Interpretable and Scalable Graphical Models for Complex Spatio-temporal Processes

by

Yu Wang

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:

Assistant Professor Yang Chen, Co-Chair
Professor Alfred O. Hero III, Co-Chair
Assistant Professor Walter Dempsey
Dr. Earl Lawrence, Los Alamos National Laboratory

Yu Wang
wayneyw@umich.edu
ORCID iD: 0000-0002-6287-4710

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my Ph.D. advisors Dr. Alfred Hero and Dr. Yang Chen. This dissertation would not have been possible without their continuous guidance, support, and encouragement. Al is an exemplary scholar, who is always hardworking and dedicated to research. I am constantly amazed by his sharpness on research and his deep insights about many different topics, ranging from physics, applied math, to statistics, and computer science. I am deeply indebted to him for devoting so much time and energy to mentoring me and guiding me through the transition from a student to a researcher. His passion for research have profoundly influenced me from both professional and personal perspectives. I am also fortunate enough to work with Yang, an inspirational advisor, who is incredibly generous with her ideas and time; and a caring mentor, who constantly offers her support and empathy. Our meetings and discussions have been a great source of inspiration and greatly shaped my way of approaching research problems.

I am also very grateful to have Dr. Walter Dempsey and Dr. Earl Lawrence serving on my doctoral dissertation committee and providing me with invaluable comments. I was fortunate to collaborate with Walter on topic models and important problems in the public health domain. Walter's enthusiasm for research has left a lot of positive impacts on me. I first met Earl during his visit to the department as a distinguished alumni speaker. Later, I had the opportunity to work with him at LANL on distributed dimensionality reduction and applications on space weather. His constant support and humor make all the research meetings there and my overall

experience at LANL enjoyable.

My thanks also go to staff members at the University of Michigan, Department of Statistics, who have helped me over the past few years. In particular, I want to thank Judy, Jean, Bebe, Virggie, Andrea, Gina, and many others, who always patiently helped with my questions and warmly welcomed me into the office with big smiles.

Additionally, I would like to express my gratitude and appreciation to Dr. Jim Zidek and Dr. Nhu Le at the University of British Columbia in Canada. My research career started with Jim and Nhu, who are both brilliant researchers and caring mentors. Jim has always been a role model to me, and without him, I would not have gone this far in this journey.

To all my friends that I made and all the people that I met throughout the Ph.D. studies: This journey would not have been so rewarding without you! Shout out to everyone in our research labs, especially Byoung and Zeyu from the Hero Group that I was fortunate to collaborate with; Leo who organized those fun board games; and Chengcheng, Cheng, Yangyi, and Ziping in my cohort. It was a great fun to spend time with you, and I have learned a lot from our interactions. Last but not least, I would like to thank my parents, Xiaoqing Tan the duck, Kitty the cat, and Larry the chinchilla for their unwavering support and unconditioned love.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

This thesis focuses on data that has complex spatio-temporal structure and on probabilistic graphical models that learn the structure in an interpretable and scalable manner. We target two research areas of interest: Gaussian graphical models for tensor-variate data and summarization of complex time-varying texts using topic models. This work advances the state-of-the-art in several directions. First, it introduces a new class of tensor-variate Gaussian graphical models via the Sylvester tensor equation. Second, it develops an optimization technique based on a fast-converging proximal alternating linearized minimization method, which scales tensor-variate Gaussian graphical model estimations to modern big-data settings. Third, it connects Kronecker-structured (inverse) covariance models with spatio-temporal partial differential equations (PDEs) and introduces a new framework for ensemble Kalman filtering that is capable of tracking chaotic physical systems. Fourth, it proposes a modular and interpretable framework for unsupervised and weakly-supervised probabilistic topic modeling of time-varying data that combines generative statistical models with computational geometric methods. Throughout, practical applications of the methodology are considered using real datasets. This includes brain-connectivity analysis using EEG data, space weather forecasting using solar imaging data, longitudinal analysis of public opinions using Twitter data, and mining of mental health related issues using TalkLife data. We show in each case that the graphical modeling framework introduced here leads to improved interpretability, accuracy, and scalability.

CHAPTER I

Introduction

Complex, structured data is ubiquitous in both industrial and academic settings and has elicited a commensurate interest in utilizing such information to assist in inference and decision making. Often, there exists simpler and interpretable underlying structure that can be exploited to make inference and summarization procedures more tractable. For large datasets, in particular, it is imperative to consider the data in the context of its structure to develop parsimonious models that represent the intrinsic form of the data well and provide computationally efficient, theoretically grounded inference procedures. On one hand, searching for such structures can help to summarize the data in a more interpretable manner and find relevant attributes of the data of interest that might otherwise go undetected. On the other hand, for some datasets the structure is explicit, and thus requires careful consideration when reasoning about modeling decisions.

Moreover, despite the fact that machine learning models have recently demonstrated great success in learning the above-mentioned complex structures that enable them to make predictions about unobserved data, the ability to interpret what a model has learned is yet to be determined and has been receiving an increasing amount of attention (Rudin et al., 2022; Murdoch et al., 2019; Du et al., 2019; Doshi-Velez and Kim, 2017; Rudin, 2019; Papernot and McDaniel, 2018). In particular, Murdoch et al.

(2019) recently introduced a unified PDR (predictive, descriptive, relevant) framework for discussing interpretations of machine learning and statistical models in general, and categorized existing techniques into model-based and post-hoc categories, with subgroups including sparsity, modularity, and simulatability. In this dissertation, the focus is on data that has temporal or spatio-temporal structure and on problems that benefit from the application of spatio-temporal based inference algorithms. In both cases, we target the overarching desiderata described in the PDR interpretability framework and introduce statistical methods that improve the overall (predictive and descriptive) accuracy and relevancy through both model-based and post-hoc approaches. Specifically, we attempt to advance two research areas. First, Gaussian graphical models for tensor-valued data is studied, and we develop a sparse multiway representation of constituent spatial and temporal processes, which enables a decomposable (i.e., spatial and temporal) and scalable framework for analyzing tensor data, especially that generated from complex dynamical systems. Second, a framework for topic modeling of time-varying texts is developed. The framework breaks previously (computationally and statistically) intractable approaches into tractable modules and utilizes computational geometric methods for extracting various (stable) forms of information from the fitted model. Overall, we improve interpretability, scalability, and accuracy throughout the full life cycle (see Figure I.1) of a data science problem with complex structure. Below, these two research areas are briefly introduced that form the backbone of this thesis.

1.1 Gaussian Graphical Models for Tensor-valued Data

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 sample size is much

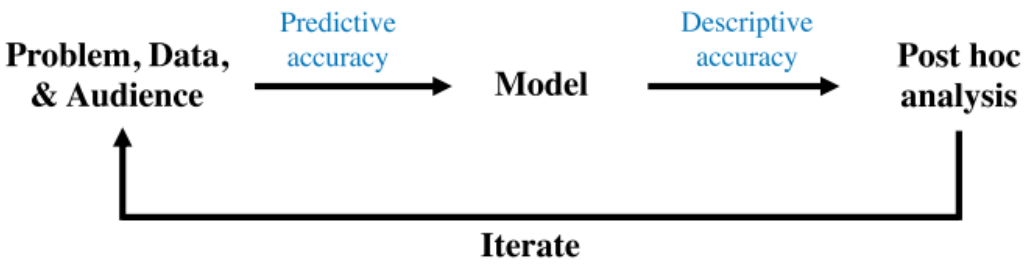


Figure I.1: An overview given by Murdoch et al. (2019) that introduces different stages (black text) in a data–science life cycle where interpretability is important.

less than the dimension of each sample ($N \ll d$). The success of the ℓ_1-penalized method for estimating conditional dependencies was demonstrated in Meinshausen and Bühlmann (2006) and Friedman et al. (2008) for the multivariate Gaussian setting. Contributing to this success is the underlying graphical structure (see Figure I.2) that facilitates simple interpretation and ties the statistical model to the mathematical field of graph theory (Lauritzen, 1996).

