \leq c_{0,\eta} \sqrt{q_{N,k} \frac{\log N}{N}}$

*(ii)* $|u^T L'_{N,\mathcal{A}_{N,k}}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}})| \leq c_{1,\eta} \|u\|_2 \sqrt{q_{N,k} \frac{\log N}{N}}$

*(iii)* $|u^T L''_{N,\mathcal{A}_{N,k}\mathcal{A}_{N,k}}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}})u - u^T \bar{L}''_{\mathcal{A}_{N,k}\mathcal{A}_{N,k}}(\bar{\boldsymbol{\beta}})u| \leq c_{2,\eta} \|u\|_2^2 q_{N,k} \sqrt{\frac{\log N}{N}}$

*(iv)* $|L''_{N,\mathcal{A}_{N,k}\mathcal{A}_{N,k}}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})u - \bar{L}''_{\mathcal{A}_{N,k}\mathcal{A}_{N,k}}(\bar{\boldsymbol{\beta}})u| \le c_{3,\eta}\|u\|_2^2 q_{N,k}\sqrt{\frac{\log N}{N}}$

*proof of Lemma B.2.* $(i)$ By Cauchy-Schwartz inequality,

$$\|L'_{N,\mathcal{A}_{N,k}}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})\|_2 \le \sqrt{q_{N,k}}\max_{i\in\mathcal{A}_{N,k}}|L'_{N,i}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})|.$$

Then note that

$$\begin{aligned}&L'_{N,i}(\boldsymbol{\mathcal{W}},\boldsymbol{\beta},\boldsymbol{\mathcal{X}})\\ &=\sum_{i_{[1:k-1,k+1:K]}}(e_{i_{[1:k-1]},p,i_{[k+1:K]}}(\boldsymbol{\mathcal{W}},\boldsymbol{\beta})\boldsymbol{\mathcal{X}}_{i_{[1:k-1]},q,i_{[k+1:K]}}+e_{i_{[1:k-1]},q,i_{[k+1:K]}}(\boldsymbol{\mathcal{W}},\boldsymbol{\beta})\boldsymbol{\mathcal{X}}_{i_{[1:k-1]},p,i_{[k+1:K]}}),\end{aligned}$$

where $e_{i_{[1:k-1]},p,i_{[k+1:K]}}\boldsymbol{\mathcal{X}}_{i_{[1:k-1]},q,i_{[k+1:K]}}(\boldsymbol{\mathcal{W}},\boldsymbol{\beta})$ is defined by

$$w_{i_{[1:k-1]},p,i_{[k+1:K]}}\boldsymbol{\mathcal{X}}_{i_{[1:k-1]},p,i_{[k+1:K]}}+\sum_{j_k\neq p}(\boldsymbol{\Psi}_k)_{p,j_k}\boldsymbol{\mathcal{X}}_{i_{[1:k-1]},j_k,i_{[k+1:K]}}+\sum_{l\neq k}\sum_{j_l\neq i_l}(\boldsymbol{\Psi}_l)_{i_l,j_l}\boldsymbol{\mathcal{X}}_{i_{[1:k-1]},p,i_{[k+1:K]}}.$$

Then evaluated at the true parameter values $(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}})$, we have $e_{i_{[1:k-1]},p,i_{[k+1:K]}}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}})$ uncorrelated with $\boldsymbol{\mathcal{X}}_{i_{[1:k-1]},\backslash p,i_{[k+1:K]}}$ and $E_{(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}})}(e_{i_{[1:k-1]},p,i_{[k+1:K]}}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}}))=0$. Also, since $\boldsymbol{\mathcal{X}}$ is subgaussian and $\mathrm{Var}(L'_{N,i}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}}))$ is bounded by Lemma C.1. $\forall i$, $L'_{N,i}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})$ has subexponential tails. Thus, by Bernstein inequality,

$$\begin{aligned}&P(\|L'_{N,\mathcal{A}_{N,k}}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})\|_2\le c_{0,\eta}\sqrt{q_{N,k}\frac{\log N}{N}})\\ &\ge P(\sqrt{q_{N,k}}\max_{i\in\mathcal{A}_{N,k}}|L'_{N,i}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})|\le c_{0,\eta}\sqrt{q_{N,k}\frac{\log N}{N}})\ge 1-O(N^{-\eta}).\end{aligned}$$

$(iii)$ By Cauchy-Schwartz,

$$\begin{aligned}&|u^TL''_{N,\mathcal{A}_{N,k}\mathcal{A}_{N,k}}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})u-u^T\bar{L}''_{\mathcal{A}_{N,k}\mathcal{A}_{N,k}}(\bar{\boldsymbol{\beta}})u|\\ &\le\|u\|_2\|u^TL''_{N,\mathcal{A}_{N,k}\mathcal{A}_{N,k}}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})-u^T\bar{L}''_{\mathcal{A}_{N,k}\mathcal{A}_{N,k}}(\bar{\boldsymbol{\beta}})\|_2\\ &\le\|u\|_2\sqrt{q_{N,k}}\max_i|u^TL''_{N,\mathcal{A}_{N,k},i}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})-u^T\bar{L}''_{\mathcal{A}_{N,k},i}(\bar{\boldsymbol{\beta}})|\\ &=\|u\|_2\sqrt{q_{N,k}}|u^TL''_{N,\mathcal{A}_{N,k},i_{\max}}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})-u^T\bar{L}''_{\mathcal{A}_{N,k},i_{\max}}(\bar{\boldsymbol{\beta}})|\\ &=\|u\|_2\sqrt{q_{N,k}}|\sum_{j=1}^{q_{N,k}}(u_jL''_{N,j,i_{\max}}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})-u_j\bar{L}''_{j,i_{\max}}(\bar{\boldsymbol{\beta}}))|\\ &\le\|u\|_2q_{N,k}|u_{j_{\max}}||L''_{N,j_{\max},i_{\max}}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})-\bar{L}''_{j_{\max},i_{\max}}(\bar{\boldsymbol{\beta}}))|\\ &\le\|u\|_2^2q_{N,k}|L''_{N,j_{\max},i_{\max}}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})-\bar{L}''_{j_{\max},i_{\max}}(\bar{\boldsymbol{\beta}}))|.\end{aligned}$$

Then by Bernstein inequality,

$$
\begin{aligned}
&P(|u^T L''_{N,\mathcal{A}_{N,k}\mathcal{A}_{N,k}}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}})u - u^T \bar{L}''_{\mathcal{A}_{N,k}\mathcal{A}_{N,k}}(\bar{\boldsymbol{\beta}})u| \leq c_{2,\eta}\|u\|_2^2 q_{N,k}\sqrt{\frac{\log N}{N}}) \\
&\geq P(\|u\|_2^2 q_{N,k}|L''_{N,j_{\max},i_{\max}}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}}) - \bar{L}''_{j_{\max},i_{\max}}(\bar{\boldsymbol{\beta}}))| \leq c_{2,\eta}\|u\|_2^2 q_{N,k}\sqrt{\frac{\log N}{N}}) \\
&\geq 1 - O(N^{-\eta}).
\end{aligned}
$$

$(ii)$ and $(iv)$ can be proved using similar arguments. □

Lemma C.3. and C.4. are used later to prove Theorem 1.

**Lemma 4.8.3.** *Assuming conditions of Theorem 1. Then there exists a constant $C_1(\bar{\boldsymbol{\beta}}) > 0$ such that for any $\eta > 0$, there exists a global minimizer of the restricted problem* (4.9) *within the disc:*

$$\{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}\|_2 \leq C_1(\bar{\boldsymbol{\beta}})\sqrt{K}\max_k \sqrt{q_{N,k}}\lambda_{N,k}\}$$

*with probability at least $1 - O(N^{-\eta})$ for sufficiently large $N$.*

*proof of Lemma B.3.* Let $\alpha_N = \max_k \sqrt{q_{N,k}}\lambda_{N,k}$. Further for $1 \leq k \leq K$ let $C_k > 0$ and $u^k \in \mathbb{R}^{m_k(m_k-1)/2}$ such that $u^k_{\mathcal{A}^c_{N,k}} = 0$, $\|u^k\|_2 = C_k$, and $u = (u_1, \ldots, u_K)$ with $\sqrt{K}\min_k C_k \leq \|u\|_2 \leq \sqrt{K}\max_k C_k$.

Then by Cauchy-Schwartz and triangle inequality, we have

$$\|\bar{\boldsymbol{\beta}}^k + \alpha_N u^k - \alpha_N u^k\|_1 \leq \|\bar{\boldsymbol{\beta}}^k + \alpha_N u^k\|_1 + \alpha_N\|u^k\|_1,$$

and

$$\|\bar{\boldsymbol{\beta}}^k\|_1 - \|\bar{\boldsymbol{\beta}}^k + \alpha_N u^k\|_1 \leq \alpha_N\|u^k\|_1 \leq \alpha_N\sqrt{q_{N,k}}\|u^k\|_2 = C_k\alpha_N\sqrt{q_{N,k}}.$$

Thus,

$$
\begin{aligned}
&Q_N(\bar{\boldsymbol{\beta}} + \alpha_N u, \boldsymbol{\mathcal{X}}, \{\lambda_{N,k}\}_{k=1}^K) - Q_N(\bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}}, \{\lambda_{N,k}\}_{k=1}^K) \\
&= L_N(\bar{\boldsymbol{\beta}} + \alpha_N u, \boldsymbol{\mathcal{X}}) - L_N(\bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}}) - \sum_{k=1}^K \lambda_{N,k}\big(\|\bar{\boldsymbol{\beta}}^k\|_1 - \|\bar{\boldsymbol{\beta}}^k + \alpha_N u^k\|_1\big) \\
&\geq L_N(\bar{\boldsymbol{\beta}} + \alpha_N u, \boldsymbol{\mathcal{X}}) - L_N(\bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}}) - \sum_{k=1}^K \lambda_{N,k}C_k\alpha_N\sqrt{q_{N,k}} \\
&\geq L_N(\bar{\boldsymbol{\beta}} + \alpha_N u, \boldsymbol{\mathcal{X}}) - L_N(\bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}}) - \alpha_N K\max_k C_k\sqrt{q_{N,k}}\lambda_{N,k} \\
&\geq L_N(\bar{\boldsymbol{\beta}} + \alpha_N u, \boldsymbol{\mathcal{X}}) - L_N(\bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}}) - K\alpha_N^2\max_k C_k.
\end{aligned}
$$

Next,

$$
\begin{aligned}
&L_N(\bar{\boldsymbol{\beta}}+\alpha_N u,\boldsymbol{\mathcal{X}})-L_N(\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})=\alpha_N u_{\mathcal{A}_N}^T L'_{N,\mathcal{A}_N}(\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})+\frac{1}{2}\alpha_N^2 u_{\mathcal{A}_N}^T L''_{N,\mathcal{A}_N\mathcal{A}_N}(\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})u_{\mathcal{A}_N}\\
&=\alpha_N\sum_{k=1}^K (u^k_{\mathcal{A}_{N,k}})^T L'_{N,\mathcal{A}_{N,k}}(\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})+\frac{1}{2}\alpha_N^2\sum_{k=1}^K (u^k_{\mathcal{A}_{N,k}})^T L''_{N,\mathcal{A}_{N,k}\mathcal{A}_{N,k}}(\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})u^k_{\mathcal{A}_{N,k}}\\
&=\alpha_N\sum_{k=1}^K (u^k_{\mathcal{A}_{N,k}})^T L'_{N,\mathcal{A}_{N,k}}(\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})+\frac{1}{2}\alpha_N^2\sum_{k=1}^K (u^k_{\mathcal{A}_{N,k}})^T (L''_{N,\mathcal{A}_{N,k}\mathcal{A}_{N,k}}(\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})-\bar{L}''_{N,\mathcal{A}_{N,k}\mathcal{A}_{N,k}}(\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}}))u^k_{\mathcal{A}_{N,k}}\\
&+\frac{1}{2}\alpha_N^2\sum_{k=1}^K (u^k_{\mathcal{A}_{N,k}})^T \bar{L}''_{N,\mathcal{A}_{N,k}\mathcal{A}_{N,k}}(\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})u^k_{\mathcal{A}_{N,k}}\\
&\geq \frac{1}{2}\alpha_N^2\sum_{k=1}^K (u^k_{\mathcal{A}_{N,k}})^T \bar{L}''_{N,\mathcal{A}_{N,k}\mathcal{A}_{N,k}}(\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})u^k_{\mathcal{A}_{N,k}}-\alpha_N K(\max_k c_{1,\eta}\|u^k_{\mathcal{A}_{N,k}}\|_2\sqrt{q_{N,k}\frac{\log N}{N}})\\
&-\frac{1}{2}\alpha_N^2 K(\max_k c_{2,\eta}\|u^k_{\mathcal{A}_{N,k}}\|_2^2 q_{N,k}\sqrt{\frac{\log N}{N}}).
\end{aligned}
$$

Here the first equality is due to the second order expansion of the loss function and the inequality is due to Lemma B.2. For sufficiently large $N$, by assumption that $\lambda_{N,k}\sqrt{N/\log N}\to\infty$ and $q_{N,k}=o(\sqrt{N/\log N})$, the second term in the last line above is $o(\alpha_N\sqrt{q_{N,k}}\lambda_{N,k})=o(\alpha_N^2)$; the last term is $o(\alpha_N^2)$. Therefore, for sufficiently large $N$

$$
\begin{aligned}
Q_N(\bar{\boldsymbol{\beta}}+\alpha_N u,\boldsymbol{\mathcal{X}},\{\lambda_{N,k}\}_{k=1}^K)-Q_N(\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}},\{\lambda_{N,k}\}_{k=1}^K)&\geq \frac{1}{2}\alpha_N^2\sum_{k=1}^K (u^k_{\mathcal{A}_{N,k}})^T \bar{L}''_{N,\mathcal{A}_{N,k}\mathcal{A}_{N,k}}(\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})u^k_{\mathcal{A}_{N,k}}\\
&-K\alpha_N^2\max_k C_k\\
&\geq \frac{1}{2}\alpha_N^2 K\min_k\big((u^k_{\mathcal{A}_{N,k}})^T \bar{L}''_{N,\mathcal{A}_{N,k}\mathcal{A}_{N,k}}(\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})u^k_{\mathcal{A}_{N,k}}\big)\\
&-K\alpha_N^2\max_k C_k,
\end{aligned}
$$

with probability at least $1-O(N^{-\eta})$. By Lemma B.1., for each $k$, $(u^k_{\mathcal{A}_{N,k}})^T \bar{L}''_{N,\mathcal{A}_{N,k}\mathcal{A}_{N,k}}(\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})u^k_{\mathcal{A}_{N,k}}\geq\Lambda^L_{\min}\|u^k_{\mathcal{A}_{N,k}}\|_2^2=\Lambda^L_{\min}(C_k)^2$. So, if we choose $\min_k C_k$ and $\max_k C_k$ such that the upper bound is minimized, then for $N$ sufficiently large, the following holds

$$
\inf_{u:u_{(\mathcal{A}_{N,k})^c}=0,\|u^k\|_2=C_k,k=1,\ldots,K} Q_N(\bar{\boldsymbol{\beta}}+\alpha_N u,\boldsymbol{\mathcal{X}},\{\lambda_{N,k}\}_{k=1}^K)>Q_N(\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}},\{\lambda_{N,k}\}_{k=1}^K),
$$

with probability at least $1-O(N^{-\eta})$, which means any solution to the problem defined in (4.9) is within the disc $\{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}\|_2\leq\alpha_N\|u\|_2\leq\alpha_N\sqrt{K}\max_k C_k\}$ with probability at least $1-O(N^{-\eta})$.

□

**Lemma 4.8.4.** *Assuming conditions of Theorems 1. Then there exists a constant $C_2(\bar{\boldsymbol{\beta}}) > 0$, such that for any $\eta > 0$, for sufficiently large $N$, the following event holds with probability at least $1 - O(N^{-\eta})$: if for any $\boldsymbol{\beta} \in S = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}\|_2 \geq C_2(\bar{\boldsymbol{\beta}})\sqrt{K}\max_k \sqrt{q_{N,k}}\lambda_{N,k}, \boldsymbol{\beta}_{\mathcal{A}_N^c} = 0\}$, then $\|L'_{N,\mathcal{A}_N}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}})\|_2 > \sqrt{K}\max_k \sqrt{q_{N,k}}\lambda_{N,k}$.*

*proof of Lemma B.4.* Let $\alpha_N = \max_k \sqrt{q_{N,k}}\lambda_{N,k}$. For $\boldsymbol{\beta} \in S$, we have $\boldsymbol{\beta} = \bar{\boldsymbol{\beta}} + \alpha_N u$, with $u_{(\mathcal{A}_N)^c}$ and $\|u\|_2 \geq C_2(\bar{\boldsymbol{\beta}})$. Note that by Taylor expansion of $L'_{N,\mathcal{A}_N}(\bar{\boldsymbol{\mathcal{W}}}, \boldsymbol{\beta}, \boldsymbol{\mathcal{X}})$ at $\bar{\boldsymbol{\beta}}$

$$\begin{aligned}
L'_{N,\mathcal{A}_N}(\bar{\boldsymbol{\mathcal{W}}}, \boldsymbol{\beta}, \boldsymbol{\mathcal{X}}) &= L'_{N,\mathcal{A}_N}(\bar{\boldsymbol{\mathcal{W}}}, \boldsymbol{\beta}, \boldsymbol{\mathcal{X}}) + \alpha_N L''_{N,\mathcal{A}_N\mathcal{A}_N}(\bar{\boldsymbol{\mathcal{W}}}, \boldsymbol{\beta}, \boldsymbol{\mathcal{X}})u_{\mathcal{A}_N} \\
&= L'_{N,\mathcal{A}_N}(\bar{\boldsymbol{\mathcal{W}}}, \boldsymbol{\beta}, \boldsymbol{\mathcal{X}}) + \alpha_N\big(L''_{N,\mathcal{A}_N\mathcal{A}_N}(\bar{\boldsymbol{\mathcal{W}}}, \boldsymbol{\beta}, \boldsymbol{\mathcal{X}}) - \bar{L}''_{N,\mathcal{A}_N\mathcal{A}_N}(\bar{\boldsymbol{\beta}})\big)u_{\mathcal{A}_N} \\
&\quad + \alpha_N \bar{L}''_{N,\mathcal{A}_N\mathcal{A}_N}(\bar{\boldsymbol{\beta}})u_{\mathcal{A}_N}.
\end{aligned}$$