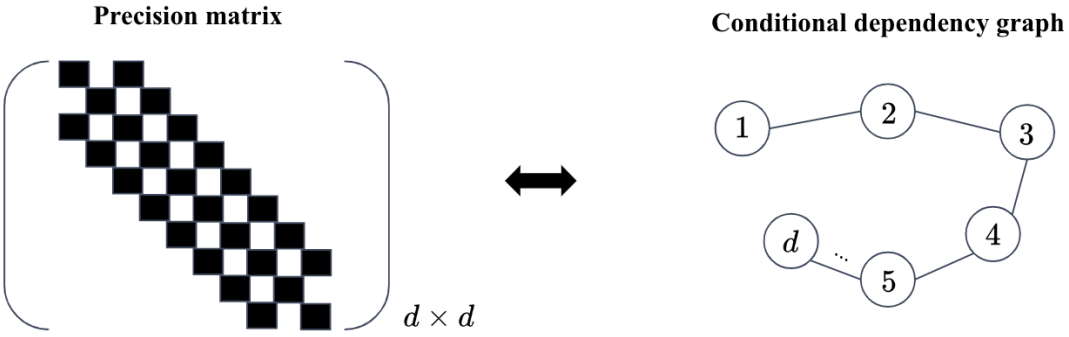


Figure I.2: For multivariate Gaussian variables, the conditional dependence structure encoded in the precision matrix (left) can be represented by a simple chain graph (right).

This success has naturally led researchers to generalize these methods to multiway tensor-valued data. Such generalizations are of benefit for many applications, including for example, 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 $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1 \times \cdots \times d_k}$ and analyzing the overall precision matrix $\boldsymbol{\Omega} \in \mathbb{R}^{d \times d}$, where $d = \prod_{k=1}^{K} d_k$. Even for a two-dimensional tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1 \times d_2}$, the computation complexity and sample complexity is high since the number of parameters in the precision matrix grows quadratically as d^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) decomposition of $\boldsymbol{\Omega}$, where each decomposed factor has an underlying graphical representation like Figure I.2 that can be modeled, estimated, and interpreted separately. The earliest and most popular form of sparse structured precision matrix estimation represents $\boldsymbol{\Omega}$ as the KP of smaller precision matrices, which corresponds to a separable structure across different modes of a data tensor (see Figure I.3). Tsiligkaridis et al. (2013) and Zhou (2014) proposed to model the precision matrix as a sparse KP of the precision 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. Lyu et al. (2019) extended the KP structured model to tensor-valued data, and provided new theoretical insights into such models. An alternative, called the Bigraphical Lasso, was proposed by Kalaitzis et al. (2013) to model conditional dependency structures of precision matrices by using a KS representation

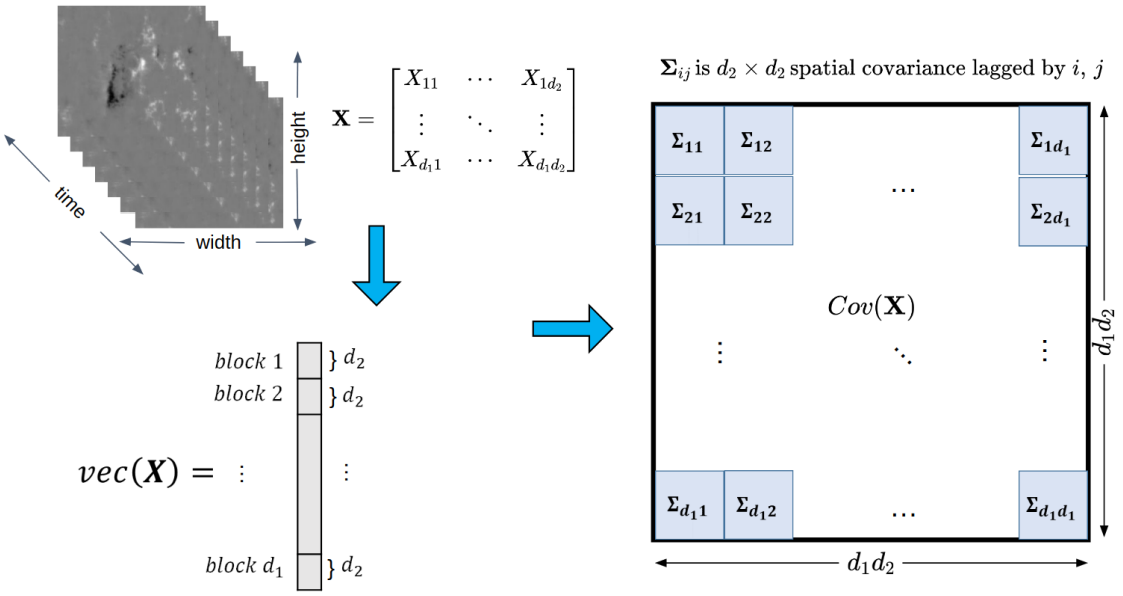


Figure I.3: Multiway data results in a patterned covariance. This structure can be exploited by assuming similar patterns within each block, such as those assumed in the Kronecker product models.

$\boldsymbol{\Omega} = \boldsymbol{\Psi}_1 \oplus \boldsymbol{\Psi}_2 = (\boldsymbol{\Psi}_1 \otimes \mathbf{I}) + (\mathbf{I} \otimes \boldsymbol{\Psi}_2)$. On the other hand, Rudelson and Zhou (2017) and Park et al. (2017) studied the KS structure on the covariance matrix $\boldsymbol{\Sigma} = \mathbf{A} \oplus \mathbf{B}$ which corresponds to errors-in-variables models. More recently, Greenewald et al. (2019) proposed a model that generalized the KS structure to model tensor-valued data, called the TeraLasso. As shown in their paper, compared to the KP structure, KS structure on the precision matrix leads to a different type of separability on the covariance matrix that provides a more parsimonious representation.

Despite being modeling choices, the KP and KS structures admit their own pros and cons. The KP model admits a simple stochastic representation, which defines a generating process for the underlying data. Unlike the KP model, the KS model does not lead to a natural generative interpretation. From another perspective, Kronecker structures can be characterized by the product graphs of the individual components. In particular, Kalaitzis et al. (2013) first motivated the KS structure on the precision matrix by relating Kronecker sum of matrices to the associated Cartesian product

graph. Thus, the overall structure of $\mathbf{\Omega}$ naturally leads to a parsimonious model that brings the individual components together. The KP, however, corresponds to the direct tensor product of the individual graphs and leads to a denser dependency structure in the precision matrix Greenewald et al. (2019). Chapter II proposes a new Kronecker-structured graphical model that admits a natural stochastic representation for precision matrices associated with tensor data. The resulted Gaussian graphical model strikes a balance between the KP- and KS- structured models. The new model poses additional challenge in computation, Chapter III proposes an estimation algorithm that utilizes state-of-the-art optimization technique and scales the method to modern big data applications. Chapter IV studies the connection between multiway Gaussian graphical models and second-order representation of spatio-temporal partial differential equations (PDE) and introduces an Kalman filtering framework for model-based physics-informed data assimilation.

1.2 Dynamic Topic Models

Probabilistic topic model is a suite of algorithms that aim to automatically discover and annotate large collections of documents that contain useful information with thematic labels. Topic modeling algorithms are statistical methods that analyze the words of the original texts to discover the themes that run through them, how those themes are connected to each other, and how they change over time. One such model that has been very successful is the Latent Dirichlet Allocation (LDA) model (Blei et al., 2003), which infers the topics (i.e., thematic information) in a corpus by assuming an underlying generative process whereby the documents are created, so that one may infer, or reverse engineer, it. The LDA model posits that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over all the words. The complete probabilistic structure can be represented by a simple graphical model shown in Figure I.4.