By triangle inequality and similar proof strategies as in Lemma B.3., for sufficiently large $N$

$$\begin{aligned}
\|L'_{N,\mathcal{A}_N}(\bar{\boldsymbol{\mathcal{W}}}, \boldsymbol{\beta}, \boldsymbol{\mathcal{X}})\|_2 &\geq \|L'_{N,\mathcal{A}_N}(\bar{\boldsymbol{\mathcal{W}}}, \boldsymbol{\beta}, \boldsymbol{\mathcal{X}})\|_2 + \alpha_N\|L''_{N,\mathcal{A}_N\mathcal{A}_N}(\bar{\boldsymbol{\mathcal{W}}}, \boldsymbol{\beta}, \boldsymbol{\mathcal{X}})u_{\mathcal{A}_N} - \bar{L}''_{N,\mathcal{A}_N\mathcal{A}_N}(\bar{\boldsymbol{\beta}})u_{\mathcal{A}_N}\|_2 \\
&\quad + \alpha_N\|\bar{L}''_{N,\mathcal{A}_N\mathcal{A}_N}(\bar{\boldsymbol{\beta}})u_{\mathcal{A}_N}\|_2 \\
&\geq \alpha_N\|\bar{L}''_{N,\mathcal{A}_N\mathcal{A}_N}(\bar{\boldsymbol{\beta}})u_{\mathcal{A}_N}\|_2 + o(\alpha_N)
\end{aligned}$$

with probability at least $1 - O(N^{-\eta})$. By Lemma B.1., $\|\bar{L}''_{N,\mathcal{A}_N\mathcal{A}_N}(\bar{\boldsymbol{\beta}})u_{\mathcal{A}_N}\|_2 \geq \Lambda^L_{\min}(\bar{\boldsymbol{\beta}})\|u_{\mathcal{A}_N}\|_2$. Therefore, taking $C_2(\bar{\boldsymbol{\beta}})$ to be $1/\Lambda^L_{\min}(\bar{\boldsymbol{\beta}}) + \epsilon$ completes the proof. □

*proof of Theorem 1.* By the Karush-Kuhn-Tucker condition, for any solution $\hat{\boldsymbol{\beta}}$ of (4.9), it satisfies $\|L'_{N,\mathcal{A}_{N,k}}(\boldsymbol{\mathcal{W}}, \hat{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}})\|_\infty \leq \lambda_{N,k}$. Thus,

$$\begin{aligned}
\|L'_{N,\mathcal{A}_N}(\boldsymbol{\mathcal{W}}, \hat{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}})\|_2 &\leq \sqrt{K}\max_k \|L'_{N,\mathcal{A}_{N,k}}(\boldsymbol{\mathcal{W}}, \hat{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}})\|_2 \\
&\leq \sqrt{K}\max_k \sqrt{q_{N,k}}\|L'_{N,\mathcal{A}_{N,k}}(\boldsymbol{\mathcal{W}}, \hat{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}})\|_\infty \\
&\leq \sqrt{K}\max_k \sqrt{q_{N,k}}\lambda_{N,k}.
\end{aligned}$$

Then by Lemmas B.4., for any $\eta > 0$, for $N$ sufficiently large, all solutions of (4.9) are inside the disc $\{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}\|_2 \leq C_2(\bar{\boldsymbol{\beta}})\max_k \sqrt{q_{N,k}}\lambda_{N,k}, \boldsymbol{\beta}_{\mathcal{A}_N^c} = 0\}$ with probability at least $1 - O(N^{-\eta})$. If we further assume that $\min_{(i,j)\in\mathcal{A}_{N,k}} |\bar{\boldsymbol{\beta}}_{i,j}| \geq 2C(\bar{\boldsymbol{\beta}})\max_k \sqrt{q_{N,k}}\lambda_{N,k}$ for each $k$, then

$$\begin{aligned}
&1-O(N^{-\eta}) \\
&\quad \leq P_{\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}}}(\|\hat{\boldsymbol{\beta}}^{\mathcal{A}_N} - \bar{\boldsymbol{\beta}}^{\mathcal{A}_N}\|_2 \leq C_2(\bar{\boldsymbol{\beta}})\max_k \sqrt{q_{N,k}}\lambda_{N,k}, \min_{(i,j)\in\mathcal{A}_{N,k}} |\bar{\boldsymbol{\beta}}_{i,j}| \geq 2C(\bar{\boldsymbol{\beta}})\max_k \sqrt{q_{N,k}}\lambda_{N,k}, \forall k) \\
&\quad \leq P_{\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}}}(\text{sign}(\hat{\boldsymbol{\beta}}^{\mathcal{A}_{N,k}}_{i_k j_k}) = \text{sign}(\bar{\boldsymbol{\beta}}^{\mathcal{A}_{N,k}}_{i_k j_k}), \forall (i_k, j_k) \in \mathcal{A}_{N,k}, \forall k).
\end{aligned}$$

□

*proof of Theorem 2.* Let $\mathcal{E}_{N,k} = \{\text{sign}(\hat{\boldsymbol{\beta}}_{i_k j_k}^{\mathcal{A}_{N,k}}) = \text{sign}(\bar{\boldsymbol{\beta}}_{i_k j_k}^{\mathcal{A}_{N,k}})\}$. Then by Theorem 1, $P_{\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}}}(\mathcal{E}_{N,k}) \geq 1 - O(N^{-\eta})$ for large $N$. On $\mathcal{E}_{N,k}$, By the KKT condition and the expansion of $L'_{N,\mathcal{A}_{N,k}}(\bar{\boldsymbol{\mathcal{W}}}, \hat{\boldsymbol{\beta}}^{\mathcal{A}_{N,k}}, \boldsymbol{\mathcal{X}})$ at $\bar{\boldsymbol{\beta}}^{\mathcal{A}_{N,k}}$

$$
\begin{aligned}
&-\lambda_{N,k}\text{sign}(\bar{\boldsymbol{\beta}}^{\mathcal{A}_{N,k}}) \\
&\quad = L'_{N,\mathcal{A}_{N,k}}(\bar{\boldsymbol{\mathcal{W}}}, \hat{\boldsymbol{\beta}}^{\mathcal{A}_{N,k}}, \boldsymbol{\mathcal{X}}) \\
&\quad = L'_{N,\mathcal{A}_{N,k}}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}^{\mathcal{A}_{N,k}}, \boldsymbol{\mathcal{X}}) + L''_{N,\mathcal{A}_{N,k}\mathcal{A}_{N,k}}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}})v_{N,k} \\
&\quad = \bar{L}''_{\mathcal{A}_{N,k}\mathcal{A}_{N,k}}v_{N,k} + L'_{N,\mathcal{A}_{N,k}}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}^{\mathcal{A}_{N,k}}, \boldsymbol{\mathcal{X}}) + (L''_{N,\mathcal{A}_{N,k}\mathcal{A}_{N,k}}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}}) - \bar{L}''_{\mathcal{A}_{N,k}\mathcal{A}_{N,k}})v_{N,k},
\end{aligned}
$$

where $v_{N,k} = \hat{\boldsymbol{\beta}}^{\mathcal{A}_{N,k}} - \bar{\boldsymbol{\beta}}^{\mathcal{A}_{N,k}}$. By rearranging the terms

$$
\begin{aligned}
&v_{N,k} = \\
&-\lambda_{N,k}[\bar{L}''_{\mathcal{A}_{N,k}\mathcal{A}_{N,k}}]^{-1}\text{sign}(\bar{\boldsymbol{\beta}}^{\mathcal{A}_{N,k}}) - [\bar{L}''_{\mathcal{A}_{N,k}\mathcal{A}_{N,k}}]^{-1}[L'_{N,\mathcal{A}_{N,k}}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}^{\mathcal{A}_{N,k}}, \boldsymbol{\mathcal{X}}) + D_{N,\mathcal{A}_{N,k}\mathcal{A}_{N,k}}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}^{\mathcal{A}_{N,k}})v_{N,k}],
\end{aligned}
\tag{4.12}
$$

where $D_{N,\mathcal{A}_{N,k}\mathcal{A}_{N,k}} = L''_{N,\mathcal{A}_{N,k}\mathcal{A}_{N,k}}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}}) - \bar{L}''_{\mathcal{A}_{N,k}\mathcal{A}_{N,k}}$. Next, for fixed $(i,j) \in \mathcal{A}^c_{N,k}$, by expanding $L'_{N,\mathcal{A}_{N,k}}(\bar{\boldsymbol{\mathcal{W}}}, \hat{\boldsymbol{\beta}}^{\mathcal{A}_{N,k}}, \boldsymbol{\mathcal{X}})$ at $\bar{\boldsymbol{\beta}}^{\mathcal{A}_{N,k}}$

$$
L'_{N,ij}(\bar{\boldsymbol{\mathcal{W}}}, \hat{\boldsymbol{\beta}}^{\mathcal{A}_{N,k}}, \boldsymbol{\mathcal{X}}) = L'_{N,ij}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}^{\mathcal{A}_{N,k}}, \boldsymbol{\mathcal{X}}) + L''_{N,ij,\mathcal{A}_{N,k}}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}^{\mathcal{A}_{N,k}}, \boldsymbol{\mathcal{X}})v_{N,k}. \tag{4.13}
$$

Then combining (4.12) and (4.13) we get

$$
\begin{aligned}
&L'_{N,ij}(\bar{\boldsymbol{\mathcal{W}}}, \hat{\boldsymbol{\beta}}^{\mathcal{A}_{N,k}}, \boldsymbol{\mathcal{X}}) \\
&= -\lambda_{N,k}\bar{L}''_{ij,\mathcal{A}_{N,k}}(\bar{\boldsymbol{\beta}}^{\mathcal{A}_{N,k}})[\bar{L}''_{\mathcal{A}_{N,k}\mathcal{A}_{N,k}}]^{-1}\text{sign}(\bar{\boldsymbol{\beta}}^{\mathcal{A}_{N,k}}) - \bar{L}''_{ij,\mathcal{A}_{N,k}}(\bar{\boldsymbol{\beta}}^{\mathcal{A}_{N,k}})[\bar{L}''_{\mathcal{A}_{N,k}\mathcal{A}_{N,k}}]^{-1}L'_{N,\mathcal{A}_{N,k}}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}^{\mathcal{A}_{N,k}}, \boldsymbol{\mathcal{X}}) \\
&+ [D_{N,ij,\mathcal{A}_{N,k}}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}^{\mathcal{A}_{N,k}}) - \bar{L}''_{ij,\mathcal{A}_{N,k}}(\bar{\boldsymbol{\beta}}^{\mathcal{A}_{N,k}})[\bar{L}''_{\mathcal{A}_{N,k}\mathcal{A}_{N,k}}]^{-1}D_{N,\mathcal{A}_{N,k}\mathcal{A}_{N,k}}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}^{\mathcal{A}_{N,k}})]v_{N,k} \\
&+ L'_{N,ij}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}^{\mathcal{A}_{N,k}}, \boldsymbol{\mathcal{X}}).
\end{aligned}
\tag{4.14}
$$

By the incoherence condition outlined in condition (A3), for any $(i,j) \in \mathcal{A}_{N,k}$,

$$
|\bar{L}''_{ij,\mathcal{A}_{N,k}}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}})[\bar{L}''_{\mathcal{A}_{N,k},\mathcal{A}_{N,k}}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}})]^{-1}\text{sign}(\bar{\boldsymbol{\beta}}_{\mathcal{A}_{N,k}})| \leq \delta < 1.
$$

Thus, following straightforwardly (with the modification that we are considering each $\mathcal{A}_{N,k}$ instead of $\mathcal{A}_N$) from the proofs of Theorem 2 of Peng et al. [2009], the remaining terms in (4.14) can be shown to be all $o(\lambda_{N,k})$, and the event $\max_{(i,j)\in\mathcal{A}^c_{N,k}} |L'_{N,ij}(\bar{\boldsymbol{\mathcal{W}}}, \hat{\boldsymbol{\beta}}^{\mathcal{A}_{N,k}}, \boldsymbol{\mathcal{X}})| < \lambda_{N,k}$ with probability at least $1 - O(N^{-\eta})$ for sufficiently large $N$. Thus, it has been proved that for sufficiently large $N$, no wrong edge will be included for each true edge set $\mathcal{A}_{N,k}$ and hence, no wrong edge will be

included in $\mathcal{A}_N = \cup_k \mathcal{A}_{N,k}$. □

*proof of Theorem 3.* By Theorem 1 and Theorem 2, with probability tending to 1, any solution to the restricted problem is also a solution to the original problem. On the other hand, by Theorem 2 and the KKT condition, with probability tending to 1, any solution to the original problem is also a solution to the restricted problem. Therefore, Theorem 3 follows. □

## 4.9 Simulated Precision Matrix

1. **AR1($\rho$)**: The covariance matrix of the form $\mathbf{A} = (\rho^{|i-j|})_{ij}$ for $\rho \in (0, 1)$.

2. **Star-Block (SB):** A block-diagonal covariance matrix, where each block's precision matrix corresponds to a star-structured graph with $(\mathbf{\Psi}_k)_{ij} = 1$. Then, for $\rho \in (0, 1)$, we have that $\mathbf{A}_{ij} = \rho$ if $(i, j) \in E$ and $\mathbf{A}_{ij} = \rho^2$ for $(i, j) \notin E$, where $E$ is the corresponding edge set.

3. **Erdos-Renyi random graph (ER):** The precision matrix is initialized at $\mathbf{A} = 0.25\mathbf{I}$, and $d$ edges are randomly selected. For the selected edge $(i, j)$, we randomly choose $\psi \in [0.6, 0.8]$ and update $\mathbf{A}_{ij} = \mathbf{A}_{ji} \to \mathbf{A}_{ij} - \psi$ and $\mathbf{A}_{ii} \to \mathbf{A}_{ii} + \psi$, $\mathbf{A}_{jj} \to \mathbf{A}_{jj} + \psi$.

# CHAPTER 5

# High-dimensional Stochastic Linear Bandit with Missing Covariates

As applications of stochastic contextual linear bandit algorithms have grown, it has become important to understand how bandit algorithms behave in high-dimensional regimes. Recent works adopted the oracle lasso convergence theory in the sequential decision-making setting, where the exploitation stage involves lasso estimation. Even when the context is fully observed, there are significant technical challenges that hinder the application of existing lasso convergence theory: 1) proving the restricted eigenvalue condition under conditionally sub-Gaussian noise and 2) accounting for the dependence between the context variables and the chosen actions. In addition, practitioners, practitioners face the additional challenges of missing values in the context vectors in real-life applications. This paper studies the effect of missing covariates on regret for stochastic linear bandit algorithms. We accommodate missing covariates using an unbiased plug-in-estimation policy. Our work provides a high-probability upper bound on the regret incurred by the proposed algorithm in terms of covariate sampling probabilities, showing that the regret degrades due to missingness by at most $\zeta_{min}^2$, where $\zeta_{min}$ is the minimum probability of observing covariates in the context vector.

## 5.1 Introduction

High-dimensional linear stochastic bandits have become of increasing interest for many applications including recommendation systems and healthcare. In order to provide algorithms with provable guarantees in the high-dimensional regime, researchers focused on solving bandit problems with linear rewards, where only a small subset of the covariates is correlated with reward. For this setting, the learner observes a context variable $X_t \in \mathbb{R}^{K\times d}$ at round t, where each arm $i$ is associated with a given feature vector $X_{t,i} \in \mathbb{R}^d$, the $i$-th row of $X_t$. Then, based on the chosen arm $a_t$ at time $t$, the learner observes a noisy reward $\hat{r}_t = X_{t,a_t}\beta^* + \varepsilon_t$ for a fixed and unknown parameter vector $\beta^* \in \mathbb{R}^d$.

For healthcare and drug discovery applications, an additional sparsity assumption is often imposed on $\beta^*$. Sparsity is an effective constraint when the feature space is a high-dimensional feature space, for which only a subset of features is correlated with the expected reward. It is thus natural to adopt the lasso framework for learning a sparse reward function $\beta^*$ with $s_0 = \|\beta^*\|_0$. The difficulty in establishing convergence of lasso bandits has been highlighted in Bastani and Bayati [2020], Kim and Paik [2019], and Oh et al. [2021].

The difficulty in proving convergence and performing regret analysis is that the observation noise $\varepsilon_t$ associated with successive pulls of the chosen arms is no longer i.i.d. As first addressed in Bastani and Bayati [2020], under mild conditions, the sequence $\varepsilon_t X_t$ is a Martingale difference sequence. This allows them to prove tail inequalities for the lasso estimator under a conditionally independent subgaussian noise assumption. Kim and Paik [2019] used these Martingale results to perform regret analysis of the Doubly-Robust (DR) Lasso bandit, and Oh et al. [2021] applied them to the sparse agnostic (SA) Lasso bandit.

These convergence results relied on technical conditions, namely the compatibility condition and the restricted eigenvalue condition. In order to establish these conditions in bandit problems, the oracle lasso convergence theory in Van De Geer et al. [2009] was modified via a Martingale concentration inequality. This required positive-definiteness of the context covariance matrix, a condition that is often violated when there are missing values in the observed contexts. This paper relaxes this requirement, allowing us to prove convergence when there are missing values in the context variables in stochastic linear bandit problems.

As Tewari and Murphy [2017] point out, missing values are common in contextual bandits, motivating this work from a practical standpoint. For example, in a clinical application of warfarin dosing with patient data [Consortium, 2009], missing values for certain genotypes are prevalent in the context vectors. These are often imputed based on demographic or other genotype information. Even after imputation, there can remain a few missing covariates.

This work adopts the regression with missing data framework from Loh and Wainwright [2012] to establish tight convergence for stochastic linear bandits with missing covariates. Specifically, we adopt the missing completely at random (MCAR) model. In the missing data model, variables in the context vector $X_{t,i} \in \mathbb{R}^d$ of the $i$-th arm at time $t$ are observed with probabilities $\boldsymbol{\zeta} = [\zeta_1, \ldots, \zeta_d] \in [0,1]^d$. As a result, we observe the context with missing entries $Z = X \odot U$, where $\odot$ is a Hadamard product and the missing indicator $U_{t,i} \in \{0,1\}^d$ based on $1 - \boldsymbol{\zeta}$ is assumed to be independent of both $X_{t,i}$ and the observation noise $\varepsilon_t$.

In the high-dimensional statistics literature, Loh and Wainwright [2012] provided a plug-in-estimator for linear regression problems with additive and/or multiplicative noise. Fan et al. [2019] tackled a similar problem for estimating sparse precision matrices with missing covariates. We use these estimators to extend the SA Lasso bandit by Oh et al. [2021] to the noisy and missing data

setting. The technical challenges that we overcome are the following.

- *Optimization convergence:* When there are missing covariates, the lasso estimation problem becomes non-convex. Furthermore, the unbiased estimator of the covariance matrix loses positive semi-definiteness and has negative eigenvalues. Thus, we must deal with the thorny problem of how a local optimum of the lasso optimization affects the regret analyses of the stochastic linear bandit algorithms.

- *Statistical convergence:* When there are missing covariates, the negative eigenvalues of the covariance matrix do not allow us to use the compatibility condition.

Both of these challenges can be addressed once we address the conditional dependence of the reward noise $\epsilon_t$. To the best of our knowledge, our work is the first to propose a lasso bandit with missing covariates and to provide theoretical guarantees. We combine the results of Loh and Wainwright [2012] in our bandit problem using martingale methods to establish convergence. Our work establishes restricted lower- and upper-restricted eigenvalue (RE) conditions on the adjusted sample covariance matrix when the observation noises are adapted to the past observations. We show that missingness in covariates inflates the previous regret bounds by a factor inversely proportional to the squared minimum sampling probability $\zeta_{min}^2$, which is reversely proportional to missingness.

### 5.1.1 Related Work

**Sparse Linear Bandit:** Interest in sparse linear bandits for high-dimensional contexts began with Abbasi-Yadkori et al. [2012], Carpentier and Munos [2012], and continued with Gilton and Willett [2017] and Bastani and Bayati [2020]. Recently, Kim and Paik [2019] applied sparse linear structure to the stochastic linear bandit problem and incorporated a doubly-robust technique to prove convergence. The regret analysis in Kim and Paik [2019] mainly adopts the procedure in Bastani and Bayati [2020] that extended the standard lasso convergence results to online regression with non-i.i.d. samples. Oh et al. [2021] introduced SA Lasso bandit under a balanced covariance assumption to circumvent the dependency problem.

**Missing data in regression and covariance estimation:** Traditional methods introduced by Städler and Bühlmann [2012] worked with the EM algorithm to perform statistical inference for missing data. However, even in the batch setting when they do converge, EM algorithms often converge slowly. Loh and Wainwright [2012] developed M-estimators that cope with missing and corrupted data by simply adjusting the sample covariance matrix. The adjusted sample covariance estimates, however, are not necessarily positive semi-definite, which makes the likelihood non-convex. Loh and Wainwright [2012] proved that the local optima of the non-convex lasso problem have comparable mean squared error as the global optimum. Thus, a simple projected

gradient algorithm for the non-convex lasso objective under missing data is enough to guarantee the convergence of their estimators.

## 5.2 Problem Setup

We address the missing context covariate problem by incorporating an adjusted covariance matrix during the lasso step and analyze the resulting regret analyses for a modified SA Lasso bandit. We start with a precise description of the linear contextual bandit problem under missing data and motivate the proposed estimator.

**Missing covariates:** In a sparse stochastic linear bandit, the reward for pulling the $i$-th arm at time $t$ is of the form $r_{t,i} = X_{t,i}\beta^*$ for $i \in [K]$ given the covariates $X_{t,i} \in \mathbb{R}^d$, called context variables. Over the time steps $[T]$, the learner estimates the unknown regression parameter $\beta^* \in \mathbb{R}^d$, which is assumed to be sparse. In a typical stochastic linear bandit problem, after pulling arm $a_t$, we observe a reward $\hat{r}_t$, linked with context vector $X_{t,a_t} \in \mathbb{R}^d$, via the noisy linear model

$$\hat{r}_t = X_{t,a_t}\beta^* + \varepsilon_t \qquad t \in [T], a_t \in [K], \tag{5.1}$$

where $\varepsilon_t \in \mathbb{R}$ is the observation noise independent of $X_{t,i}$. Instead of directly observing $X_{t,i}$, we observe $Z_{t,i} = [Z_{t,i1}, \ldots, Z_{t,id}]$ with missing entries defined as follows

$$Z_{t,ij} = \begin{cases} X_{t,ij} & \text{if the entry is not missing} \\ 0 & \text{if the entry is missing} \end{cases}. \tag{5.2}$$

Equivalently, the learner observes the context for the $i$-th arm at time $t$ as $Z_{t,i} = X_{t,i} \odot U_{t,i}$, where $U_{t,i} \in \{0,1\}^d$ and $\odot$ is the Hadamard product. Each entry $U_{t,ij}$ is an independent Bernoulli random variable with sampling probability parameter $\zeta_j$, the probability of observing the $j$-th covariate of the context vector $X_{t,i}$ The estimation goal is to recover $\beta^*$ as we receive the context vectors $Z_{t,i}$ with missing entries.

**Multi-armed bandit setting with missing covariates:** Based on the observed contexts with missing entries, the learner is repeatedly faced with the problem of deciding which of $K$ available arms to pull based on the observed contexts $Z_{t,i} = X_{t,i} \odot U_{t,i} \in \mathbb{R}^d$, $i \in [K]$. That is, at time $t$, we observe $\boldsymbol{Z}_t \in \mathbb{R}^{K\times d}$ matrix-variate data and the missingness pattern $\boldsymbol{U}_t \in \mathbb{R}^{K\times d}$, where the $i$-th row of these variables corresponds to the $i$-th arm. Based on the context variables for each arm, the learner pulls an arm and incurs a reward. Note that when $\zeta_j = 1$ for all covariates $j \in [d]$, our problem setting is the same as the fully observed case of Kim and Paik [2019] and Oh et al. [2021].

When missing values exist, we incorporate the observed missingness pattern $\boldsymbol{U}_t$ to pull the arm

maximizing the estimated reward

$$a_t = \arg\max_{i \in [K]} (Z_{t,i} \oslash \hat{\boldsymbol{\zeta}}) \hat{\beta}_{t-1}, \qquad t \in [T], \tag{5.3}$$

where $\hat{\boldsymbol{\zeta}} \in \mathbb{R}^d$ is an estimate of the sampling probabilities based on the $U_t$'s, and $\hat{\beta}_t$ is an estimate of $\beta^*$. The policy defined by (5.3) is called the plug-in-estimation policy.

We define the optimal arm at time $t$ as

$$a_t^* = \arg\max_{1 \le i \le K} X_{t,i} \beta^*, \qquad t \in [T], \tag{5.4}$$

and the *regret(t)* as the difference between the expected reward of the optimal arm and the expected reward of the chosen arm at time $t$.

$$\begin{aligned} regret(t) &= \mathbb{E}[r_{t,a_t^*} - r_{t,a_t} | \{X_{i,t}\}_{i=1}^K, a_t, U_t] \\ &= X_{t,a_t^*} \beta - X_{t,a_t} \beta, \qquad t \in [T], \end{aligned}$$

where $r_{t,a_t^*}$ and $r_{t,a_t}$ are the maxima achieved in (5.3) and (5.4). The learner aims to minimize the cumulative regret over $T$ steps.

For the rest of the paper, we define the filtration $\mathcal{F}_{t-1}$ as the union of all observations up to time $t-1$ including rewards, missingness patterns, and contexts:

$$\mathcal{F}_{t-1} = \{(Z_\tau, \hat{r}_{\tau,a_\tau}, U_\tau)\}_{\tau=1}^{t-1}.$$

Given $\mathcal{F}_{t-1}$, the learner selects the arm $a_t$ according to the current estimate $\hat{\beta}_t$.

## 5.3 Lasso bandit with missing covariates

We introduce covariate missingness into the SA Lasso bandit [Oh et al., 2021] when the contexts are corrupted with missing values. We analyze the regret when the plug-in-estimation policy (5.3) is used in the modified lasso bandit algorithm.

### 5.3.1 Bandit with missing covariates

Compared to Kim and Paik [2019], where the uneven sampling of the covariates is addressed by taking the average of the contexts at each round and calculating the corresponding pseudo-rewards, Oh et al. [2021] introduced the SA Lasso bandit by making an additional assumption on the distribution of the contexts such that the covariance matrix $\Sigma_t = \mathbb{E}[X_t^T X_t | \mathcal{F}_{t-1}]$ behaves

sufficiently well enough to converge to the marginal context covariance matrix $\Sigma = \mathbb{E}[X_t^T X_t]$. More specifically, Oh et al. [2021] showed that

$$\sum_{i=1}^{k} \mathbb{E}_{\mathcal{X}_t}\left[X_{t,i}^T X_{t,i} \mathbf{1}(X_{t,i} = \arg\max_{X \in \mathcal{X}_t} X\beta^*)\right] \succcurlyeq (2\nu C_{\mathcal{X}})^{-1}\Sigma,$$

under the balanced covariance assumption (Assumption.6) stated later in this paper, where $\boldsymbol{X}_t = [X_{t,1}, \ldots, X_{t,K}]^T$ (See Lemma 3 and Lemma 10 of Oh et al. [2021] for more details). This paper will show that under the balanced covariance assumption, we can achieve a similar result as Oh et al. [2021] even in the presence of missing covariates.

For the SA Lasso bandit with missing covariates algorithm, we directly use the observed rewards $\hat{r}_{t,a_t}$ and the incompletely observed contexts $Z_{t,a_t} = X_{t,a_t} \odot U_{t,a_t}$. Thus, in the case of our modification, we define $\mathbf{Z}_t = [Z_{1,a_1}, \ldots Z_{t,a_t}] \in \mathbb{R}^{t \times d}$ and $\boldsymbol{r}_t = [\hat{r}_1, \ldots, \hat{r}_t] \in \mathbb{R}^t$, where $Z_{\tau,a_\tau} = X_{\tau,a_\tau} \odot U_{\tau,a_\tau}$ and $\hat{r}_\tau = X_{\tau,a_\tau}\beta + \varepsilon_\tau$.

### 5.3.2 Lasso estimation with adjusted covariance matrix

Based on the observed $\mathbf{Z}_t$, we optimize

$$\hat{\beta}_t \in \arg\min_{\|\beta\|_1 < R} \left\{ \frac{1}{2}\beta^T \hat{\Gamma}_{miss,t}\beta - \langle \hat{\gamma}_{miss,t}, \beta \rangle + \eta_t \|\beta\|_1 \right\} \tag{5.5}$$

where $\|\cdot\|_1$ is an $\ell_1$ norm and

$$\begin{aligned} \hat{\Gamma}_{miss,t} &= \left(\frac{1}{t}\mathbf{Z}_t^T \mathbf{Z}_t\right) \oslash \hat{M} \\ \hat{\gamma}_{miss,t} &= \left(\frac{1}{t}\mathbf{Z}_t^T \mathbf{r}_t\right) \oslash \hat{\boldsymbol{\zeta}} \\ \hat{M}_{ij} &= \begin{cases} \hat{\boldsymbol{\zeta}}_i & \text{if } i = j \\ \hat{\boldsymbol{\zeta}}_i \hat{\boldsymbol{\zeta}}_j & \text{if } i \neq j \end{cases} \end{aligned}, \tag{5.6}$$

where $\hat{\boldsymbol{\zeta}} \in \mathbb{R}^d$ is the sampling probability of each of the covariates. This type of lasso estimator, under noisy and missing data, was first introduced in Loh and Wainwright [2012] for the regression problem. Our analysis adopts this estimator for the high-dimensional stochastic linear bandit problem.

While the application of Loh and Wainwright [2012]'s approach seems simple enough, several challenges arise due to the sequential-decision making setting. As the noise $\varepsilon_t$ is not i.i.d., we cannot directly apply the convergence results from Loh and Wainwright [2012] to the lasso bandit with covariate missingness.

We show that the optimal regularization path depends on the minimum sampling probability of the covariates $\zeta_{min} = \min_j \hat{\boldsymbol{\zeta}}_j$. Our theorem shows that $\eta_t$ should scale with $\frac{\log d}{t \cdot \zeta_{min}^2}$. This regularization path agrees with the noiseless setting in Oh et al. [2021]. Intuitively, the effective sample size of the covariates is $t \cdot \zeta_{min}^2$, which is the number of time steps needed to reliably estimate the off-diagonal entries of $\mathbb{E}[X_t^T X_t | \mathcal{F}_{t-1}]$. It will be easily seen that all of our results will reduce to Oh et al. [2021] for the case that $\zeta_j = 1$ for all $j \in [K]$.

## 5.4 Algorithm

Algorithm 3 solves the lasso bandit problem under the covariate missingness. The key differences compared to the fully-observed counterpart are 1) the use of adjusted plug-in estimators $\hat{\Gamma}_{miss,t}$ and $\hat{\gamma}_{miss,t}$ and 2) the theoretically justified regularization parameter $\eta_t$. An additional tuning parameter $R$ is introduced to the non-convexity of the problem (5.5) to constrain $\beta$ in an $\ell_1$ ball and is motivated by a similar approach proposed by Loh and Wainwright [2012] and Rudelson et al. [2017].

**Algorithm 3:** SA Lasso bandit with missing covariates

**Input:** $\eta_1$, $R$
Initialize $\beta_0 = 0$, $\hat{\boldsymbol{\zeta}}_0 = 1$
**for** $t = 1, \ldots, T$ **do**
  Observe contexts $Z_t \sim \mathcal{P}_{K \times d}$
  and the missing pattern $U_t$
  Update $\hat{\boldsymbol{\zeta}}_t = \hat{\boldsymbol{\zeta}}_{t-1} + \frac{1}{t}\left(\frac{1}{K}\sum_{i=1}^{K} U_{t,i} - \hat{\boldsymbol{\zeta}}_{t-1}\right)$
  Pull arm $a_t = \arg\max_{i \in [k]} (Z_{t,i} \oslash \hat{\zeta}_t)\hat{\beta}_t$
  Observe $\hat{r}_t$ for the arm $a_t$
  Update $\eta_t = \eta_1 \sqrt{\frac{4\log(t\,\zeta_{min}^2) + \log d}{t\,\zeta_{min}^2}}$
  Updated $\hat{\Gamma}_{miss,t}$ and $\hat{\gamma}_{miss,t}$ based on (5.6)
  Update $\hat{\beta}_t$ based on (5.5)
**end**

For Algorithm 3, we consider the constrained program as introduced in Loh and Wainwright [2015a], given $\hat{\Gamma}_{miss,t}$ and $\hat{\gamma}_{miss,t}$,

$$\hat{\beta}_t \in \underset{\|\beta\|_1 \leq R}{\arg\min} \left\{ \frac{1}{2} \beta^T \widehat{\Gamma}_{miss,t} \beta - \langle \widehat{\gamma}_{miss,t}, \beta \rangle + \eta_t \|\beta\|_1 \right\} \quad , \tag{5.7}$$

for some constant $R = b\sqrt{s_0}$, where $b = \max_{i \in [d]} \beta^*$. As we do not have a priori knowledge of the sparsity level $s_0$, $R$ is user-defined parameter. While typical gradient descent methods fail to converge to a global optimum due to local minima, Loh and Wainwright [2012] proved that a

simple projected gradient descent algorithm for (5.7) converges within the statistical tolerance of the global optimum even when $\widehat{\Gamma}_{miss,t} \not\succcurlyeq 0$.

To iteratively solve (5.7), we apply the following Lagrangian update with $\ell_1$-ball penalty parameter $\mu$ and sparsity parameter $\eta_t$.

$$\beta^{i+1} = \underset{\|\beta\|_1 \leq R}{\arg\min} \Big\{ \mathcal{L}(\beta^i) + \langle \nabla \mathcal{L}(\beta^i), \beta - \beta^i \rangle + \frac{\mu}{2}\|\beta - \beta^i\|_2^2 + \lambda_t \|\beta\|_1 \Big\}, \tag{5.8}$$

where $\mu > 0, \eta_t > 0$, $\mathcal{L}(\beta) = \frac{1}{2}\beta^T \widehat{\Gamma}_{miss,t}\beta - \langle \widehat{\gamma}_{miss,t}, \beta \rangle$ is the likelihood, and $\nabla \mathcal{L}(\beta) = \widehat{\Gamma}_{miss,t}\beta - \widehat{\gamma}_{miss,t}$ is its gradient.

## 5.5 Regret analysis under missing data

We derive an upper bound on the regret of the SA Lasso bandit with missing covariates defined in Section 5.3.1. We emphasize that the regret analysis is not limited to our problem but can be derived for any linear bandits using Proposition 5.5.1. This key result extends Proposition 1 in Bastani and Bayati [2020] to the case where there are missing values. We first state the standard assumptions we make to derive the upper bound for the cumulative regret and the relevant definitions that will extend the compatibility assumption to the case of missing covariates.

**Assumption 1** (Feature set and parameter)**.** *There exists a positive constant $x_{max}$ such that $\|X_{t,i}\|_2 \leq x_{max}$ for all $X_{t,i} \in \mathbb{R}^d$ and a positive constant $b$ such that $\|\beta^*\|_2 \leq b$ and $\|\beta^*\|_0 = s_0$.*

**Assumption 2** (i.i.d. context)**.** *The context variables $X_t \in \mathbb{R}^{K \times d}$ are i.i.d. and follow matrix-variate distribution $\mathcal{P}_X$ at every time $t$:*

**Assumption 3** (Sub-Gaussian error)**.** *The error $\varepsilon_t = \hat{r}_{t,a_t} - X_{t,a_t}\beta^*$ is $\sigma_\varepsilon$-sub-Gaussian adapted to $\mathcal{F}_{t-1}$ for some $\sigma_\varepsilon > 0$. In other words, for every $\alpha \in \mathbb{R}$, $\mathbb{E}[e^{\alpha\varepsilon_t}|\mathcal{F}_{t-1}] \leq e^{\sigma_\varepsilon^2\alpha^2/2}$.*

Assumption 1-3 are standard assumptions in the stochastic linear bandit literature [Bastani and Bayati, 2020, Oh et al., 2021, Kim and Paik, 2019].