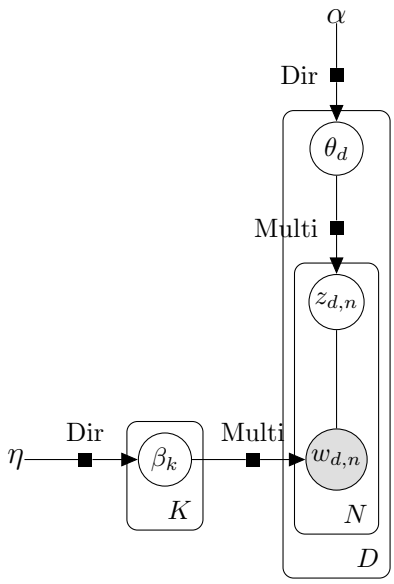


Figure I.4: Graphical representation of the standard Latent Dirichlet Allocation (LDA) model. Here nodes are random variables; edges indicate dependence through probability distributions (e.g., Dirichlet or multinomial). Shaded nodes are observed; unshaded nodes are latent. Plates indicate replicated variables.

Numerous extensions of the original LDA model have been proposed to handle more complex data that exhibits serial dependencies. In particular, Blei and Lafferty (2006) proposed a Dynamic Topic Model (DTM) that models time-varying corpus (e.g., archive of articles published on the Science journal from 1990 to 2020), and the alignment among topics across time steps is captured by a Kalman filter procedure with a Markov assumption where the state (of topics) at time $t + 1$ is independent of all other history given the state at time t. Wang and McCallum (2006) deals with similar data and introduced a non-Markov continuous-time model called the Topics-over-Time (TOT), which captures changes in the occurrence (and co-occurrence conditioned on time) of the topics themselves, not changes in the word distribution of each topic. Wang et al. (2008) further improved the DTM with a continuous time variant called cDTM that uses Brownian motion to model the latent topics in a sequential collection of documents, where a topic is a pattern of word use that is expected to evolve over the course of the collection.

All the methods mentioned above try to build certain dynamical or flexible struc-

tures explicitly into the probabilistic model. Besides the fact that these methods generally rely on complex stochastic process specifications to model temporal or other dependency structures, they all suffer from the following: 1. natural interpretation comes at the cost of correct model assumption: DTM and its variants achieve interpretability under the assumption of model being correct, which is restrictive as complicated real world applications tend not to follow the models perfectly and any abrupt change in the data makes modeling results hard to interpret; 2. computational instability: as many of these methods rely on either expensive MCMC sampling schemes or variational approximations as inference algorithms, they face the issue of getting trapped into local minimum/maximum of their objectives, which makes the results hard to interpret with confidence; 3. there is inherent inflexibility to diverse dynamical structures, as most methods are developed for handling specific temporal dynamics and are not able to capture all types (e.g., abnormality, clustering, etc) of variations jointly. In Chapter V, a scalable and interpretable framework is developed that attempts to overcome those issues in traditional dynamic topic models. Additionally, the proposed temporal topic modeling approach is extended to incorporate side information via weak supervision.

1.3 Outline and Contributions

This section lists the chapters and corresponding contributions in this dissertation. Each chapter aims to be a self contained exposition on a specific topic; as a result, some introductory material for particular chapters are similar in scope.

Chapter II describes a structured Gaussian graphical model for tensor-valued data. Here, we consider the underlying generating process of the data to be governed by a Sylvester equation. We show that this leads to a Kronecker sum structural assumption on the square root factor of the precision matrix of the data. The resulted modeling approach is able to simultaneously improve robustness, richness, and interpretability

of existing Kronecker-structured models. This chapter is based on Wang et al. (2020c) and was published in the *Proceedings of the* 23^{rd} *International Conference on Artificial Intelligence and Statistics*.

Chapter III tackles a challenging optimization problem posed by the Sylvester graphical model. An algorithm based on the proximal alternating linearized minimization is proposed to estimate generating factors of the model. State-of-the-art convergence rate is achieved and a comprehensive convergence analysis is done via recent development of optimization theory. Practically, we apply the new procedure to astrophysics-related application in solar flare prediction, where we model the solar magnetogram and atmosphere as Guassian Markov Random Field (GMRF) induced by a Sylvester-structured precision matrix. The utility of the estimated precision matrix is demonstrated via a linear prediction of evolution of the solar active regions. The chapter is based on the work of Wang and Hero (2021b) that was published in the *Proceedings of the* 38^{th} *International Conference on Machine Learning*.

Chapter IV connects Kronecker-structured (inverse) covariance modeling and spatio temporal PDEs via the ensemble Kalman filter (EnKF) framework for data assimilation. A new EnKF algorithm is introduced and the emergence of sparsity and multiway structures in second-order statistical characterizations of dynamical processes governed by PDEs is studied. We demonstrate promises of the new approach for tracking complex spatio-temporal systems. The chapter is based on the work of Wang and Hero (2021a) presented in the *Workshop on Machine Learning and the Physical Sciences at the* 35^{th} *Conference on Neural Information Processing Systems* and the joint work with Zeyu Sun, Dogyoon Song, and Alfred Hero that was under revision at *Statistics Surveys*. Additionally, a `Julia` software package called `TensorGraphicalModels` (Wang et al., 2022) has been developed to accompany this work.

Chapter V introduces a simple and modular approach for modeling time-varying

texts that combines standard LDA, shortest path algorithms on neighborhood graphs, and geometric embedding. This approach enables interpretation and visualization of latent thematic information that are intrinsically temporally dependent. We demonstrate that the framework is able to capture perceptually natural temporal trajectories of latent topics with minimal modeling assumptions. Further, we show that the framework is able to incorporate side information (e.g., labels) via weak supervision. Two important applications are considered: analysis of Twitter data for understanding COVID-19 related public discourse; and analysis of TalkLife data for understanding mental health related issues and aiding early detection and intervention. The work is partially based on the work of Wang et al. (2021) published in the *Harvard Data Science Review*.

CHAPTER II

The Sylvester Graphical Lasso

In this chapter we introduce 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) which imposes a Kronecker sum model for the inverse covariance matrix, by providing an alternative Kronecker sum model that is generative and interpretable. The interpretability follows from the Sylvester generative model on which SyGlasso is based: the model is exact for any observation process that is a solution of a diffusion-based partial differential equation. 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 the brain connectivity and its temporal dependencies.

2.1 Introduction

To address the sample complexity challenges that arise in modern multivariate analysis of tensor-variate data, sparsity can be imposed on the second order information - the covariance $\mathbf{\Sigma}$ or the inverse covariance $\mathbf{\Omega}$ - by using a sparse Kronecker product (KP) or Kronecker sum (KS) decomposition of $\mathbf{\Sigma}$ or $\mathbf{\Omega}$. The earliest and most popular form of sparse structured precision matrix estimation approaches represent $\mathbf{\Omega}$ as the KP of smaller precision matrices, which means that the resulting $\mathbf{\Sigma}$ also composes of KP of smaller covariance matrices due to the property of KP. Tsiligkaridis et al. (2013); Zhou (2014); Lyu et al. (2019) have developed estimation and statistical inference procedures under the KP structure and showed that the underlying true precision matrix can be estimated efficiently with high-dimensional consistency guarantees with single matrix or tensor sample. Alternatively, Kalaitzis et al. (2013); Greenewald et al. (2019) propose to model conditional dependency structures of precision matrices by using a KS representation. Rudelson and Zhou (2017); Park et al. (2017) studied the KS structure on the covariance matrix which corresponds to errors-in-variables models.