We additionally assume the compatibility condition on the true Gram matrix $\Sigma := \frac{1}{K}\mathbb{E}[X^T X]$. This is a standard assumption in high-dimensional regression literature [Van De Geer et al., 2009] and stochastic linear bandit problems [Wang et al., 2018, Bastani and Bayati, 2020, Kim and Paik, 2019, Oh et al., 2021]. Before we introduce the compatibility condition, we first define the active set $S_0 = \{j : \beta_j^* \neq 0\}$ as the set of indices that correspond to non-zero values of $\beta_j^*$. Thus, the true

$\beta^*$ can be divided into the following

$$\beta^*_{j,S_0} = \beta_j \mathbb{1}(j \in S_0) \text{ and } \beta^*_{j,S_0^c} = \beta_j \mathbb{1}(j \notin S_0)$$

Let $\mathbb{C}(S_0)$ be the set of vectors $\beta \in \mathbb{R}^d$ defined as

$$\mathbb{C}(S_0) = \{\beta \in \mathbb{R}^d | \|\beta_{S_0^c}\|_1 < 3\|\beta_{S_0}\|_1\}. \tag{5.9}$$

Then, we can define the compatibility condition.

**Assumption 4** (Compatibility Condition)**.** *For an active set $S_0$, there exists a compatibility constant $\phi^2 > 0$ such that*

$$\phi_0^2 \|\beta_{S_0}\|_1^2 \leq s_0 \beta^T \Sigma \beta \qquad \forall \beta \in \mathbb{C}(S_0)$$

The compatibility condition generalizes the positive-definite assumption on the population covariance matrix and ensures that a Lasso estimator will converge to the true parameter $\beta$ with high probability as the sample size grows to infinity. While the compatibility condition seems technical at first sight, it allows us to bound the $\ell_1$-norm with the $\ell_2$-norm. It can be easily seen that the positive-definite assumption in OLS satisfies the compatibility condition with $\phi_0 = \sqrt{\lambda_{min}(\Sigma)}$. Despite the compatibility condition on the true $\Sigma$, our unbiased estimator for $\Sigma$ always contains negative eigenvalues when there are missing covariates in the context vector. Thus, we later introduce restricted strong convexity to prove convergence.

**Assumption 5** (Relaxed symmetry)**.** *For a joint distribution $\mathcal{P}_X$, there exists $\nu < \infty$ such that $\frac{\mathcal{P}_X(-\mathbf{x})}{\mathcal{P}_X(\mathbf{x})} \leq \nu$ for all $\mathbf{x} \in \mathbb{R}^d$.*

Relaxed symmetry assumption is satisfied by a wide range of distribution including the Gaussian distribution and the uniform distribution. Note that for symmetric distributions, Assumption 5 is satisfied with $\nu = 1$.

**Assumption 6** (Balanced covariance)**.** *Consider a permutation $(i_1, \ldots, i_K)$ of $(1, \ldots, K)$. For any integer $k \in \{2, \ldots, K-1\}$ and fixed vector $\beta$, there exists $C_{\mathcal{X}}$ such that*

$$\begin{aligned}&\mathbb{E}\left[X_{i_k}^T X_{i_k} \mathbb{1}(X_{i_1}\beta^* < \cdots < X_{i_K}\beta^*)\right] \\ &\preccurlyeq C_X \mathbb{E}\left[(X_{i_1}^T X_{i_1} + X_{i_K}^T X_{i_K}) \mathbb{1}(X_{i_1}\beta^* < \cdots < X_{i_K}\beta^*)\right]\end{aligned}$$

In bandit problems, the sample covariance we work with at time $t$ is $\mathbb{E}[X^T X | \mathcal{F}_{t-1}]$. As we are selecting arms $a_t = \arg\max_{i \in [K]} (Z_{t,i} \oslash \hat{\boldsymbol{\zeta}})\beta_t$ based on the current estimate $\beta_t$, our algorithm may not evenly sample from the whole distribution. As introduced in Oh et al. [2021], the balanced

covariance assumption implies that we can control the covariance matrix based on the extreme selections of the arms.

For example, if the arms are completely correlated, $C_X$ is constant independent of dimensions. In a more general setting, Oh et al. [2021] proved that the balanced covariance condition is satisfied with $C_X = \binom{K-1}{K_0}$ with $K_0 = \lceil \frac{K-1}{2} \rceil$ when the arms are independent and identically distributed from Gaussian distribution. While the balanced covariance condition was first introduced from the proof technique in Oh et al. [2021], the value of $C_X$ gives an insight into the behavior of the population covariance matrix.

As $\widehat{\Gamma}_{miss,t}$ is not positive semi-definite, we introduce lower-restricted eigenvalue (RE) and upper-RE conditions from Loh and Wainwright [2012], which are also known as restricted strong convexity and restricted strong smoothness conditions in the optimization literature [Agarwal et al., 2012, Negahban et al., 2012].

**Definition 5.5.1.** (Lower-RE condition) The matrix $\widehat{\Gamma}$ satisfies a lower restricted eigenvalue condition with curvature $\alpha_1 > 0$ and tolerance $\tau(t, d) > 0$ if

$$\beta^T \widehat{\Gamma} \beta \geq \alpha_1 \|\beta\|_2^2 - \tau(t, d) \|\beta\|_1^2 \ \ \forall \beta \in \mathbb{R}^d. \tag{5.10}$$

This condition is also used in Loh and Wainwright [2012]. In the fully observed case, the standard covariance matrix $\frac{1}{K} \boldsymbol{X}_t^T \boldsymbol{X}_t$ will satisfy the lower-RE condition with $\alpha_1 = \frac{1}{2} \lambda_{min} \left( \frac{1}{K} \boldsymbol{X}_t^T \boldsymbol{X}_t \right)$ and $\tau(t, d) \asymp \frac{\log d}{t}$ (See Loh and Wainwright [2012] for more details). This condition is useful to study the statistical aspect of the proposed method, as $\widehat{\Gamma}_{miss,t}$ has negative eigenvalues. We extend the results of Loh and Wainwright [2012] and show that the estimator for the covariance matrix with missing covariates in bandit problems satisfies the lower-RE condition.

**Definition 5.5.2.** (Upper-RE condition) The matrix $\widehat{\Gamma}$ satisfies an upper restricted eigenvalue condition with smoothness $\alpha_2 > 0$ and tolerance $\tau(t, d) > 0$ if

$$\beta^T \widehat{\Gamma} \beta \leq \alpha_2 \|\beta\|_2^2 + \tau(t, d) \|\beta\|_1^2 \ \ \forall \beta \in \mathbb{R}^d. \tag{5.11}$$

A popular example that satisfies the lower- and upper-RE condition is a Toeplitz matrix $\Sigma \in \mathbb{R}^{d \times d}$, where $\Sigma_{ij} = \rho^{|i-j|}$ with $\rho \in [0, 1)$. Such covariance structure arises from autoregressive processes, and the parameter $\rho$ determines the memory in the process. In this setting, it can be shown that the minimum eigenvalue $\lambda_{min}(\Sigma)$ is $1 - \rho > 0$ regardless of the dimension $d$, and thus a Toeplitz matrix satisfies the lower- and upper-RE condition. Thus, the sample covariance matrix for autoregressive processes will also satisfy the RE condition with high probability. Besides Toeplitz matrices, Raskutti et al. [2010] showed that a wide range of random Gaussian matrices that satisfy (5.10) and (5.11), and Rudelson and Zhou [2012] established similar results for random

matrices with dependent entries in the sub-Gaussian setting.

### 5.5.1 Lasso convergence for bandit problems with missing data

Given the observed covariates $Z_t = X_t \odot U_t$, we have the following proposition.

**Proposition 5.5.1.** *Let $X_t \in \mathbb{R}^{K\times d}$ be sub-Gaussian with parameters $(\Sigma_x, \sigma_x^2)$ and $Z_t = X_t \odot U_t$ be the missing data matrix with parameter $\boldsymbol{\zeta} = [\zeta_1, \ldots, \zeta_d] \in [0,1]^d$. Also, define $\mathcal{F}_{\tau-1}$ to be the filtration up to time $\tau - 1$ in the bandit setting with $\hat{r}_\tau = X_{\tau,a_\tau}\beta^* + \varepsilon_\tau$. Suppose $\varepsilon_\tau|\mathcal{F}_{\tau-1}$ is $\sigma_\varepsilon$-sub-Gaussian for $\tau = 1, \ldots, t$. If $t > \max\left(\frac{1}{\zeta_{min}^4}\frac{\sigma_x^4}{k^2\lambda_{min}^2(\Sigma_x)}, 1\right)\cdot s_0 \log d$, then for any vector $\beta^*$ with sparsity at most $s_0$, there is a universal positive constant $c_0$ such that any global optimum $\hat{\beta}$ of* (5.5) *with any $\|\beta^*\|_2 \le R$ and $\eta_t \ge 4\phi(\mathbb{Q}, \sigma_\varepsilon)\sqrt{\frac{\log d}{t}}$ satisfies the bound*

$$\|\widehat{\beta}_t - \beta^*\|_2 \le \frac{c_0\sqrt{s_0}}{\alpha_1}\max\left\{\phi(\mathbb{Q}, \sigma_\varepsilon)\sqrt{\frac{\log d}{t}}, \eta_t\right\}, \tag{5.12}$$

*with probability at least $1 - c_1\exp(-c_2 \log d)$ for $\alpha_1 = \frac{1}{2}\lambda_{min}(\Sigma_x)$ and $\phi(\mathbb{Q}, \sigma_\varepsilon) = c_0\frac{\sigma_x}{\sqrt{k}\zeta_{min}}(\sigma_\varepsilon + \frac{\sigma_x}{\sqrt{k}\zeta_{min}})\|\beta^*\|_2$.*

This convergence result does not follow directly from Loh and Wainwright [2012] since $\{\varepsilon_\tau Z_{\tau,a_\tau j}\}_{\tau=1}^t$ for $j \in [d]$ are not i.i.d. However, they are a martingale difference sequence adapted to the filtration $\mathcal{F}_{t-1}$, as shown in Appendix C.

The technical challenges in proving (5.12) when $\varepsilon_\tau Z_{\tau,a_\tau j}$ is a martingale difference sequence are 1) bounding $\left\|\frac{\boldsymbol{\epsilon}_t^T \mathbf{Z}_t}{t}\right\|_\infty$ and 2) proving $\|\hat{\gamma}_{miss,t} - \hat{\Gamma}_{miss,t}\beta^*\|_\infty \le \phi(\mathbb{Q}, \sigma_\varepsilon)\sqrt{\frac{\log p}{t}}$ in the bandit setting. Once these are bounded with high-probability, the rest of the proof of Proposition 5.5.1 follows the proof method of Loh and Wainwright [2012] (provided in Appendix C).

Note that Proposition 5.5.1 only bounds the statistical error of the global optimizer. As the problem (5.5) is non-convex, we also need to address the optimization error of our estimator. Based on Lemma 5.8.6 and Lemma 5.8.8, we can adopt Theorem 2 of Loh and Wainwright [2012] to obtain the following theorem.

**Theorem 5.5.1** (Theorem 2, Loh and Wainwright [2012])**.** *Denote $\psi$ as the objective function of Lagrangian program* (5.8) *with global optimum $\widehat{\beta}_t$ after applying the updates* (5.8)*. Under the conditions of Proposition 5.5.1, there are universal positive constants $(c_1, c_2)$ and a contraction coefficient $\omega \in (0,1)$, independent of $(t, d, s_0)$, such that*

$$\|\beta^i - \widehat{\beta}_t\|_2^2 \le \underbrace{c_1\|\widehat{\beta}_t - \beta^*\|_2^2}_{\delta^2} \qquad \textit{for all iterates } i \ge \mathcal{I} \tag{5.13}$$

*where* $\mathcal{I} := c_2 \log \frac{(\psi(\beta^0) - \psi(\widehat{\beta}))}{\delta^2} / \log(1/\omega)$.

The standard regret analyses of the SA bandits without missing covariates assume that $\hat{\beta}$ is the global optimum of the problem (5.7). However, in reality, as our objective function is non-convex, only a local optimum $\widetilde{\beta}$ of (5.7) may be available.

Lemma 5.8.6 and 5.8.8 allow us to adopt the optimization result from Loh and Wainwright [2012], and we can bound the regret for $t > T_0$ by

$$\begin{aligned} \|\widetilde{\beta}_t - \beta^*\|_2 &\leq \|\widetilde{\beta}_t - \hat{\beta}_t\|_2 + \|\hat{\beta}_t - \beta^*\| \\ &\leq c_0 \|\hat{\beta}_t - \beta^*\|_2 \end{aligned}$$

for some constant $c_0 > 1$. Such an extension is possible since the lower-RE condition of the covariance matrix allows our objective function to be slightly non-convex, where all local optimums are within the statistical tolerance of the problem for $t > T_0$.

### 5.5.2 Regret Analyses with missing values

In this section, we provide our regret analysis for Algorithm 3.

**Theorem 5.5.2** (SA Lasso bandit with missing values)**.** *Suppose Assumption 1, 2, 3, 4 and 5 hold. Then, for some constant $c_0, c_1 > 0$, the cumulative regret of the SA Lasso bandit with missing values is $\mathcal{O}\left(\frac{1}{\zeta_{min}^2}\sqrt{s_0 T \log(dT)}\right)$ with probability at least $1 - c_0 \exp(-c_1 \log d)$.*

As pointed out in Oh et al. [2021], the learner does not have to go through the exploration phase due to the balanced covariance assumption (Assumption 6). Compared to the fully observed setting, the regret is increased by $\frac{1}{\zeta_{min}^2}$ due to the missing covariates. This matches our intuition that extra arm pulls are needed to accurately estimate the off-diagonal entries of $\widehat{\Gamma}_{miss,t}$ when covariates are missing.

## 5.6 Simulation study

To demonstrate the benefit of incorporating the missing pattern in the observed context, we performed two sets of simulations.

1. **Convergence**: For the first set of experiments, we first compare the modified SA Lasso to existing methods as we increase the missingness.

2. **Comparison to Imputation**: As there is no systematic way of incorporating the missing values, practitioners often impute the missing values by columns. We compare our method to the SA lasso bandit with imputed context variables.

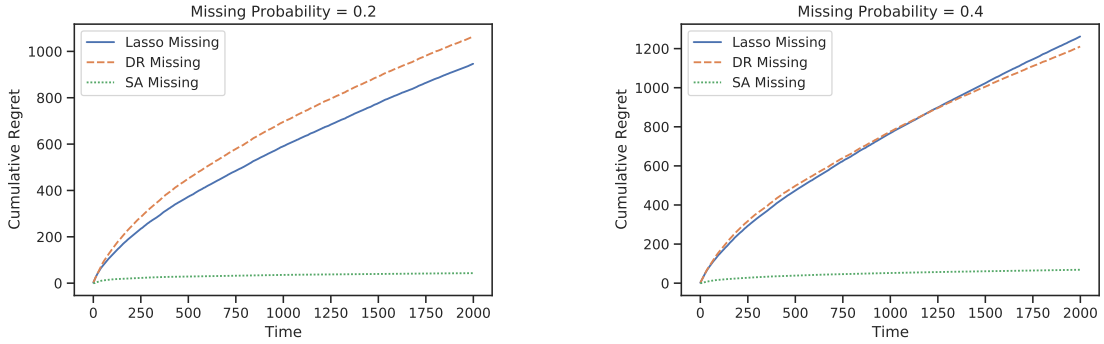



Figure 5.1: Cumulative regret over time with $1 - \zeta_j \in \{0.2, 0.3, 0.4\}$ for $j \in [K]$, $k = 20$, and $d = 200$ for the lasso bandit, the DR Lasso bandit, and the SA Lasso bandit with the adjusted estimator $\hat{\Gamma}_{t,miss}$ and $\hat{\gamma}_{t,miss}$.

### 5.6.1 Convergence Simulation

We conduct simulations to evaluate the improvement in cumulative regret based on our modifications on the lasso bandit, the DR lasso bandit, and the SA lasso bandit, presented in Figure 5.1. The details of the DR lasso modification are included in Appendix 5.8.5. We set $k = 20$, $d = 200$ and the missing probability $1-\zeta_j \in \{0.2, 0.4\}$ for all $j \in [K]$. The sparsity was set to $s_0 = c\sqrt{d}$, where $c$ is constant. At each round $t$, we generate the true contexts $X_t \in \mathbb{R}^{K\times d}$, where $X_{t,i} \sim N(0, \Sigma)$ and $\Sigma$ is a Toeplitz matrix. The learners observe $Z_t = X_t \odot U_t$, where all entries $U_{t,ij} \sim Ber(\zeta_j)$. We could have also varied the sampling probability $\zeta_j$ for each covariate, but Theorems 5.5.2 shows that only the minimum sampling probability plays a role in the convergence rate. In this setting, the Gaussian model for $X_t$ satisfies the symmetry condition (Assumption 5). Lastly, we generate $\varepsilon_t$ from the normal distribution, and the reward is observed based on the approximation $Z_t \oslash \hat{\boldsymbol{\zeta}}$ of the context $X_t$.

In order to verify the rates predicted by Theorem 5.5.2, we perform the same simulations for $\zeta_j \in [0.65, 0.9]$. Given our specified simulation of $(T, d, s_0, \zeta)$, Theorem 5.5.2 says that the regret is $\mathcal{O}\left(\frac{1}{\zeta_{min}^2}\sqrt{s_0 T \log(dT)}\right)$ for the SA Lasso bandit. Thus, the rescaled cumulative regret $\frac{\text{regret}(t)\cdot\zeta_{min}^2}{\sqrt{s_0 T \log(dT)}}$ for the SA Lasso bandit have similar values for different $\zeta_{min} \in [0.65, 0.9]$ as shown in Figure 5.2.