KP vs KS: One of the advantages of the KP model is that it admits a simple stochastic representation as $\mathbf{X} = \mathbf{C}^{-1}\mathbf{Z}\mathbf{D}^{-1}$, where $\mathbf{A} = \mathbf{C}\mathbf{C}^T, \mathbf{B} = \mathbf{D}\mathbf{D}^T$, and $\mathbf{Z}$ is white Gaussian. It can be shown using properties of KP that $\mathbf{X} \sim \mathcal{N}(0, (\mathbf{A} \otimes \mathbf{B})^{-1})$. Unlike the KP model, the KS model does not have a simple stochastic representation. From another perspective, the Kronecker structures can be characterized by different types of product graphs of the individual component graphs. Specifically, Kalaitzis et al. (2013) relates $(\mathbf{\Psi}_1 \oplus \cdots \oplus \mathbf{\Psi}_K)$ to the associated Cartesian product graph. As a result, the overall number of edges (active conditional dependencies) is additive in the number of edges in the individual graphs. The KP, however, corresponds to the direct tensor product of the individual graphs and leads to a denser dependency structure in the precision matrix, as the number of overall edges is multiplicative in

the number of individual edges [1].

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 KS or KP model for the precision matrix, the Sylvester structured graphical model uses a KS model for the square root factor of the precision matrix. The model is estimated by joint sparse regression models that impose sparsity on the individual components $\mathbf{\Psi}_k$ for $k = 1, \ldots, K$. The Sylvester model reduces to a squared KS 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. SyGlasso is the first KS-based graphical lasso model that admits a stochastic representation (i.e., Sylvester). Thus, our proposed SyGlasso puts the KS representations on similar ground as the KP representations in terms of interpretablility.

2.1.1 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 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 $\text{vec}(\boldsymbol{\mathcal{X}}) := (\boldsymbol{\mathcal{X}}_{1,1,\ldots,1}, \boldsymbol{\mathcal{X}}_{2,1,\ldots,1}, \ldots, \boldsymbol{\mathcal{X}}_{m_1,1,\ldots,1}, \boldsymbol{\mathcal{X}}_{1,2,\ldots,1}, \ldots, \boldsymbol{\mathcal{X}}_{m_1,m_2,\ldots,m_k})^T \in \mathbb{R}^p$ with $p = \prod_{k=1}^K m_k$.

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

[1] From Greenewald et al. (2019) KS (Cartesian product graph) edges: $\sum_{k=1}^K |E_k| \prod_{i \neq k} |V_i|$; KP (direct product graph) edges: $\frac{1}{2} \prod_{k=1}^K (2|E_k| + |V_k|) - \prod_{k=1}^K |V_k|$; where E_k and V_k denote the edge and vertex sets, respectively for component k.

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 $\bigotimes_{k=1}^{K} \boldsymbol{\Psi}_k = \boldsymbol{\Psi}_1 \otimes \cdots \otimes \boldsymbol{\Psi}_K$, and the equivalent notation for the Kronecker sum as $\bigoplus_{k=1}^{K} \boldsymbol{\Psi}_k = \boldsymbol{\Psi}_1 \oplus \cdots \oplus \boldsymbol{\Psi}_K = \sum_{k=1}^{K} \mathbf{I}_{[m_{k+1:K}]} \otimes \boldsymbol{\Psi}_k \otimes \mathbf{I}_{[m_{1:k-1}]}$, where $\mathbf{I}_{[m_{k:\ell}]} = \mathbf{I}_{m_k} \otimes \cdots \otimes \mathbf{I}_{m_\ell}$.

2.1.2 Outline

We briefly outline the structure of this chapter. Section 2.2 introduces the Sy-Glasso method in details. Section 2.3 studies the statistical convergence of the Sy-Glasso. Section 2.4 provides numerical illustrations of the method using synthetic data. Section 2.5 provides numerical illustrations of the method using real data that arises from Solar flare prediction problems. Section 2.6 concludes the chapter.

2.2 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:

$$\boldsymbol{\mathcal{X}} \times_1 \boldsymbol{\Psi}_1 + \cdots + \boldsymbol{\mathcal{X}} \times_K \boldsymbol{\Psi}_K = \boldsymbol{\mathcal{T}}, \tag{2.1}$$

where $\boldsymbol{\Psi}_k \in \mathbb{R}^{m_k \times m_k}, k = 1, \ldots, K$ are sparse symmetric positive definite matrices and $\boldsymbol{\mathcal{T}}$ is a random tensor of the same order as $\boldsymbol{\mathcal{X}}$. Equation (2.1) 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 $\mathbf{\Psi_1 X} + \mathbf{X \Psi_2}^T = \mathbf{T}$ 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 (2.1) is equivalent to the following system of linear equations:

$$\left(\bigoplus_{k=1}^{K} \mathbf{\Psi}_k \right) \operatorname{vec}(\boldsymbol{\mathcal{X}}) = \operatorname{vec}(\boldsymbol{\mathcal{T}}), \tag{2.2}$$

If $\boldsymbol{\mathcal{T}}$ is a random tensor such that $\operatorname{vec}(\boldsymbol{\mathcal{T}})$ has zero mean and identity covariance, it follows from (2.2) that any $\boldsymbol{\mathcal{X}}$ generated from the stochastic relation (2.1) 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 (2.1). 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 (2.2) element-wise, we first observe that

$$\begin{aligned} &\left(\sum_{k=1}^{K} (\mathbf{\Psi}_k)_{i_k, i_k} \right) \boldsymbol{\mathcal{X}}_{i_{[1:K]}} \\ &= -\sum_{k=1}^{K} \sum_{j_k \neq i_k} (\mathbf{\Psi}_k)_{i_k, j_k} \boldsymbol{\mathcal{X}}_{i_{[1:k]}, j_k, i_{[k+1:K]}} + \boldsymbol{\mathcal{T}}_{i_{[1:K]}}. \end{aligned} \tag{2.3}$$

Note that the left-hand side of (2.3) 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 (2.3) can be interpreted as an autogregressive model re-

lating 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 $\boldsymbol{\Psi}'s$ act as regression coefficients. The formulation in (2.3) naturally leads us to consider a pseudolikelihood-based estimation procedure (Besag, 1977) for estimating $\boldsymbol{\Omega}$. It is known that inference using pseudo-likelihood is consistent and enjoys the same $\sqrt{N}$ convergence rate as the MLE in general (Varin et al., 2011). This procedure can also be more robust to model misspecification. Specifically, we define the sparse estimate of the underlying precision matrices along each axis of the data as the solution of the following convex optimization problem:

$$\begin{aligned} \min_{\substack{\boldsymbol{\Psi}_k \in \mathbb{R}^{m_k \times m_k} \\ k=1,\ldots K}} & -N \sum_{i_1,\ldots,i_K} \log \boldsymbol{\mathcal{W}}_{i_{[1:K]}} \\ & + \frac{1}{2} \sum_{i_1,\ldots,i_K} \|(I) + (II)\|_2^2 + \sum_{k=1}^{K} P_{\lambda_k}(\boldsymbol{\Psi}_k). \end{aligned} \tag{2.4}$$

where $P_{\lambda_k}(\cdot)$ is a penalty function indexed by the tuning parameter λ_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 ℓ_1-norm penalty, i.e., $P_{\lambda_k}(\boldsymbol{\Psi}_k) = \lambda_k \|\boldsymbol{\Psi}_k\|_{1,\text{off}}$.