### 5.6.2 Comparison to Imputation

In real-life applications, practitioners are often faced with missing entries and resort to imputing the missing entries with the observed column average. The resulting sample covariance matrix with imputed covariates, however, becomes a biased estimator of the population covariance matrix. With the same set of parameters from the first experiments, we compare our method to the imputation method. Figure 5.3 shows the results for our method and imputation. For a smaller

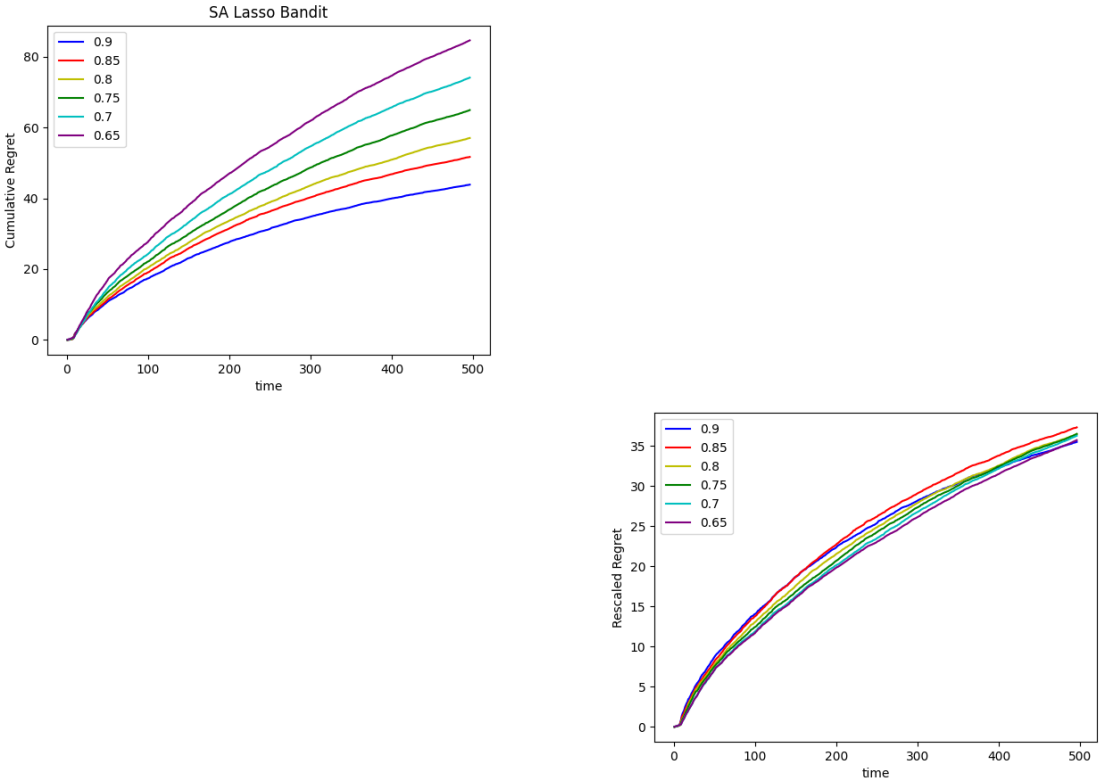



Figure 5.2: **(Top)** Cumulative regret over time with $\zeta \in [0.65, 0.9]$, $k = 20$, and $d = 200$ for the proposed method. **(Bottom)** The rescaled cumulative regret as $regret(t) \cdot \frac{\zeta_{min}^2}{\sqrt{s_0 T \log(dT)}}$, based on the rates of Theorem 5.2.

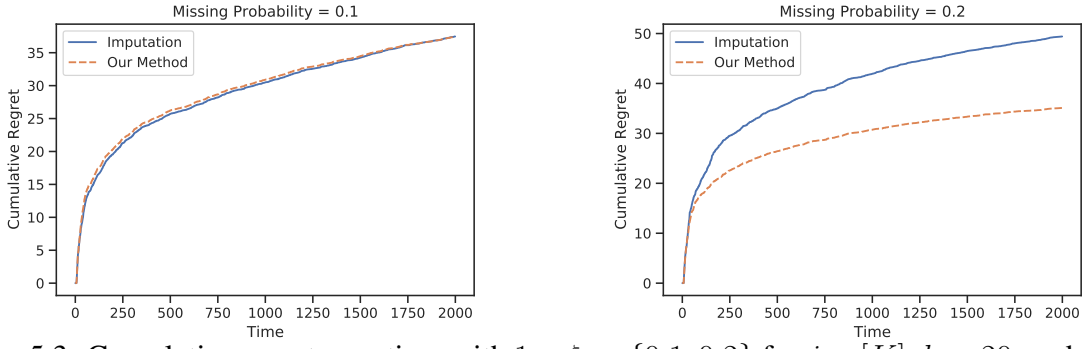



Figure 5.3: Cumulative regret over time with $1 - \zeta_j \in \{0.1, 0.2\}$ for $j \in [K]$, $k = 20$, and $d = 200$ for the proposed method and the SA Lasso Bandit with imputed covariates.

missing probability of $0.1$, the number of imputed covariates is small enough that the performance between our method and the imputation method is about the same. However, as we increase the missing probability of the covariates to $1 - \zeta_{min} = 0.2$, the difference in performance starts to become clear. With the unbiased estimator of $\hat{\Gamma}_{t,miss}$, we can quickly recover the $\beta^*$ with our method.

As a result, the resulting cumulative regret performance has an edge over the imputation method.

### 5.6.3 Case Study: Experimental Design

Understanding the interactions among microbiomes in microbial communities is important to many applications in health, ecology, and antibiotic design. Recently, Lozano et al. [2019] introduced a model microbiome community for the rhizosphere, called THOR or BFK, that combines three microbial species *Bacillus cereus (B), Flavobacterium Johnsonian (F)*, and *Pseudomonas koreensis* in order to study the complex interactions between them under different conditions (classes). We will use data from this model to illustrate the application of the proposed contextual bandit with missingness to the sequential design of experiments for discovering the gene probes that best discriminate between two experimental conditions: a BFK community having a wildtype strain of F (class 1) vs a community having a mutant strain of F (class 2).

**Gene probe Selection Problem**: Often only a few key genes among the thousands of genes play important roles in the behavior of microbial species under varying experimental conditions. However, simultaneously collecting all available gene expression data for multiple experiments could be costly for researchers. We reformulate the DNA probe selection problem as a sequential design problem using contextual bandits. The objective is to sequentially select gene probes (arms) to discover a few genes that best discriminate between the classes. We aim to establish proof of concept that such discriminative genes can be discovered sequentially without the need to sequence the entire genome at once.

**Dataset**: Experimental microbiome data was collected and processed in the lab of one of the co-authors. We performed gene sequencing on each species, yielding the three gene expression datasets $\mathbf{X}_i \in \mathbb{R}^{n_i \times (m_1+m_2)}$ for $i \in \{B, F, K\}$, where $n_i$ is the number of genes for species $i$ and $m_1$ and $m_2$ are the number of replicates (samples) for conditions (classes) 1 and 2, respectively. There are $6179$ gene probes for *Bacillus*, $5198$ genes in *Flavobacterium*, and $5864$ genes in *Pseudomonas* with $m_1 = 38$ and $m_2 = 34$.

**Bandit Formulation**: We formulate the sequential gene selection problem with the contextual bandit having the following components

- *Arms*: The arms correspond to the genomes of $B, F, K$ for the three species, denoted $\mathbf{X} \in \mathbb{R}^{k \times (m_1+m_2)}$ where $k$ represents the number of selectable DNA probes, which depends on $B$, $F$, or $K$.

- *Covariates*: The covariates of the $i$-th arm at time $t$ $\mathbf{X}_{t,i} \in \mathbb{R}^{m_1+m_2}$ are the gene expressions of the $i$-th probe for the set of samples for both experimental conditions.

- *Reward*: The reward is the observed discrimination provided by the selected gene probe

(arm). We use Welch's t-statistic, specifically the $p$-value of this statistic to measure discrimination. The reward is defined as

$$\hat{r}_t = \log\left(\frac{1 - p_{a_t}}{p_{a_t}}\right) = X_{a_t}\beta^* + \epsilon_t \tag{5.14}$$

where $p_{a_t}$ denotes the *p-value* of the Welch's test of the null hypothesis that the arm $a_t$ is a non-discriminative gene whose means are identical in each class.

**Evaluation and Results**: Based on the reward function (5.14), we apply our contextual bandit and treat the zero expression values in the data $X$ as missing entries. As the goal of the bandit problem is to select the most discriminating DNA probes at each time point, we evaluate the fraction of the probe selections that correctly lead to a statistically significant ($\alpha = 0.05$). To simulate noisy rewards in terms of *p-value*, each probe is sampled with replacement at each time $t$.

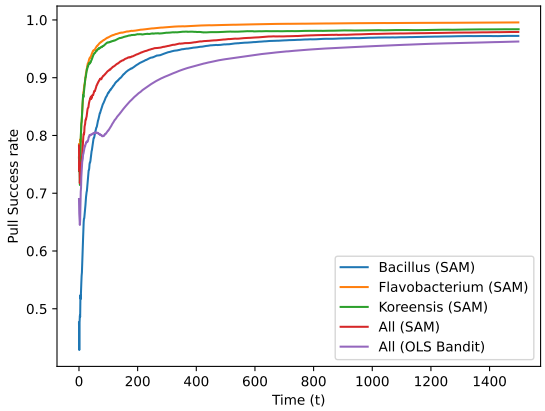



Figure 5.4: Success rate of proposed contextual bandit as measured by the fraction of probes selected at each time (arm pull) that highly discriminate between microbiome classes as measured by *p-value* ($\alpha < 0.05$) of Welch's test of significance for testing that class means are identical. Each result represents an average over 100 trials. The proposed sparse agnostic with missingness bandits (SAM) more rapidly achieve 100% success rate than the standard bandit (OLS).

Figure 5.4 shows the pull success rate for our proposed bandit. For each species, the success rate rapidly reaches around 0.95 for SAM. The difference between the class distributions is the largest for *Flavobacterium* in the overall data, and the bandit quickly discriminates this in its early set of selected probes for $F$ (shown in orange in Figure 5.4).

## 5.7 Conclusion

The missing value problem is challenging in high-dimensional stochastic linear bandit problems. Missing covariates often result from the high cost of collecting context vectors, or alternatively, sensor failures. This paper presented a modification of the SA Lasso bandit when the context covariates may be incompletely observed. We modeled the problem as missing completely at random but with the possibly different covariate missingness probabilities. Even in this simple setting, the unbiased plug-in estimators of the lasso estimation result in a non-convex objective of the lasso estimation, which was first addressed in Loh and Wainwright [2012] for regression problems and extended in this paper for bandit problems.

A natural extension to our paper is to relax the MCAR assumption. Often in clinical trials, the missing pattern itself has an implication for a given patient. For example, a patient is more likely to complete a health survey if he or she has a relevant medical condition. In such a setting, we can have an additional layer of modeling the missing probabilities based on the meta-data of the patients or the groups defined by the observed data. Then, the data are considered missing at random (MAR), which corresponds to a much broader than MCAR. This direction gives us an additional group information of the observed data to improve the convergence rate of our covariance matrix and in turn result in a faster regret convergence.

Another direction is to incorporate the noisy data into the learning framework based on errors-in-variables (EIV) models, introduced in Loh and Wainwright [2012] and Rudelson et al. [2017] for regression problems. In a bandit EIV model, the column-correlated contexts $X_t \in \mathbb{R}^{K\times t}$ at time $t$ would be observed with the row-correlated measurement noise matrix $W_t \in \mathbb{R}^{K\times d}$. The intuition behind the EIV-model for the bandit problems is to take the arm-correlated measurement noise into account. We can effectively reduce the noise by looking at the covariance structures. Thus, similar to the problem in this paper, the problem can be formulated as recovering $\beta^*$ based on the noisy context covariates $Z_t = X_t + W_t$.

Besides EIV models, matrix-variate modeling for bandit problems could be used to improve the performance of the bandit algorithms when the arms are correlated or specific group means are present in the context vectors. We hope our work promotes additional interest in theoretically exploring bandit algorithms with noisy and missing covariates.

## 5.8 Detailed Proofs for stochastic multi-armed bandit with missing covariates

### 5.8.1 Outline

The structure of the appendix is as follows. First, we present the regret analysis for Sparse Agnostic Lasso Bandit with missing covariates. In Appendix 5.8.4, we state the technical lemmas needed to complete the regret analysis proofs in this paper.

For ease of notations in the proofs, we let $\|X_{t,i}\|_2 \leq x_{max} = 1$ for all $X_{t,i} \in \mathbb{R}^d$ and $\|\beta^*\|_2 \leq b = 1$.

#### 5.8.1.1 Notations

Recall that at time $t$, we observe the contexts of each arm $Z_t = X_t \odot U_t \in \mathbb{R}^{K \times d}$, where $K$ is the number of arms and $d$ is the dimension of the context covariate vector. Based on the arm, the environment provides a noisy reward $\hat{r}_t = X_{t,a_t}\beta^* + \varepsilon_t$. At each round, the learner observes $\hat{r}_t$ and performs lasso on $\{Z_{\tau,a_\tau} \oslash \hat{\boldsymbol{\zeta}}_t, \hat{r}_\tau\}_{\tau=1}^t$, where $Z_{\tau,a_\tau}$ is the context covariate for arm $a_\tau$ at time $\tau$.

### 5.8.2 Regret analysis for SA Lasso bandit with missing contexts

The proof structure of the regret analysis is inspired by the regret analysis of Kim and Paik [2019] and Oh et al. [2021] for stochastic linear bandits but is different in two aspects.

- Kim and Paik [2019], Bastani and Bayati [2020], and Oh et al. [2021] work with positive semi-definite $\Sigma$ and thus do not have to consider the optimization error as their sub-problems are convex. With additional insight from Loh and Wainwright [2012], we work with the local approximation $\widetilde{\beta}_t$ instead of the global optimum $\hat{\beta}_t$ of the non-convex sub-problems in our analysis. The technical difficulty is addressed in Loh and Wainwright [2012], but we bring it here for regret analysis for completeness.
- While we could consider the missing probability as a constant over all covariates, we keep track of the minimum sampling probability $\zeta_{min}$ to provide more insight to practitioners for systematically dealing with missing covariates.

The technical challenge with regret analysis for the SA Lasso bandit comes from proving the lower- and upper-RE condition for $\widehat{\Gamma}_t$, shown in Appendix 5.8.3. The rest of the regret analysis follows the proof in Oh et al. [2021], but we keep track of the sampling probability $\zeta_{min}$. We include the proof of the regret analysis here for completeness.

#### 5.8.2.1 Proof of Theorem 5.5.2

*Proof.* For $T_0 = \frac{2\log(2d^2)}{C_0(s_0)^2}$ where $C_0(s_0) = \min\left(\frac{1}{2}, \frac{\lambda_{min}(\Sigma)}{216 s_0 \nu C_{\mathcal{X}} x_{max}^2} \frac{\zeta_{min}^2}{1+\zeta_{min}^2}\right)$, we first define the event $\mathcal{E}_t$ as follows

$$\mathcal{E}_t = \left\{ \|\Sigma_t - \widehat{\Gamma}_{miss,t}\|_\infty \leq \frac{\lambda_{min}(\Sigma)}{54 s_0 \nu C_{\mathcal{X}}} \right\}$$

In other words, $\mathcal{E}_t$ corresponds to the event where $\widehat{\Gamma}_{miss,t}$ is close enough to $\Sigma_t$. Then, the structure of the proof can be decomposed as follows

(a) $t \leq T_0$: The first part of the learning phase is an exploration phase based on $\hat{\beta}_t$. The bandit algorithm accumulates data to estimate $\widehat{\Gamma}_{miss,t}$ during this phase.

(b) $t > T_0$ and $\mathcal{E}_t^c$: $\widehat{\Gamma}_{miss,t}$ does not converge to $\Sigma_t$ and thus does not satisfy the lower- and upper-RE condition.

(c) $t > T_0$ and $\mathcal{E}_t$: SA Lasso bandit algorithm correctly estimates $\beta$ with missing covariates.

We use regret($t$) as the regret at time $t$, and $\mathcal{R}(t) = \mathbb{E}[\text{regret}(t)]$. Then, by Assumption 1 and 2, we can bound regret($t$) by

$$\text{regret}(t) \leq X_{t,a_t^*}\beta^* - X_{t,a_t}\beta^* \leq \|X_{t,a_t^*} - X_{t,a_t}\|_2 \|\beta^*\|_2 \leq 2x_{max}b$$

For ease of notation, we will set $x_{max} = 1$ and $b = 1$.

As noted above, we can divide regret($t$) into three parts

$$\begin{aligned}
\text{regret}(t) &= \text{regret}(t)\mathbb{1}(t \leq T_0) + \text{regret}(t)\mathbb{1}(t > T_0, \mathcal{E}_t) + \text{regret}(t)\mathbb{1}(t > T_0, \mathcal{E}^c) \\
&\leq 2\mathbb{1}(t \leq T_0) + \text{regret}(t)\mathbb{1}(t > T_0, \mathcal{E}_t) + 2x_{max}b\mathbb{1}(t > T_0, \mathcal{E}^c) \\
&= \underbrace{2\mathbb{1}(t \leq T_0)}_{(a)} + \underbrace{\text{regret}(t)\mathbb{1}\left((Z_{t,a_t} \oslash \hat{\boldsymbol{\zeta}}_t)\hat{\beta}_t \geq (Z_{t,a_t^*} \oslash \hat{\boldsymbol{\zeta}}_t)\hat{\beta}_t, t > T_0, \mathcal{E}_t\right)}_{(b)} + \underbrace{2\mathbb{1}(t > T_0, \mathcal{E}_t^c)}_{(c)}
\end{aligned}$$

We first relate part $(b)$ with the lasso convergence result for our case for $t > T_0$ and $\mathcal{E}_t$ detailed in Lemma 5.8.2. That is,

$$\text{regret}(t)\mathbb{1}\left((Z_{t,a_t} \oslash \hat{\boldsymbol{\zeta}}_t)\hat{\beta}_t \geq (Z_{t,a_t^*} \oslash \hat{\boldsymbol{\zeta}}_t)\hat{\beta}_t, t > T_0, \mathcal{E}_t\right) \leq \|\hat{\beta}_{t-1} - \beta^*\|_2 \leq d_t$$

where $d_t = \frac{c_0\sqrt{s_0}}{\alpha_1} \max\left\{\phi(\mathbb{Q}, \sigma_\varepsilon)\sqrt{\frac{\log d}{t}}, \lambda_t\right\}$.

For part (c), we have that by Lemma 5.8.3 the event $\mathcal{E}_t^c$ does not happen with probability at least $1 - c_0\exp(-c_1 \log d)$ for $t > T_0$.