The optimization problem (2.4) can be put into the following matrix form:

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

where $\operatorname{diag}(\boldsymbol{\Psi}_k) \in \mathbb{R}^{m_k \times m_k}$ 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., $\mathbf{S} = \frac{1}{N}\operatorname{vec}(\boldsymbol{\mathcal{X}})^T \operatorname{vec}(\boldsymbol{\mathcal{X}})$. Note that the pseudolikelihood above approximates the ℓ_1-penalized Gaussian negative log-likelihood 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 (2.4) 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 (2.1) holds.

Remark II.1. *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* $\boldsymbol{\Omega}$*. To address this non-identifiability, we rewrite the overall precision matrix* $\boldsymbol{\Omega}$ *as*

$$\boldsymbol{\Omega} = \left(\bigoplus_{k=1}^{K} \boldsymbol{\Psi}_k\right)^2 = \left(\bigoplus_{k=1}^{K} \boldsymbol{\Psi}_k^{off} + \bigoplus_{k=1}^{K} \operatorname{diag}(\boldsymbol{\Psi}_k)\right)^2,$$

where $\boldsymbol{\Psi}_k^{off} = \boldsymbol{\Psi}_k - diag(\boldsymbol{\Psi}_k)$*, and estimate the diagonal and off-diagonal entries* $\boldsymbol{\Psi}_k$*'s separately. This allows us to reconstruct the overall precision matrix* $\boldsymbol{\Omega}$ *when* $\boldsymbol{\Psi}_k^{off}$ *is penalized with an* ℓ_1 *penalty.*

2.2.1 Estimation of the graphical model

Let $Q_N(\boldsymbol{\mathcal{W}}, \{\boldsymbol{\Psi}_k^{\text{off}}\}_{k=1}^K)$ denote the objective function defined in (2.4). Here, $\boldsymbol{\mathcal{W}} = \bigoplus_{k=1}^K \operatorname{diag}(\boldsymbol{\Psi}_k)$. We adopt a convergent alternating minimization approach (Khare and Rajaratnam, 2014) that cycles between optimizing $\boldsymbol{\Psi}_k$ 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

$$\begin{aligned} T_{i_k j_k}(\boldsymbol{\Psi}_k^{\text{off}}) &= \operatorname{argmin}_{\substack{(\tilde{\boldsymbol{\Psi}}_l)_{m,n}=(\boldsymbol{\Psi}_l)_{m,n} \\ \forall (l,m,n)\neq(k,i_k,j_k)}} Q_N(\tilde{\boldsymbol{\mathcal{W}}}, \{\tilde{\boldsymbol{\Psi}}_k^{\text{off}}\}_{k=1}^K) \\ T(\boldsymbol{\mathcal{W}}) &= \quad \operatorname{argmin}_{\substack{\tilde{\boldsymbol{\Psi}}_k^{\text{off}}=\boldsymbol{\Psi}_k^{\text{off}} \\ \forall k}} \quad Q_N(\tilde{\boldsymbol{\mathcal{W}}}, \{\tilde{\boldsymbol{\Psi}}_k^{\text{off}}\}_{k=1}^K). \end{aligned} \tag{2.5}$$

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

Algorithm II.1: Nodewise SyGlasso

Input: Standardized data $\boldsymbol{\mathcal{X}}$, penalty parameter λ_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 (1.1) in Appendix A

      end

    end

  end

  # *Update diagonal elements*;

  $\hat{\boldsymbol{\mathcal{W}}}^{(\text{t}+1)} \leftarrow T(\boldsymbol{\mathcal{W}}^{(\text{t})})$ from (1.2) in Appendix A

end

2.3 Large Sample Properties

We show that under suitable conditions, the Sylvester graphical lasso (SyGlasso) estimator (Algorithm II.1) achieves both model selection consistency and estimation

consistency. As in other studies (Khare et al., 2015; Peng et al., 2009)[2], for the convergence analysis 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 II.1, 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 ℓ_1 penalized problem:

$$\min_{\boldsymbol{\beta}} L_N\Big(\boldsymbol{\mathcal{W}}, \boldsymbol{\beta}, \boldsymbol{\mathcal{X}}\Big) + \sum_{k=1}^{K} \lambda_k \|\boldsymbol{\Psi}_k\|_{1,\text{off}}, \tag{2.6}$$

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

$$\begin{aligned} L\Big(\boldsymbol{\mathcal{W}}, \boldsymbol{\beta}, \boldsymbol{\mathcal{X}}^s\Big) = -N \sum_{i_{[1:K]}} \log \boldsymbol{\mathcal{W}}_{i_{[1:K]}} \\ + \frac{1}{2} \sum_{i_1,\ldots,i_K} ((I) + (II))^2. \end{aligned} \tag{2.7}$$

where

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

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 $\mathcal{A}_k := \{(i,j) : (\boldsymbol{\Psi}_k)_{i,j} \neq 0, i \neq j\}$ and $q_k := |\mathcal{A}_k|$ for $k = 1, \ldots, K$ be the true edge set and the number of edges, respectively. Let

[2]When $K = 1$ it is possible to relax this assumption to require only accurate estimates of the diagonals, see Khare et al. (2015); Peng et al. (2009) for details.

$\mathcal{A} = \cup_{k=1}^{K} \mathcal{A}_k$. We use $\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\Omega}}, \bar{\boldsymbol{\mathcal{W}}}$ to emphasize that they are the true values of the corresponding parameters.

(A1 - Subgaussianity) The data $\{\boldsymbol{\mathcal{X}}^s\}_{s=1}^N$ are i.i.d subgaussian random tensors, that is, $\text{vec}(\boldsymbol{\mathcal{X}}^s) \sim \mathbf{x}$, where $\mathbf{x}$ is a subgaussian random vector in $\mathbb{R}^p$, i.e., there exist a constant $c > 0$, such that for every $\mathbf{a} \in \mathbb{R}^p$, $\mathbb{E}e^{\mathbf{a}^T\mathbf{x}} \leq e^{c\mathbf{a}^T\bar{\boldsymbol{\Sigma}}\mathbf{a}}$, 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 $(i, j) \in \mathcal{A}_k$

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

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 when this condition is satisfied.

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

1. Theorem 2.3.1 establishes estimation consistency and sign consistency for the nodewise SyGlasso restricted to the true support, i.e., $\boldsymbol{\beta}_{\mathcal{A}^c} = 0$,

2. Theorem 2.3.2 shows that no wrong edge is selected with probability tending to one,

3. Theorem 2.3.3 establishes consistency result of the nodewise SyGlasso.

Theorem 2.3.1. *Suppose that conditions (A1-A2) are satisfied. Suppose further that* $\lambda_{N,k} = O(\sqrt{\frac{m_k \log p}{N}})$ *for all* k *and* $N > O(\max_k q_k m_k \log p)$ *as* $N \to \infty$. *Then there exists a constant* $C(\bar{\boldsymbol{\beta}})$, *such that for any* $\eta > 0$, *the following hold with probability at least* $1 - O(\exp(-\eta \log p))$*:*

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

$$\min_{\boldsymbol{\beta}:\boldsymbol{\beta}_{\mathcal{A}^c}=0} L_N(\bar{\boldsymbol{\mathcal{W}}}, \boldsymbol{\beta}, \boldsymbol{\mathcal{X}}) + \sum_{k=1}^{K} \lambda_k \|\boldsymbol{\Psi}_k\|_{1,off}. \tag{2.8}$$

- *(Estimation consistency) Any solution* $\hat{\boldsymbol{\beta}}_{\mathcal{A}}$ *of* (2.8) *satisfies:*

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

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

Theorem 2.3.2. *Suppose that the conditions of Theorem 2.3.1 and (A3) are satisfied. Suppose further that* $p = O(N^{\kappa})$ *for some* $\kappa \geq 0$. *Then for* $\eta > 0$, *for* N *sufficiently large, the solution of* (2.8) *satisfies:*

$$P_{\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}}}\Big(\max_{(i,j)\in\mathcal{A}_k^c} |L'_{N,ij}(\bar{\boldsymbol{\mathcal{W}}}, \hat{\boldsymbol{\beta}}_{\mathcal{A}_k}, \boldsymbol{\mathcal{X}})| < \lambda_{N,k} \Big)$$
$$\geq 1 - O(\exp(-\eta \log p))$$

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

Theorem 2.3.3. *Assume the conditions of Theorem 2.3.2. Then there exists a constant* $C(\bar{\boldsymbol{\beta}}) > 0$ *such that for any* $\eta > 0$ *the following events hold with probability at least* $1 - O(\exp(-\eta \log p))$*:*

- *There exists a global minimizer* $\hat{\boldsymbol{\beta}}$ *to problem* (2.4).
- *(Estimation consistency) Any minimizer* $\hat{\boldsymbol{\beta}}$ *of* (2.4) *satisfies:*

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

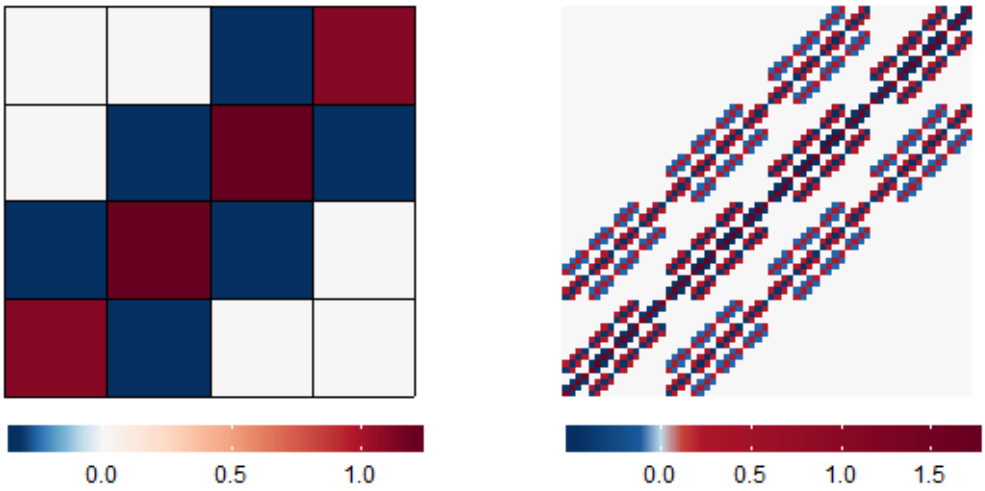
- *(Sign consistency) If further the minimal signal strength:* $\min_{(i,j)\in\mathcal{A}_k} |(\boldsymbol{\Psi}_k)_{i,j}| \geq 2C(\bar{\boldsymbol{\beta}}) \max_k \sqrt{q_k}\lambda_{N,k}$ *for each* k*, then sign(*$\hat{\boldsymbol{\beta}}$*)=sign(*$\bar{\boldsymbol{\beta}}$*).*

Proofs of the above theorems are given in Appendix A.

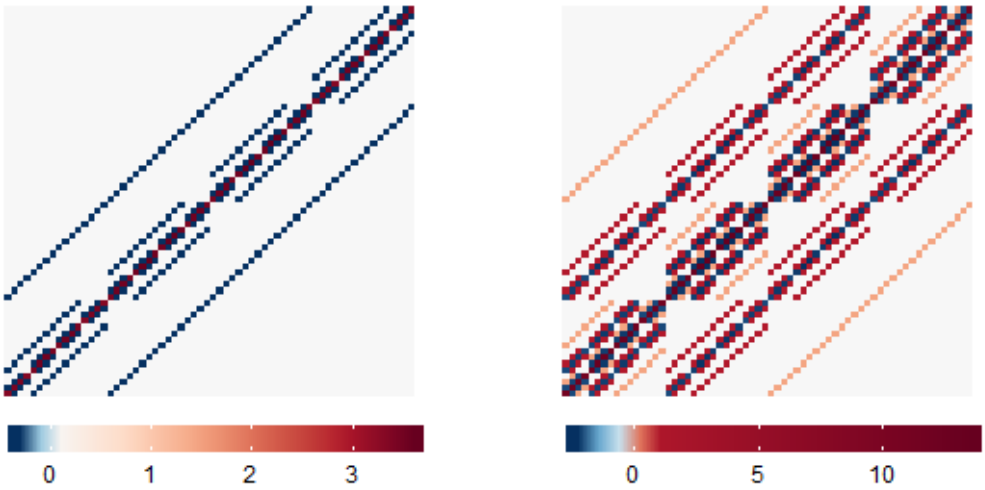
2.4 Numerical Illustrations

We evaluate the proposed SyGlasso estimator 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 the sparsity pattern of $\boldsymbol{\Omega}$ with $K = 3$ and $m_k = 4, \forall k$. For simplicity, we generate $\boldsymbol{\Psi}_k$ for $k = 1, 2, 3$ as identical 4×4 precision matrices that follow a one dimensional autoregressive-1 (AR1) process. We recall the KP and KS models:

Kronecker Product (KP): The KP model restricts the precision matrix and the covariance 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 $\boldsymbol{\Omega}$ 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 II.1 (c) and (d).



(a) $\boldsymbol{\Psi}_k$ (b) KP $\boldsymbol{\Omega}$

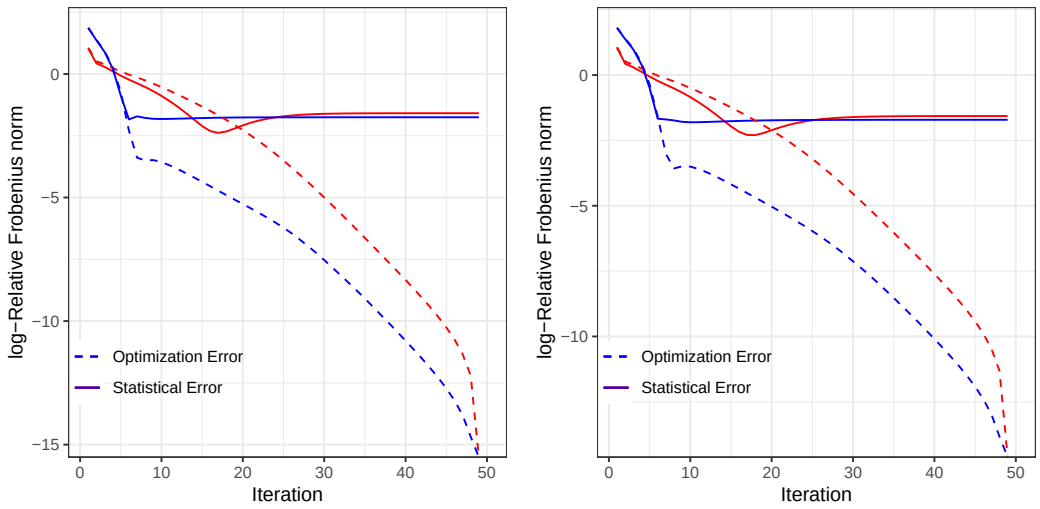


(c) KS $\boldsymbol{\Omega}$ (d) SyGlasso $\boldsymbol{\Omega}$

Figure II.1: Comparison of SyGlasso to Kronecker sum (KS) and product (KP) structures. All models are composed of the same components $\boldsymbol{\Psi}_k$ 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 64×64 precision matrix $\boldsymbol{\Omega}$ following (b) the KP structure with $\boldsymbol{\Omega} = \bigotimes_{k=1}^{3} \boldsymbol{\Psi}_k$, (c) the KS structure with $\boldsymbol{\Omega} = \bigoplus_{k=1}^{3} \boldsymbol{\Psi}_k$, and (d) the proposed Sylvester model with $\boldsymbol{\Omega} = \left(\bigoplus_{k=1}^{3} \boldsymbol{\Psi}_k\right)^2$. 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 Sum (KS): The covariance matrix under the KS precision matrix assumption is nonseparable across K data dimensions, and the KS-structured models can be motivated from a maximum entropy point of view. Contrary to the KP structure, the number of edges in the KS structure grows as the sum of the edges of the individual graphs (as a result of Cartesian product of the associated graphs), which leads to a more controllable number of edges in $\mathbf{\Omega}$.