Based on the regret at time $t$, we can calculated the expected regret for $t > T_0$ as

$$\begin{aligned}
\mathcal{R}(t) &\leq \mathbb{E}\left[\text{regret}(t)\mathbb{1}\left(2\|\widehat{\beta}_t - \beta^*\|_1 \geq \text{regret}(t), \mathcal{E}_t\right)\right] + 2x_{max}b\mathbb{P}(\mathcal{E}^c) \\
&= \mathbb{E}\left[\text{regret}(t)\mathbb{1}\left(2\|\widehat{\beta}_t - \beta^*\|_1 \geq \text{regret}(t), \text{regret}(t) \leq d_t, \mathcal{E}_t\right)\right] \\
&\quad + \mathbb{E}\left[\text{regret}(t)\mathbb{1}\left(2\|\widehat{\beta}_t - \beta^*\|_1 \geq \text{regret}(t), \text{regret}(t) > d_t, \mathcal{E}_t\right)\right] + 2x_{max}b\mathbb{P}(\mathcal{E}^c) \\
&\leq d_t + \mathbb{P}\left(2\|\widehat{\beta}_t - \beta^*\|_1 \geq d_t, \mathcal{E}_t\right) + 2\mathbb{P}(\mathcal{E}^c)
\end{aligned}$$

Then, with probability at least $1 - c_0\exp(-c_1\log d)$, we have that the regret is bounded by

$$\begin{aligned}
\text{regret}(t) &\leq T_0 + \sum_{t=T_0}^{T} d_t \\
&\leq \frac{2\log(2d^2)}{C_0(s_0)^2} + \frac{c_0\sqrt{s_0}}{\alpha_1}\phi(\mathbb{Q}, \sigma_\varepsilon)\sqrt{\log d}\sqrt{\log T}\sqrt{T} \\
&\asymp \mathcal{O}\left(\frac{1}{\zeta_{min}^2}\sqrt{s_0 T\log(dT)}\right)
\end{aligned}$$

□

**Lemma 5.8.1** (Kim and Paik [2019], Lemma 4.3)**.** *Suppose Assumption 1, 2, 3, and 4 hold, and* $d_t = \frac{c_0\sqrt{s_0}}{\alpha_1}\max\left\{\phi(\mathbb{Q}, \sigma_\varepsilon)\sqrt{\frac{\log d}{t}}, \lambda_{2t}\right\}$*. Then, for* $t \geq T_0$*,*

$$\mathbb{P}\left(\|\widehat{\beta}_t - \beta^*\|_2 \leq d_t \;\middle|\; \|\widehat{\beta}_{t-1} - \beta^*\|_2 \leq d_{t-1}, \ldots, \|\widehat{\beta}_{T_0} - \beta^*\|_2 \leq d_{T_0}\right) \leq c_1\exp(-c_2\log d)$$

*Thus, with probability at least* $1 - c_1\exp(-c_2\log d)$*,*

$$\|\hat{\beta}_t - \beta\|_2 \leq d_t \qquad \text{for every } t \geq T_0$$

*Proof.* The proof is same as Kim and Paik [2019] except we use Proposition 5.5.1 for the lasso convergence using the martingale difference sequence. □

**Lemma 5.8.2.** *With probability at least* $1 - c_1\exp(-c_2\log d)$*,*

$$regret(T, b) \leq \sum_{t=T_0}^{T} d_t = \frac{c_0\sqrt{s_0}}{\alpha_1}\phi(\mathbb{Q}, \sigma_\varepsilon)\sqrt{\log d}\sqrt{\log T}\sqrt{T}$$

*Proof.* The proof of the final lemma follows from Lemma 4.4 of Kim and Paik [2019] by combining the result of Proposition 5.5.1, and we include the proof here for completeness. Suppose $t \geq T_0$. Then, by Lemma 5.8.1, we have that $\|\hat{\beta}_{t-1} - \beta^*\|_2 \leq d_t$ with probability at least

$1 - c_1 \exp(-c_2 \log d)$. Also, by the definition of $a_t$, we have that

$$
\begin{aligned}
&(Z_{t,a_t} \oslash \hat{\boldsymbol{\zeta}}_t - Z_{t,a_t^*} \oslash \hat{\boldsymbol{\zeta}}_t)\hat{\beta}_{t-1} \geq 0 \\
&\Rightarrow \mathbb{E}[(Z_{t,a_t} \oslash \hat{\boldsymbol{\zeta}}_t - Z_{t,a_t^*} \oslash \hat{\boldsymbol{\zeta}}_t)\hat{\beta}_{t-1} | \mathcal{F}_{t-1}] = (X_{t,a_t} - X_{t,a_t^*})\hat{\beta}_{t-1} \geq 0
\end{aligned}
$$

as $X_t$'s and $U_t$'s are independent and $\mathbb{E}[U_{t,i}] = \hat{\boldsymbol{\zeta}}_t$ for all $i \in [K]$. Then,

$$
\begin{aligned}
regret(t,b) &\leq regret(t,b) + (X_{t,a_t} - X_{t,a_t^*})\hat{\beta}_{t-1} \\
&= (X_{t,a_t} - X_{t,a_t^*})(\hat{\beta}_{t-1} - \beta) \\
&\leq \|X_{t,a_t} - X_{t,a_t^*}\|_2 \|\hat{\beta}_{t-1} - \beta^*\|_2 \\
&\leq \|\hat{\beta}_{t-1} - \beta^*\|_2 \leq d_t
\end{aligned}
$$

Thus, we have the cumulative regret is at most $\sum_{t=1}^{T} d_t$.

$$
\begin{aligned}
\sum_{t=1}^{T} d_t &= \sum_{t=1}^{T} \frac{c_0\sqrt{s_0}}{\alpha_1} \max\left\{\phi(\mathbb{Q}, \sigma_\varepsilon)\sqrt{\frac{\log d}{t}}, \lambda_{2t}\right\} \\
&= \frac{c_0\sqrt{s_0}}{\alpha_1} \max\left\{\phi(\mathbb{Q}, \sigma_\varepsilon)\sqrt{\log d}, \lambda_{2t}\right\} \sum_{t=1}^{T} \sqrt{\frac{1}{t}} \\
&\leq \frac{c_0\sqrt{s_0}}{\alpha_1} \phi(\mathbb{Q}, \sigma_\varepsilon)\sqrt{\log d}\sqrt{\log T}\sqrt{T}
\end{aligned}
$$

□

### 5.8.3 Technical Proofs for SA Lasso bandit with missing covariates

Instead of dealing with the uneven sampling based on $\beta^*$ with pseudo-reward construction as in Kim and Paik [2019], Oh et al. [2021] address this problem by utilizing Assumption 5 on the symmetry of the covariate distribution. The core of the proof is adopted from Oh et al. [2021], but additional efforts are needed to prove similar results based on the non-positive semi-definite matrix $\widehat{\Gamma}_{miss,t}$. Since the important parts of proving the convergence of Lasso estimators include bounding $\|\widehat{\Gamma}_{miss,t} - \Sigma\|_\infty$, we address this problem first.

#### 5.8.3.1 RE condition of for SA Lasso bandit with missing covariates

In order to use the Bernstein inequality for adapted samples in Oh et al. [2021], we need to bound the infinity norm. Define the following

**Definition 5.8.1.** For all $i, j$ with $1 \leq i \leq j \leq d$, we define $\delta_t^{ij}(Z_t)$ to be a real-value function with random variable $Z_t = X_t \odot U_t \in \mathbb{R}^d$ as input:

$$\delta_t^{ij}(Z_t) = \begin{cases} \frac{1}{x_{max}^2} \frac{\zeta_i \zeta_j}{1+\zeta_i \zeta_j} \left( \frac{Z_t^{(i)} Z_t^{(j)}}{\hat{\zeta}_i \hat{\zeta}_j} - \mathbb{E}[X_t^{(i)} X_t^{(j)} | \mathcal{F}_{t-1}] \right) & i \neq j \\ \frac{1}{x_{max}^2} \frac{\zeta_i}{1+\zeta_i} \left( \frac{Z_t^{(i)} Z_t^{(i)}}{\hat{\zeta}_j} - \mathbb{E}[X_t^{(i)} X_t^{(i)} | \mathcal{F}_{t-1}] \right) & 1 \leq i \leq d \end{cases} \tag{5.15}$$

where $Z_t^{(i)}$ is the $i-th$ element of $Z_t$ and $U_{t,ij} \in \{0, 1\}$.

Note that this is a generalization of the deviation definition in Oh et al. [2021], where the definition is equivalent for the case of fully observed covariates with $\zeta_i = 1 \; \forall i \in [d]$ As we cannot directly use the convergence results from Loh and Wainwright [2012], the deviation proofs have to modify the proof structure of Oh et al. [2021] to incorporate the missing value correction of the proposed modified SA and DR Lasso bandits. Then, it follows that 1) $\mathbb{E}[\delta_t^{ij}(Z_t)|\mathcal{F}_{t-1}] = 0$ and 2) $\mathbb{E}[|\delta_t^{ij}|^m|\mathcal{F}_{t-1}] \leq 1$ for all $m \geq 2$. Therefore, similarly to the fully observed covariates in Oh et al. [2021], Lemma 5.8.10 can be applied for $\widehat{\Gamma}_t$.

**Lemma 5.8.3.** *For* $\tau \geq \frac{2\log(2d^2)}{C_0(s_0)^2}$, *where* $C_0(s_0) = \min\left(\frac{1}{2}, \frac{\lambda_{min}(\Sigma)}{216 s_0 \nu C_{\mathcal{X}} x_{max}^2} \frac{\zeta_{min}^2}{1+\zeta_{min}^2}\right)$, *we have that*

$$\mathbb{P}\left( \|\widehat{\Gamma}_{miss,t} - \Sigma_t\| \geq \frac{\lambda_{min}(\Sigma)}{54 s_0 \nu C_{\mathcal{X}}} \right) \leq \exp\left( -\frac{\tau C_0(s_0)^2}{2} \right)$$

*Proof.* $\delta_t^{ij}(Z_t)$ is defined for the diagonal entries and off-diagonal entries. We show the arguments for diagonal entries first. Note that we have

$$\frac{\zeta_{min}}{1+\zeta_{min}} \frac{\|\widehat{\Gamma}_{t,miss} - \Sigma_\tau\|_\infty}{x_{max}^2} = \max \frac{1}{\tau} \left| \sum_{t=1}^{\tau} \delta_t^{ii}(Z_t) \right|$$

Then, Lemma 5.8.10 gives us that

$$\mathbb{P}\left( \frac{\zeta_{min}}{1+\zeta_{min}} \frac{\|\widehat{\Gamma}_{miss,t} - \Sigma_\tau\|_\infty}{x_{max}^2} \geq w + \sqrt{2w} + \sqrt{\frac{8\log(2d)}{\tau}} + \frac{4\log(2d)}{\tau} \right) \leq \exp\left( -\frac{\tau w}{2} \right)$$

Then for $\tau \geq \frac{4\log(2d)}{C_0(s_0)^2}$, where $C_0(s_0) = \min\left(\frac{1}{2}, \frac{\lambda_{min}(\Sigma)}{216 s_0 \nu x_{max}^2} \frac{\zeta_{min}}{1+\zeta_{min}}\right)$,

$$\begin{aligned} w + \sqrt{2w} + \sqrt{\frac{8\log(2d)}{\tau}} + \frac{4\log(2d)}{\tau} &\leq 4C_0(s_0) \\ &\leq \frac{\lambda_{min}(\Sigma)}{54 s_0 \nu x_{max}^2} \end{aligned}$$

Thus, we have that

$$\begin{aligned} \mathbb{P}\left(\|\Sigma_\tau - \widehat{\Gamma}_{miss,t}\|_\infty \geq \frac{\lambda_{min}(\Sigma)}{54 s_0 \nu x_{min}^2}\right) &\leq \exp\left(-\frac{\tau w}{2}\right) \\ &= \exp\left(-\frac{\tau C_0(s_0)^2}{2}\right) \end{aligned}$$

The argument for off-diagonal entries are the same. □

**Lemma 5.8.4.** *For $\tau \geq \frac{2\log(2d^2)}{C_0(s_0)^2}$, where $C_0(s_0) = \min\left(\frac{1}{2}, \frac{\lambda_{min}(\Sigma)}{216 s_0 \nu C_{\mathcal{X}} x_{max}^2} \frac{\zeta_{min}^2}{1+\zeta_{min}^2}\right)$, there are universal positive constants $c_i$ such that $\widehat{\Gamma}_{miss,t}$ satisfies the lower- and upper-RE conditions with $\alpha_1 = \frac{\lambda_{min}(\Sigma_\tau)}{2}$, $\alpha_2 = \frac{3}{2}\lambda_{max}(\Sigma_\tau)$, and $\tau(t,d) = \frac{\lambda_{min}(\Sigma_t)}{2s_0}$.*

*Proof.* For $v \in \mathbb{K}(2s)$, we have that

$$\begin{aligned} |v^T(\widehat{\Gamma}_{miss,t} - \Sigma_\tau)v| &\leq \|\widehat{\Gamma}_{miss,t} - \Sigma_\tau\|_\infty \|v\|_0 \\ &\leq \frac{\lambda_{min}(\Sigma)}{54 s_0 \nu C_{\mathcal{X}}} s_0 \\ &\leq \frac{\lambda_{min}(\Sigma_\tau)}{54} \end{aligned}$$

Thus, it follows that

$$\mathbb{P}\left(|v^T(\widehat{\Gamma}_{miss,t} - \Sigma_\tau)v| \leq \frac{\lambda_{min}(\Sigma)}{54}\right) \leq 1 - \exp\left(-\frac{\tau C_0(s_0)^2}{2}\right)$$

Therefore, Lemma 5.8.11 gives us that $\widehat{\Gamma}_{miss,t}$ satisfies the RE-condition with $\alpha_1 = \frac{\lambda_{min}(\Sigma_\tau)}{2}$, $\alpha_2 = \frac{3}{2}\lambda_{max}(\Sigma_\tau)$, and $\tau(t,d) = \frac{\lambda_{min}(\Sigma_t)}{2s_0}$. □

#### 5.8.3.2 Technical Lemmas

We adopt the strategy in Bastani and Bayati [2020] and use the Bernstein concentration inequality (Lemma 5.8.5) for the martingale difference sequence adapted to a given filtration.

**Lemma 5.8.5.** *Let $\{D_k, \mathcal{F}_k\}_{k=1}^{\infty}$ be a martingale difference sequence, and suppose that $D_k$ is $\sigma$-subgaussian in an adapted sense, i.e., for all $\alpha \in \mathbb{R}$, $\mathbb{E}[e^{\alpha D_k}|\mathcal{F}_k] \leq e^{\alpha^2\sigma^2/2}$ almost surely. Then, for all $t \geq 0$, $\mathbb{P}[|\sum_{k=1}^n D_k| \geq t] \leq 2\exp[-t^2/(2n\sigma^2)]$.*

Lemma 5.8.5 follows from Theorem 2.19 of Wainwright [2019] when $\alpha_* = \alpha_k = 0$ and $\nu_k = \sigma$ for all $k$.

**Lemma 5.8.6.** *Under the assumption of Proposition 5.5.1,*

$$\mathbb{P}\left(\left\|\frac{\boldsymbol{\epsilon}_t^T \mathbf{Z}_t}{t}\right\|_\infty \geq c_0 \frac{\sigma_x \sigma_\epsilon}{\sqrt{k}\zeta_{min}}\sqrt{\frac{\log d}{t}}\right) \leq c_1 e^{-c_2 \log d/\zeta_{min}^2},$$

*where $\mathbf{Z}_t = \mathbf{X}_t \odot \mathbf{U}_t \in \mathbb{R}^{t\times d}$ is the observed matrix, $\boldsymbol{\epsilon}_t = [\varepsilon_1, \ldots, \varepsilon_t]^T \in \mathbb{R}^t$, and $\zeta_{min} = \min_{i\in[d]} \zeta_i$ is the minimum sampling probability of the covariates.*

*Proof.* The key component of the proof follows from the observation that $\mathbf{Z}_t$ is sub-Gaussian with the parameter at most $(\frac{1}{K}\Sigma_z, \frac{1}{K}\sigma_x^2)$. From our assumption, at time $\tau$, $X_\tau \in \mathbb{R}^{K\times d}$ is a sub-Gaussian matrix with parameter $\sigma_x^2$. Then, as noted in Loh and Wainwright [2012], $Z_\tau = X_\tau \odot U_\tau$ is a sub-Gaussian matrix with parameter at most $\sigma_x^2$.

Note that for any vector $v \in \mathbb{R}^d$ and any missing pattern of pattern of $X_{\tau,i}$, it follows that

$$\mathbb{E}[\exp(\alpha Z_{\tau,i} v)|\text{missing pattern}] = \mathbb{E}[\exp(\alpha X_{\tau,i} u)] \leq \exp\left(\frac{\sigma_x^2\alpha^2}{2}\right)$$

where the vector $u \in \mathbb{R}^p$ has entries $u_i = v_i$ if the $i$-th entry is observed and $u_i = 0$ if not observed. Thus, it follows that $Z_\tau = X_\tau \odot U_\tau$ is a sub-Gaussian matrix with parameter at most $\sigma_x^2$. Furthermore, since the $i$-th row of $\mathbf{Z}_\tau$ is a row-mean of $Z_\tau$, i.e. $\bar{Z}_\tau$, we have that $\mathbf{Z}_t$ is a sub-Gaussian matrix with parameter at most $\frac{1}{K}\sigma_x^2$.

Let $\mathcal{F}_t$ be the sigma algebra generated by random variables $\mathbf{Z}_t, \boldsymbol{r}_t$, and $U_t$. Also, for each $i \in [d]$, define $D_{\tau,i} = \varepsilon_{\tau,i}\mathbf{Z}_{t,\tau i}$. Then,

$$\mathbb{E}[\exp(\alpha D_{\tau,i})|\mathcal{F}_{t-1}] \leq \mathbb{E}_{\mathbf{Z}_t}[\exp(\alpha^2 \mathbf{Z}_{t,\tau i}\sigma_\varepsilon^2/2)|\mathcal{F}_{t-1}] \leq \exp(\frac{\alpha^2}{2}\cdot\frac{\sigma_x^2\sigma_\varepsilon^2}{k})$$

Then, it follows that $D_{1,i}, \ldots, D_{t,i}$ is a martingale difference sequence adapted to the filtration $\mathcal{F}_1 \subset \cdots \subset \mathcal{F}_t$ since $\mathbb{E}[\varepsilon_\tau \mathbf{Z}_{t,\tau i}|\mathcal{F}_t] = 0$ and $D_{\tau,i}$ is $\frac{(\sigma_x\sigma_\varepsilon)^2}{K}$-sub-Gaussian adapted to $\{\mathcal{F}_\tau\}_{\tau=1}^t$.

In order to prove the bound on $\|\cdot\|_\infty$, we rewrite the following probability using a union bound

$$\begin{aligned}\mathbb{P}\left(\left\|\frac{\boldsymbol{\epsilon}_t^T\mathbf{Z}_t}{t}\right\|_\infty \leq c_0\frac{\sigma_x\sigma_\varepsilon}{\sqrt{k}\zeta_{min}}\sqrt{\frac{\log d}{t}}\right) &\geq 1-\sum_{i=1}^{d}\mathbb{P}\left(|\boldsymbol{\epsilon}_t^T\mathbf{Z}_{t,\cdot i}| > \frac{tc_0\sigma_x\sigma_\varepsilon}{\sqrt{k}\zeta_{min}}\sqrt{\frac{\log d}{n}}\right)\\ &\geq 1-2d\exp\left(-c_0^2\frac{\log d}{\zeta_{min}^2}\right)\\ &= 1-2\exp\left(-c_1\frac{\log d}{\zeta_{min}^2}\right)\end{aligned}$$

for some constant $c_0 > 1$ and $c_1 > 0$. The second inequality follows from the Bernstein inequality stated in Lemma 5.8.5. □

**Lemma 5.8.7.** *(Lemma 3 of Loh and Wainwright [2012]) Under the condition of Proposition 5.5.1, there are universal positive constant $c_i$ such that $\hat{\Gamma}_{miss,t}$ satisfies lower- and upper- RE condition with $\alpha_1 = \lambda_{min}(\Sigma)/2$, $\alpha_2 = \frac{3}{2}\lambda_{max}(\Sigma_x)$, and*

$$\tau(t,d) = c_0\lambda(\Sigma_x)\max\left(\frac{\sigma_x^4}{\zeta_{min}^4k^2\lambda_{min}^2(\Sigma_x)},1\right)\cdot\frac{\log d}{t}$$

*with probability at least* $1-c_1\exp\left(-c_2t\cdot\min\left(\frac{\zeta_{min}^4k^2\lambda_{min}^2(\Sigma_x)}{\sigma_x^4},1\right)\right)$

*Proof.* The proof of this Lemma follows from Loh and Wainwright [2012] tracking all relative constants for the bandit problem. □

**Lemma 5.8.8.** *Under the conditions of Proposition 5.5.1, there are universal positive constants $c_i$ such that*

$$\|\hat{\gamma}_{miss,t} - \widehat{\Gamma}_{miss,t}\beta^*\|_\infty < \phi(\mathbb{Q},\sigma_\varepsilon)\sqrt{\frac{\log d}{t}}, \tag{5.16}$$

*holds, with parameter*

$$\phi(\mathbb{Q},\sigma_\varepsilon) = c_0\frac{\sigma_x}{\sqrt{k}\zeta_{min}}\left(\sigma_\varepsilon + \frac{\sigma_x}{\sqrt{k}\zeta_{min}}\right),$$

*with probability at least* $1-c_1\exp(-c_2\log d)$

*Proof.* The deviation condition (5.16) has the same proof structure as Loh and Wainwright [2012], but adopts Lemma 5.8.6 instead of Lemma 14 of Loh and Wainwright [2012]. The technical challenge from directly adopting Lemma 4 of Loh and Wainwright [2012] in our analysis comes from the assumption that $\varepsilon_\tau$ is conditionally sub-Gaussian adapted to $\mathcal{F}_{\tau-1}$ in bandit settings, but this can be easily addressed by using Lemma 5.8.6.

Note that

$$\begin{aligned}\|\hat{\gamma}_{miss,t} - \Sigma_x \beta^*\|_\infty &= \left\| \left( \frac{1}{t} \left( \mathbf{Z}_t^T \boldsymbol{r}_t - cov(\mathbf{Z}_{t,\cdot i}, \boldsymbol{r}_t) \right) \oslash \boldsymbol{\zeta} \right) \beta^* \right\|_\infty \\ &\leq \frac{1}{\zeta_{min}} \left\| \frac{1}{t} \left( \mathbf{Z}_t^T \boldsymbol{r}_t - cov(\mathbf{Z}_{t,\cdot i}, \boldsymbol{r}_t) \right) \beta^* \right\|_\infty \\ &\leq \frac{1}{\zeta_{min}} \left( \underbrace{\left\| \frac{1}{t} \left( \mathbf{Z}_t^T X_t - cov(\mathbf{Z}_{t,\cdot i}, X_{t,\cdot i}) \right) \beta^* \right\|_\infty}_{I} + \underbrace{\left\| \frac{\boldsymbol{\epsilon}_t^T \mathbf{Z}_t}{t} \right\|_\infty}_{II} \right)\end{aligned}$$

The component $I$ can be bounded by Lemma 14 of Loh and Wainwright [2012], and $II$ is bounded by Lemma 5.8.6. In other words,

$$\begin{aligned}\mathbb{P}\left( I \geq c_0 \frac{\sigma_x^2}{\zeta_{min}^2 \sqrt{k}} \sqrt{\frac{\log d}{t}} \right) &\leq c_1 \exp(-c_2 \log d) \\ \mathbb{P}\left( II \geq c_0 \frac{\sigma_x \sigma_\varepsilon}{\zeta_{min}^2 \sqrt{k}} \sqrt{\frac{\log d}{t}} \right) &\leq c_1 \exp(-c_2 \log d)\end{aligned} \tag{5.17}$$

Similarly, we have that

$$\begin{aligned}\|(\hat{\Gamma}_{miss,t} - \Sigma_x)\beta^*\|_\infty &= \left\| \left( \left( \frac{\mathbf{Z}_t^T \mathbf{Z}_t}{t} - \Sigma_z \right) \oslash M \right) \beta^* \right\|_\infty \\ &\leq \frac{1}{\zeta_{min}^2} \left\| \left( \frac{\mathbf{Z}_t^T \mathbf{Z}_t}{t} - \Sigma_z \right) \beta^* \right\|_\infty\end{aligned} \tag{5.18}$$

Then, using Lemma 3 of Loh and Wainwright [2012], we have that

$$\mathbb{P}\left( \|(\hat{\Gamma}_{miss,t} - \Sigma_x)\beta^*\|_\infty \leq c_0 \frac{\sigma_x^2}{k \zeta_{min}^2} \sqrt{\frac{\log d}{t}} \right) \leq 1 - c_1 \exp(-c_2 \log d) \tag{5.19}$$

Combining the inequalities (5.17) and (5.18) completes the proof. □

#### 5.8.3.3 Proof of proposition

*Proof.* With the previous results in hand, the proof of the proposition follows from Loh and Wainwright [2012], which requires 1) RE-condition (Lemma 5.8.7) and 2) deviation bounds (Lemma 5.8.8). □

### 5.8.4 Technical Lemmas

**Lemma 5.8.9** (Oh et al. [2021], Lemma 10)**.** *Suppose Assumption 6 holds. For a fixed vector $\beta \in \mathbb{R}^d$, we have*

$$\sum_{i=1}^{K} \mathbb{E}_{\mathcal{X}_t}\left[X_{t,i}^T X_{t,i} \mathbb{1}(X_{t,i} = \arg\max_{X \in \mathcal{X}_t} X\beta)\right] \succcurlyeq (2\nu C_{\mathcal{X}})^{-1}\Sigma$$

**Lemma 5.8.10** (Oh et al. [2021], Lemma 9)**.** *Suppose $\mathbb{E}[\delta_t^{ij}(Z_t)|\mathcal{F}_{t-1}] = 0$ and $\mathbb{E}[|\delta_t^{ij}(Z_t)|^m|\mathcal{F}_{t-1}] \leq m!$ for all integer $m \geq 2$, all $t \geq 1$, and all $1 \leq i \leq j \leq d$. Then, for all $w > 0$, we have*

$$\mathbb{P}\left(\max_{1\leq i\leq j\leq d}\left|\frac{1}{\tau}\sum_{t=1}^{\tau}\delta_t^{ij}(Z_t)\right| \geq w + \sqrt{2w} + \sqrt{\frac{4\log(2d^2)}{\tau} + \frac{2\log(2d^2)}{\tau}}\right) \leq \exp\left(-\frac{\tau w}{2}\right)$$

**Lemma 5.8.11** (Loh and Wainwright [2012], Lemma 13)**.** *Suppose $s \geq 1$ and $\widehat{\Gamma}_t$ is an estimator of $\Sigma_t$ satisfying the deviation condition*

$$|v^T(\widehat{\Gamma}_t - \Sigma_t)v| \leq \frac{\lambda_{min}(\Sigma_t)}{54} \qquad \forall v \in \mathbb{K}(2s)$$

*Then, we have the lower-RE condition*

$$v^T\widehat{\Gamma}_t v \geq \frac{\lambda_{min}(\Sigma_t)}{2}\|v\|_2^2 - \frac{\lambda_{min}(\Sigma_t)}{2s}\|v\|_1^2$$

*and the upper-RE condition*

$$v^T\widehat{\Gamma}_t v \leq \frac{3}{2}\lambda_{max}(\Sigma_t)\|v\|_2^2 + \frac{\lambda_{min}(\Sigma_t)}{2s}\|v\|_1^2$$

### 5.8.5 DR Lasso Bandit Modification

In this subsection, we introduce the modification on DR Lasso bandit proposed by Kim and Paik [2019]. This modification was implemented for comparison in the empirical studies (subsection **??**)

#### 5.8.5.1 Missing covariates DR Lasso bandit

For the DR Lasso bandit by Kim and Paik [2019], we observe the reward $\hat{r}_t$ after pulling arm $a_t$ based on (5.3) using the estimate $\hat{\beta}_{t-1}$ of $\beta$ given $\mathcal{F}_{t-1}$ and define the doubly-robust pseudo-reward

with covariate missingness as follows

$$\widetilde{r}_t = (\bar{Z}_t \oslash \hat{\boldsymbol{\zeta}})\hat{\beta}_{t-1} + \frac{1}{K}\frac{\hat{r}_t - (Z_{t,a_t} \oslash \hat{\boldsymbol{\zeta}})\hat{\beta}_{t-1}}{\pi_{t,a_t}} \tag{5.20}$$

where $\bar{Z}_t = \frac{1}{K}\sum_{i=1}^{K} Z_{t,i} \in \mathbb{R}^d$. The pseudo-reward (5.20) differs from the original DR Lasso bandit with full data in that $\bar{Z}_t \oslash \hat{\boldsymbol{\zeta}}$ replaces the fully observed mean context $\bar{X}_t = \frac{1}{K}\sum_{i=1}^{K} X_{t,i} \in \mathbb{R}^d$. As we are dealing missingness, we adjust the weights of the context vector in order to get the linear relationship $\mathbb{E}[\widetilde{r}_t|\mathcal{F}_{t-1}] = (\bar{Z}_t \oslash \zeta)^T\beta$.

Based on the pseudo-reward defined in (5.20), the DR Lasso bandit with missing context applies Lasso regression to the pair $(\bar{Z}_t \oslash \hat{\boldsymbol{\zeta}}, \widetilde{r}_t)$. For ease of notation, at round $t$, we define $\mathbf{Z}_t \in \mathbb{R}^{t\times d}$ and $\mathbf{r}_t = [\widetilde{r}_1, \ldots, \widetilde{r}_t]^T \in \mathbb{R}^t$, where $\mathbf{Z}_{t,\tau}$ is the $\tau$-th row vector defined by

$$\mathbf{Z}_{t,\tau} = \bar{Z}_\tau = \frac{1}{k}\sum_{k=1}^{K} Z_{\tau,k}.$$

This accomplishes context averaging for the case of covariate missingness, in analogy to the averaging of fully observed context introduced in the original DR Lasso bandit of Kim and Paik [2019]. By averaging the context matrix, Kim and Paik [2019] circumvent the violation of the i.i.d. condition and the uneven sampling of the contexts in the stochastic linear bandit setting. As a result, the extended oracle lasso convergence from Bastani and Bayati [2020] could be adopted for the regret analysis of DR Lasso bandit. Such averaging of $Z_{t,i}$ will have analogous benefits in our analysis of regret when there exists covariate missingness.

As noted in Oh et al. [2021], DR lasso bandit has to explore the sample space with an explicit exploration phase. Furthermore, the stringent assumption on the noise distribution in the original paper remains one of the main factors that make SA lasso bandit a better method to utilize. Furthermore, the calculation of the pseudo-reward has a high variance when the standard assumption on the noise variable is imposed.

**Algorithm 4:** DR Lasso bandit with missing covariates

**Input:** $\eta_1$, $\eta_2$, $T_0$, $R$
Initialize $\beta_0 = 0$, $\hat{\boldsymbol{\zeta}}_0 = 1$
**for** $t = 1, \ldots, T$ **do**
  Observe contexts $Z_t \sim \mathcal{P}_{K\times d}$
  and the missing pattern $U_t$
  Update $\hat{\boldsymbol{\zeta}}_t = \hat{\boldsymbol{\zeta}}_{t-1} + \frac{1}{k}\left(\sum_{i=1}^{k} U_{t,i} - \hat{\boldsymbol{\zeta}}_{t-1}\right)$
  **if** $t \leq T_0$ **then**
    Pull arm $a_t = i$ with probability $\frac{1}{K}$
    Update $\pi_{a_t} = \frac{1}{K}$
  **else**
    $\eta_{1t} = \eta_1\sqrt{\frac{\log(t\,\zeta_{min}^2)+\log d}{t\,\zeta_{min}^2}}$
    Sample $m_t \sim Ber\,(\eta_{1t})$
    **if** $m_t = 1$ **then**
      Pull arm $a_t = i$ with probability $\frac{1}{K}$
    **else**
      Pull arm $a_t = \arg\max_{i\in[K]}\{(Z_{t,i} \oslash \hat{\boldsymbol{\zeta}}_t)\hat{\beta}_{t-1}\}$
    **end**
    $\pi_{a_t} = \frac{\eta_{1t}}{K} + (1-\eta_{1t}) \cdot I(a_t = i)$
  **end**
  Observe $\hat{r}_{t,a_t}$ and calculate the pseudo-reward $\tilde{r}_t$ based on (5.20)
  Updated $\hat{\Gamma}_{miss,t}$ and $\hat{\gamma}_{miss,t}$ based on (5.6)
  $\eta_{2t} = \eta_2\sqrt{\frac{\log(t\,\zeta_{min}^2)+\log d}{t\,\zeta_{min}^2}}$
  Update $\hat{\beta}_t$ based on (5.5)
**end**

# CHAPTER 6

# Conclusion

Multi-modal and high-dimensional data from complex systems have called for scalable and interpretable models. In this thesis, we tackled several problems in various machine learning applications by imposing decomposable structures. Topic modeling and multi-spectral Brainbow images have a common generative model that is driven by latent topics or neuronal processes. Methods in these applications, including the proposed models in this thesis, look into the estimation of the decomposable means with multi-variate noises. In real-life applications, however, it is possible to exhibit dependencies among observations and variables. While the simultaneous estimation of dependencies in multiple modes of the matrix- and tensor-variate data has been studied, it is not straightforward how the existing methods can be combined to jointly estimate the mean and covariance structures.

On one hand, we have a decomposable mean structure that can be often motivated by the data generating process. Low-rank matrix decomposition methods have been well studied as it has desirable theoretical results, and an increasing body of machine learning researcher shows the benefit of using such models even for the tensor-variate data. However, there are no theoretical or empirical results to provide guidance for practitioners on how to decide on which Kronecker structures to use for estimating tensor- covariance models in a given application. An insight into this open question would allow us to have a better understanding of how to tackle tensor-variate problems. Furthermore, this would allow researchers to explicitly incorporate structured covariance models with decomposable means. In particular, combining the Kronecker covariance model and the low-rank mean structure could be powerful for modeling non-stationary spatio-temporal data as the Kronecker model effectively formulate spatio-temporal behavior by viewing the data as a tensor.

## 6.1 Future Works

### 6.1.1 Time Varying Topic Modeling with kernel estimator

The majority of approaches for estimating topic polytope focus on the mean structure of the word distributions in a given corpus. As an alternative, there has been a line of work that looks at the second- and/or third-order co-occurrence tensor to estimate the topic vertices [Anandkumar et al., 2012, Fu et al., 2018]. Extending these approaches, one could study the time-varying topic modeling with the correlation among words by looking at the kernelized co-occurrence matrix defined as follows

$$\hat{\Sigma}_n(t) = \frac{\sum_s \alpha_{st} \boldsymbol{w}_s \boldsymbol{w}_s^T}{\sum_s \alpha_{st}} \tag{6.1}$$

where $\boldsymbol{w}_s \in [0,1]^V$ is the word distribution for the document at time $s$ and $\alpha_{st} = \kappa\left(\frac{s-t}{h_n}\right)$ is a symmetric non-negative kernel function over time. Fu et al. [2018] connected this co-occurrence matrix with a geometric view similar to our method in Chapter 2. The consistency results can be derived based on the theoretical analyses in Zhou et al. [2010] with the estimation procedures from Anandkumar et al. [2012] and Fu et al. [2018]. Such a model would allow practitioners to explicitly build a continuous evolution of latent topics over time with theoretical guarantees.

### 6.1.2 Parallelizing Brainbow Tracing

The hidden Markov model formulation of the neuron tracing problem is a building block for tracing neurons in the whole brain. When the researchers in neuroimaging capture the Brainbow images, only a small portion of the brain is recorded at a time. Therefore, the tracing algorithm is only based on a small portion of the brain. Although we provide some guidance on the tracing problems in Chapter 3, there remain numerous computational challenges when it comes to stitching together tracing results from different parts of the brain. In theory, by calculating the global adjacency matrix with edges between adjacent states of the supervoxels, one can naively apply our method. However, the computational burden of calculating and storing the adjacency matrix for the entire brain is prohibitive. Instead, one could impose auto-regressive structures on the adjacency matrix and sequentially trace the entire brain.

### 6.1.3 Relaxing MCAR assumption in the bandit problem

In Chapter 5, we modeled the missing values in the contextual linear bandit problem as missing completely at random (MCAR) with the possibility of different covariate missing probabilities. While MCAR assumption is standard and leads to convenient theoretical analyses, MCAR can be

overly simplifying in certain applications. For example, in clinical trials, a patient is more likely to complete a health survey if he or she has a relevant medical condition. In such a setting, it would be more realistic to look into the underlying structure of the missing mechanism to improve the learner's performance.