We compare these methods under different model assumptions to explore the flexibility of the proposed SyGlasso model under model mismatch. 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(ρ), 2) Star-Block (SB), or 3) Erdos-Renyi (ER) random graph models described in Appendix A.

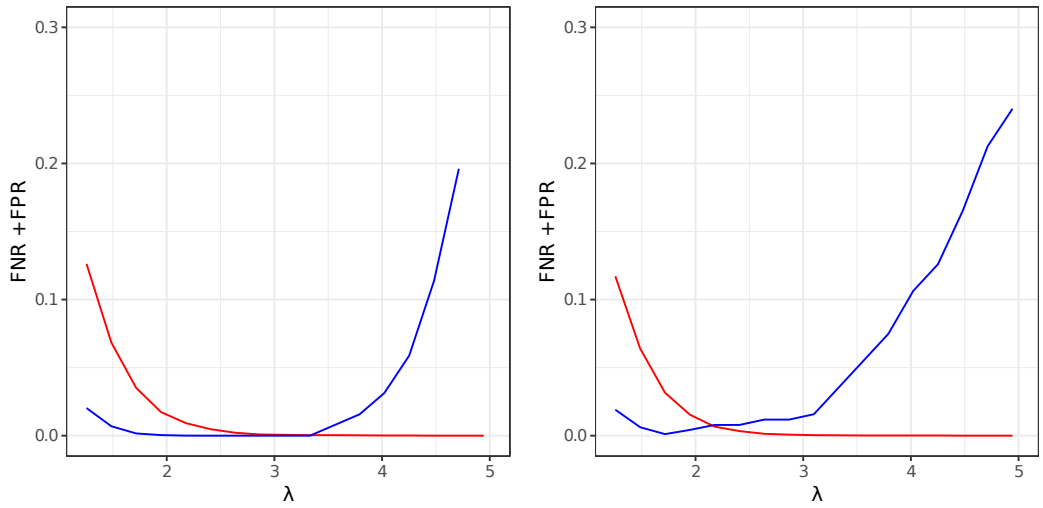


(a) SB and AR (b) SB and ER

Figure II.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 $\log\left(\|\hat{\mathbf{\Psi}}_k^{(t)} - \mathbf{\Psi}_k\|_F \backslash \|\mathbf{\Psi}_k\|_F\right)$, and the dotted line shows the optimization error $\log\left(\|\hat{\mathbf{\Psi}}_k^{(t)} - \hat{\mathbf{\Psi}}_k\|_F \backslash \|\hat{\mathbf{\Psi}}_k\|_F\right)$, 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 with $K = 2$ 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 II.2 shows the iterative optimization performance of Algorithm II.1. All the plots for the various scenarios exhibit iterative optimization approximation errors that quickly converge to values 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 II.3 displays the sum of false positive rate and false negative rate (FPR+FNR), it 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 II.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 λ, 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 16×16 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 $\boldsymbol{\Psi}_k$ as a single edge.

The results shown in Figure II.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 II.1. This follows from the observation that $(\boldsymbol{\Psi}_1 \oplus \boldsymbol{\Psi}_2)^2 = \mathbf{I}_{m_1} \otimes \boldsymbol{\Psi}_1^2 + \boldsymbol{\Psi}_2^2 \otimes \mathbf{I}_{m_2} + 2\boldsymbol{\Psi}_1 \otimes \boldsymbol{\Psi}_2 = \boldsymbol{\Psi}_1^2 \oplus \boldsymbol{\Psi}_2^2 + 2\boldsymbol{\Psi}_1 \otimes \boldsymbol{\Psi}_2$.

2.5 EEG Analysis

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 α 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 α-band signals. Previous Gaussian graphical

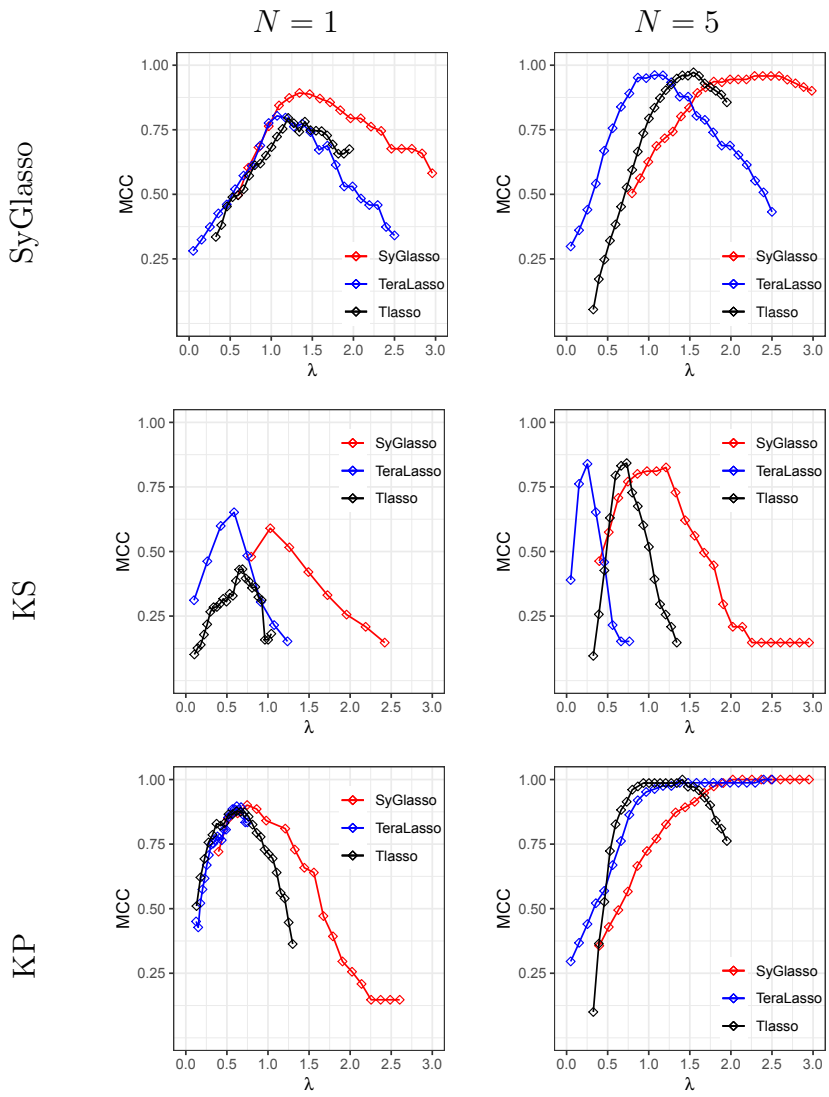
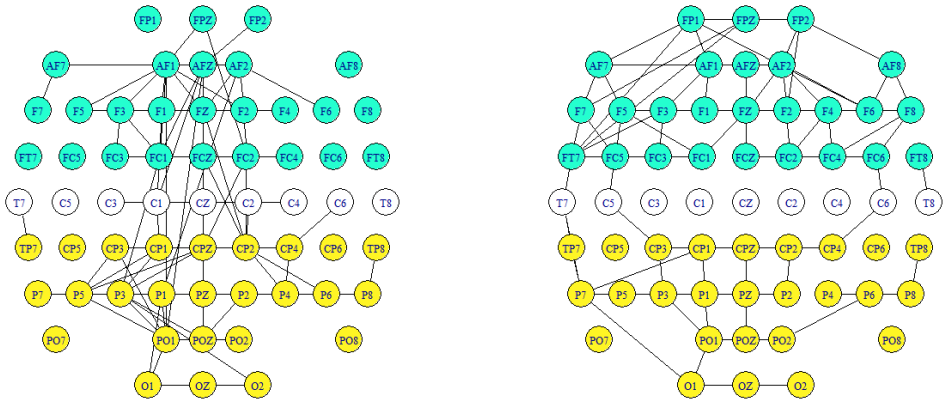