# BIBLIOGRAPHY


Samaneh Abbasi-Sureshjani, Marta Favali, Giovanna Citti, Alessandro Sarti, and Bart M ter Haar Romeny. Curvature integration in a 5d kernel for extracting vessel connections in retinal images. *IEEE Transactions on Image Processing*, 27(2):606–621, 2017.

Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Online-to-confidence-set conversions and application to sparse stochastic bandits. In *Artificial Intelligence and Statistics*, pages 1–9, 2012.

Naoki Abe, Alan W Biermann, and Philip M Long. Reinforcement learning with immediate rewards and linear hypotheses. *Algorithmica*, 37(4):263–293, 2003.

Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics*, pages 2452–2482, 2012.

Anima Anandkumar, Dean P Foster, Daniel J Hsu, Sham M Kakade, and Yi-Kai Liu. A spectral algorithm for latent dirichlet allocation. *Advances in neural information processing systems*, 25, 2012.

Sanjeev Arora, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. In *International Conference on Machine Learning*, pages 280–288, 2013.

Thomas L Athey, Daniel J Tward, Ulrich Mueller, and Michael I Miller. Hidden markov modeling for maximum likelihood neuron reconstruction. *arXiv preprint arXiv:2106.02701*, 2021.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Zhong-Zhi Bai, Gene H Golub, and Michael K Ng. Hermitian and skew-hermitian splitting methods for non-hermitian positive definite linear systems. *SIAM Journal on Matrix Analysis and Applications*, 24(3):603–626, 2003.

Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine learning research*, 9(Mar):485–516, 2008.

Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020.

Erik Bekkers, Remco Duits, Tos Berendschot, and Bart ter Haar Romeny. A multi-orientation analysis approach to retinal vessel tracking. *Journal of Mathematical Imaging and Vision*, 49 (3):583–610, 2014.

Michele Benzi and Christine Klymko. Total communicability as a centrality measure. *Journal of Complex Networks*, 1(2):124–149, 2013.

Michele Benzi and Valeria Simoncini. Approximation of functions of large matrices with kronecker structure. *Numerische Mathematik*, 135(1):1–26, 2017.

Julian Besag. Efficiency of pseudolikelihood estimation for simple gaussian fields. *Biometrika*, pages 616–618, 1977.

José M Bioucas-Dias. A variable splitting augmented lagrangian approach to linear spectral unmixing. In *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, 2009. WHISPERS'09. First Workshop on*, pages 1–4. IEEE, 2009.

David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Patrick Breheny and Jian Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, 5(1): 232, 2011.

Dawen Cai, Kimberly B Cohen, Tuanlian Luo, Jeff W Lichtman, and Joshua R Sanes. Improved tools for the brainbow toolbox. *Nature methods*, 10(6):540, 2013.

Alexandra Carpentier and Rémi Munos. Bandit theory meets compressed sensing for high dimensional stochastic linear bandit. In *Artificial Intelligence and Statistics*, pages 190–198. PMLR, 2012.

Jiansheng Chen, Zhengqin Li, and Bo Huang. Linear spectral clustering superpixel. *IEEE Transactions on Image Processing*, 26(7):3317–3330, 2017.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

International Warfarin Pharmacogenetics Consortium. Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England Journal of Medicine*, 360(8):753–764, 2009.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Annual Conference on Learning Theory*, 2008.

A Philip Dawid. Some matrix-variate distribution theory: notational considerations and a bayesian application. *Biometrika*, 68(1):265–274, 1981.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

Remco Duits. *Perceptual organization in image analysis*. PhD thesis, PhD thesis, Technische Universiteit Eindhoven, 2005.

Remco Duits, Maurice Duits, Markus van Almsick, and Bart ter Haar Romeny. Invertible orientation scores as an application of generalized wavelet theory. *Pattern Recognition and Image Analysis*, 17(1):42–75, 2007a.

Remco Duits, Michael Felsberg, Gösta Granlund, and Bart ter Haar Romeny. Image analysis and reconstruction using a wavelet transform constructed from a reducible representation of the euclidean motion group. *International Journal of Computer Vision*, 72(1):79–102, 2007b.

Ernesto Estrada, Naomichi Hatano, and Michele Benzi. The physics of communicability in complex networks. *Physics reports*, 514(3):89–119, 2012.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Roger Fan, Byoungwook Jang, Yuekai Sun, and Shuheng Zhou. Precision matrix estimation with noisy and missing data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2810–2819. PMLR, 2019.

Marta Favali, Samaneh Abbasi-Sureshjani, Bart ter Haar Romeny, and Alessandro Sarti. Analysis of vessel connectivities in retinal images by cortically inspired spectral clustering. *Journal of Mathematical Imaging and Vision*, 56(1):158–172, 2016.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008a.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008b.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Applications of the lasso and grouped lasso to the estimation of sparse graphical models. Technical report, Technical report, Stanford University, 2010.

Drew Friedmann, Albert Pun, Eliza L Adams, Jan H Lui, Justus M Kebschull, Sophie M Grutzner, Caitlin Castagnola, Marc Tessier-Lavigne, and Liqun Luo. Mapping mesoscale axonal projections in the mouse brain using a 3d convolutional network. *Proceedings of the National Academy of Sciences*, 117(20):11068–11075, 2020.

Xiao Fu, Kejun Huang, Nicholas D Sidiropoulos, Qingjiang Shi, and Mingyi Hong. Anchor-free correlated topic modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

Davis Gilton and Rebecca Willett. Sparse linear contextual bandits via relevance vector machines. In *2017 International Conference on Sampling Theory and Applications (SampTA)*, pages 518–522. IEEE, 2017.

Gene Golub, Stephen Nash, and Charles Van Loan. A hessenberg-schur method for the problem ax+ xb= c. *IEEE Transactions on Automatic Control*, 24(6):909–913, 1979.

Nathan W Gouwens, Staci A Sorensen, Jim Berg, Changkyu Lee, Tim Jarsky, Jonathan Ting, Susan M Sunkin, David Feng, Costas A Anastassiou, Eliza Barkan, et al. Classification of electrophysiological and morphological neuron types in the mouse visual cortex. *Nature neuroscience*, page 1, 2019.

Lars Grasedyck. Existence and computation of low kronecker-rank approximations for large linear systems of tensor product structure. *Computing*, 72(3-4):247–265, 2004.

Kristjan Greenewald, Shuheng Zhou, and Alfred Hero III. Tensor graphical lasso (teralasso). *arXiv preprint arXiv:1705.03983*, 2017.

Kristjan Greenewald, Shuheng Zhou, and Alfred Hero III. Tensor graphical lasso (teralasso). *To appear in JRSS-B. arXiv preprint arXiv:1705.03983*, 2019.

Elizabeth P Hayden, Ryan E Wiegand, Eric T Meyer, Lance O Bauer, Sean J O'Connor, John I Nurnberger Jr, David B Chorlian, Bernice Porjesz, and Henri Begleiter. Patterns of regional brain activity in alcohol-dependent subjects. *Alcoholism: Clinical and Experimental Research*, 30(12):1986–1991, 2006.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Kejun Huang, Xiao Fu, and Nikolaos D Sidiropoulos. Anchor-free correlated topic modeling: Identifiability and algorithm. In *Advances in Neural Information Processing Systems*, pages 1786–1794, 2016.

David H Hubel and Torsten N Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3):574–591, 1959.

Byoungwook Jang and Alfred Hero. Minimum volume topic modeling. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3013–3021. PMLR, 2019.

Michiel HJ Janssen, Augustus JEM Janssen, Erik J Bekkers, J Oliván Bescós, and Remco Duits. Design and processing of invertible orientation scores of 3d images. *Journal of Mathematical Imaging and Vision*, 60(9):1427–1458, 2018.

Alfredo Kalaitzis, John Lafferty, Neil D Lawrence, and Shuheng Zhou. The bigraphical lasso. In *International Conference on Machine Learning*, pages 1229–1237, 2013.

Kshitij Khare, Sang-Yun Oh, and Bala Rajaratnam. A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(4):803–825, 2015.

Gi-Soo Kim and Myunghee Cho Paik. Doubly-robust lasso bandit. In *NeurIPS 2019 : Thirty-third Conference on Neural Information Processing Systems*, pages 5877–5887, 2019.

Gennady G Knyazev. Motivation, emotion, and their inhibitory control mirrored in brain oscillations. *Neuroscience & Biobehavioral Reviews*, 31(3):377–395, 2007.

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

Daniel Kressner and Christine Tobler. Krylov subspace methods for linear systems with tensor product structure. *SIAM journal on matrix analysis and applications*, 31(4):1688–1714, 2010.

John D Lafferty and David M Blei. Correlated topic models. In *Advances in neural information processing systems*, pages 147–154, 2006.

Aaron Q Li, Amr Ahmed, Sujith Ravi, and Alexander J Smola. Reducing the sampling complexity of topic models. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 891–900. ACM, 2014.

Rongjian Li, Tao Zeng, Hanchuan Peng, and Shuiwang Ji. Deep learning segmentation of optical microscopy images improves 3-d neuron reconstruction. *IEEE transactions on medical imaging*, 36(7):1533–1541, 2017.

Tongliang Liu, Mingming Gong, and Dacheng Tao. Large-cone nonnegative matrix factorization. *IEEE transactions on neural networks and learning systems*, 28(9):2129–2142, 2017.

Jean Livet, Tamily A Weissman, Hyuno Kang, Ryan W Draft, Ju Lu, Robyn A Bennis, Joshua R Sanes, and Jeff W Lichtman. Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature*, 450(7166):56, 2007.

Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664, 2012.

Po-Ling Loh and Martin J Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *The Journal of Machine Learning Research*, 16(1): 559–616, 2015a.

Po-Ling Loh and Martin J Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16:559–616, 2015b.

Po-Ling Loh and Martin J Wainwright. Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics*, 45(6):2455–2482, 2017.

Gabriel L Lozano, Juan I Bravo, Manuel F Garavito Diago, Hyun Bong Park, Amanda Hurley, S Brook Peterson, Eric V Stabb, Jason M Crawford, Nichole A Broderick, and Jo Handelsman. Introducing thor, a model microbiome for genetic dissection of community behavior. *MBio*, 10 (2):e02846–18, 2019.

Xiang Lyu, Will Wei Sun, Zhaoran Wang, Han Liu, Jian Yang, and Guang Cheng. Tensor graphical model: Non-convex optimization and statistical inference. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.

Nicolai Meinshausen and Peter Buhlmann. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, pages 1436–1462, 2006.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462, 2006.

José MP Nascimento and José M Bioucas-Dias. Hyperspectral unmixing based on mixtures of dirichlet components. *IEEE Transactions on Geoscience and Remote Sensing*, 50(3):863–878, 2012.

José MP Nascimento and José MB Dias. Vertex component analysis: A fast algorithm to unmix hyperspectral data. *IEEE transactions on Geoscience and Remote Sensing*, 43(4):898–910, 2005.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, Bin Yu, et al. A unified framework for high-dimensional analysis of $m$-estimators with decomposable regularizers. *Statistical science*, 27(4):538–557, 2012.

XuanLong Nguyen. Posterior contraction of the population polytope in finite admixture models. *Bernoulli*, 21(1):618–646, 2015.

Min-hwan Oh, Garud Iyengar, and Assaf Zeevi. Sparsity-agnostic lasso bandit. In *International Conference on Machine Learning*, pages 8271–8280. PMLR, 2021.

Seyoung Park, Kerby Shedden, and Shuheng Zhou. Non-separable covariance models for spatio-temporal data, with applications to neural encoding analysis. *arXiv preprint arXiv:1705.05265*, 2017.

Hanchuan Peng, Zongcai Ruan, Deniz Atasoy, and Scott Sternson. Automatic reconstruction of 3d neuron structures using a graph-augmented deformable model. *Bioinformatics*, 26(12):i38–i46, 2010.

Jie Peng, Pei Wang, Nengfeng Zhou, and Ji Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009.

Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

Xinghao Qiao, Shaojun Guo, and Gareth M James. Functional graphical models. *Journal of the American Statistical Association*, 114(525):211–222, 2019.

Miroslav Radojević and Erik Meijering. Automated neuron tracing using probability hypothesis density filtering. *Bioinformatics*, 33(7):1073–1080, 2017.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.

Douglas H Roossien, Benjamin V Sadis, Yan Yan, John M Webb, Lia Y Min, Aslan S Dizaji, Luke J Bogart, Cristina Mazuski, Robert S Huth, Johanna S Stecher, Sriakhila Akula, Fred Shen, Ye Li, Tingxin Xiao, Madeleine Vandenbrink, Jeff W Lichtman, Takao K Hensch, Erik D Herzog, and Dawen Cai. Multispectral tracing in densely labeled mouse brain with ntracer. *Bioinformatics*, 35(18):3544–3546, September 2019. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btz084.

Adam J. Rothman, Peter J. Bickel, Elizaveta Levina, and Ji Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

Mark Rudelson and Shuheng Zhou. Reconstruction from anisotropic random measurements. In *Conference on Learning Theory*, pages 10–1. JMLR Workshop and Conference Proceedings, 2012.

Mark Rudelson and Shuheng Zhou. Errors-in-variables models with dependent measurements. *Electronic Journal of Statistics*, 11(1):1699–1797, 2017.

Mark Rudelson, Shuheng Zhou, et al. Errors-in-variables models with dependent measurements. *Electronic Journal of Statistics*, 11(1):1699–1797, 2017.

Fred Y. Shen, Margaret M. Harrington, Logan A. Walker, Hon Pong Jimmy Cheng, Edward S. Boyden, and Dawen Cai. Light microscopy based approach for mapping connectivity with molecular specificity. *bioRxiv*, 2020. doi: 10.1101/2020.02.24.963538. URL https://www.biorxiv.org/content/early/2020/02/25/2020.02.24.963538.

Nicolas Städler and Peter Bühlmann. Missing values: sparse inverse covariance estimation and an extension to sparse regression. *Statistics and Computing*, 22(1):219–235, 2012.

Jian Tang, Zhaoshi Meng, Xuanlong Nguyen, Qiaozhu Mei, and Ming Zhang. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *International Conference on Machine Learning*, pages 190–198, 2014.

Yee W Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Sharing clusters among related groups: Hierarchical dirichlet processes. In *Advances in neural information processing systems*, pages 1385–1392, 2005.

Ambuj Tewari and Susan A Murphy. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, pages 495–517. Springer, 2017.

Theodoros Tsiligkaridis, Alfred O Hero III, and Shuheng Zhou. On convergence of kronecker graphical lasso algorithms. *IEEE transactions on signal processing*, 61(7):1743–1755, 2013.

Engin Turetken, Fethallah Benmansour, Bjoern Andres, Hanspeter Pfister, and Pascal Fua. Reconstructing loopy curvilinear structures using integer programming. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1822–1829, 2013.

Sara A Van De Geer, Peter Bühlmann, et al. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Xue Wang, Mingcheng Wei, and Tao Yao. Minimax concave penalized multi-armed bandit model with high-dimensional covariates. In *International Conference on Machine Learning*, pages 5200–5208, 2018.

Yu Wang and Alfred Hero. Sg-palm: a fast physically interpretable tensor graphical model. In *International Conference on Machine Learning*, pages 10783–10793. PMLR, 2021.

Yu Wang, Arunachalam Narayanaswamy, Chia-Ling Tsai, and Badrinath Roysam. A broadly applicable 3-d neuron tracing method based on open-curve snake. *Neuroinformatics*, 9(2):193–217, 2011.

Yu Wang, Byoungwook Jang, and Alfred Hero. The sylvester graphical lasso (syglasso). In *International Conference on Artificial Intelligence and Statistics*, pages 1943–1953. PMLR, 2020.

Yun Wang, Peng Xie, Hui Gong, Zhi Zhou, Xiuli Kuang, Yimin Wang, An-an Li, Yaoyao Li, Lijuan Liu, Matthew B Veldman, et al. Complete single neuron reconstruction reveals morphological diversity in molecularly defined claustral and cortical neuron types. *BioRxiv*, page 675280, 2019.

Ami Wiesel and Alfred O Hero. Decomposable principal component analysis. *IEEE Transactions on Signal Processing*, 57(11):4369–4377, 2009.

Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273. ACM, 2003.

Jinhui Yuan, Fei Gao, Qirong Ho, Wei Dai, Jinliang Wei, Xun Zheng, Eric Po Xing, Tie-Yan Liu, and Wei-Ying Ma. Lightlda: Big topic models on modest computer clusters. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1351–1361. International World Wide Web Conferences Steering Committee, 2015.

Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

Mikhail Yurochkin and XuanLong Nguyen. Geometric dirichlet means algorithm for topic inference. In *Advances in Neural Information Processing Systems*, pages 2505–2513, 2016.

Mikhail Yurochkin, Aritra Guha, and XuanLong Nguyen. Conic scan-and-cover algorithms for nonparametric topic modeling. In *Advances in Neural Information Processing Systems*, pages 3881–3890, 2017.

Cun-Hui Zhang and Tong Zhang. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, pages 576–593, 2012.

Cun-Hui Zhang et al. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.

Jiong Zhang, Behdad Dashtbozorg, Erik Bekkers, Josien PW Pluim, Remco Duits, and Bart M ter Haar Romeny. Robust retinal vessel segmentation via locally adaptive derivative frames in orientation scores. *IEEE transactions on medical imaging*, 35(12):2631–2644, 2016.

Xiao Lei Zhang, Henri Begleiter, Bernice Porjesz, Wenyu Wang, and Ann Litke. Event related potentials during object recognition tasks. *Brain Research Bulletin*, 38(6):531–538, 1995.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563, 2006.

Shuheng Zhou. Gemini: Graph estimation with matrix variate normal instances. *The Annals of Statistics*, 42(2):532–562, 2014.

Shuheng Zhou, John Lafferty, and Larry Wasserman. Time varying undirected graphs. *Machine Learning*, 80(2-3):295–319, 2010.

Shuheng Zhou, Philipp Rütimann, Min Xu, and Peter Bühlmann. High-dimensional covariance estimation based on gaussian graphical models. *Journal of Machine Learning Research*, 12 (Oct):2975–3026, 2011.

Zhi Zhou, Hsien-Chi Kuo, Hanchuan Peng, and Fuhui Long. Deepneuron: an open deep learning toolbox for neuron tracing. *Brain informatics*, 5(2):1–9, 2018.

Hongxiao Zhu, Nate Strawn, and David B Dunson. Bayesian graphical models for multivariate functional data. *The Journal of Machine Learning Research*, 17(1):7157–7183, 2016.