Figure II.4: Performance of SyGlasso, TeraLasso (KS), and Tlasso (KP) measured by MCC under model misspecification. MCC of 1 represents a perfect recovery of the sparsity pattern in $\mathbf{\Omega}$, and MCC of 0 corresponds to random guess. From top to bottom, the synthetic data were generated with 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. Note that the SyGlasso has better performance for a single sample (left column) when data is generated from the matched Kronecker model and as good performance for the mismatched Kronecker models.

models applied to such α 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 similar dependency structure as reported in Zhu et al. (2016)

and Qiao et al. (2019) while recovering the chain structure of the temporal relationship.

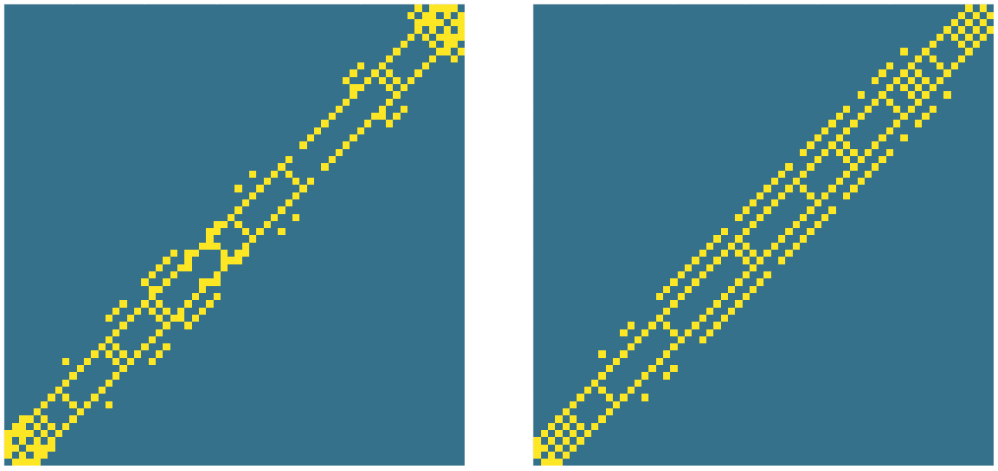
Specifically, after the band-pass filter was applied, we work with the tensor data $\boldsymbol{\mathcal{X}}_{alcoholic}, \boldsymbol{\mathcal{X}}_{control} \in \mathbb{R}^{m_{nodes} \times m_{time} \times m_{trial}}$ corresponding to an alcoholic subject and a control subject. We simultaneously estimate $\boldsymbol{\Psi}_{node} \in \mathbb{R}^{m_{node} \times m_{node}}$ that encodes the dependency structure among electrodes and $\boldsymbol{\Psi}_{time} \in \mathbb{R}^{m_{time} \times m_{time}}$ 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 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.



(a) Alcoholic subject (b) Control subject

Figure II.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.

Figure II.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 also consistent with the result in Hayden et al. (2006) and Zhu et al. (2016) that showed frontal asymmetry of the alcoholic subjects.



(a) Alcoholic subject (b) Control subject

Figure II.6: Support (off-diagonals) of SyGlasso-estimated temporal Sylvester factor $\hat{\boldsymbol{\Psi}}_{time}$ of the precision matrix for (a) the alcoholic subject and (b) the control subject. Both subjects exhibit banded conditional dependency structures over time.

While previous analyses on this EEG data using graphical models only focused on the precision matrix of the electrodes, here we exhibit in Figure II.6 the second precision matrix that encodes temporal dependency. Overall both subjects exhibit banded dependency structures over time, since adjacent timepoints are conditionally dependent. However, note that the conditional dependency structure of the timepoints for the alcoholic subject appears to be more chaotic.

2.6 Conclusion

This chapter proposed a Sylvester-structured graphical model and an inference algorithm, the SyGlasso, that can be applied to tensor-valued data. The current frameworks 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.

There are several promising future directions. First is to relax the assumption that the diagonals of the factors are fixed - an assumption that is standard among the Kronecker structured models for theoretical analysis. Practically, SyGlasso is able to recover the off-diagonals of the individual $\mathbf{\Psi}_k$ and the diagonal of $\mathbf{\Omega}$, which only requires to estimating $\bigoplus_{k=1}^{K} \text{diag}(\mathbf{\Psi}_k)$ instead of all diagonal entries $\text{diag}(\mathbf{\Psi}_k)$ for all k. Secondly, in terms of the statistical properties, our theoretical results guarantee sparsistency of the individual graphs with a slower convergence rate than that is proposed in Greenewald et al. (2019), while empirical evidence suggests that a faster rate can be achieved. Improvement of this statistical convergence rate analysis will be worthwhile. Also, our results do not guarantee statistical convergence of individual $\mathbf{\Psi}_k$'s nor $\mathbf{\Omega}$ with respect to the operator norm. Similar to the solution proposed in Zhou et al. (2011), we plan to adopt a two-step procedure using SyGlasso for variable selection followed by refitting the precision matrix $\mathbf{\Omega}$ using maximum likelihood estimation with edge constraint.

CHAPTER III

A Proximal Alternating Linearized Minimization Method for Tensor Graphical Models

In this chapter, we extend the Sylvester graphical model introduced in Chapter II to incorporate a new inference procedure, called SG-PALM, for learning conditional dependency structure of high-dimensional tensor-variate data. Unlike the SyGlasso, the new method is computationally scalable to ultra-high dimension. Scalability of SG-PALM follows from the fast proximal alternating linearized minimization (PALM) procedure that SG-PALM uses during training. We establish that SG-PALM converges linearly (i.e., geometric convergence rate) to a global optimum of its objective function. We demonstrate the scalability and accuracy of SG-PALM for an important but challenging climate prediction problem: spatio-temporal forecasting of solar flares from multimodal imaging data.

3.1 Introduction

A common challenge for structured tensor graphical models is the efficient estimation of the underlying (conditional) dependency structures. KP-structured models are generally estimated via extension of GLasso (Friedman et al., 2008) that iteratively minimize the ℓ_1-penalized negative likelihood function for the matrix-normal

data with KP covariance. This procedure was shown to converge to some local optimum of the penalized likelihood function (Yin and Li, 2012; Tsiligkaridis et al., 2013). Similarly, Kalaitzis et al. (2013) further extended GLasso to the KS-structured case for 2-way tensor data. Greenewald et al. (2019) extended this to multiway tensors, exploiting the linearity of the space of KS-structured matrices and developing a projected proximal gradient algorithm for KS-structured inverse covariance matrix estimation, which achieves linear convergence (i.e., geometric convergence rate) to the global optimum. In Chapter II, the Sylvester-structured graphical model is estimated via a nodewise regression approach inspired by algorithms for estimating a class of vector-variate graphical models (Meinshausen and Bühlmann, 2006; Khare et al., 2015). However, no theoretical convergence result for the algorithm was established nor did they study the computational efficiency of the algorithm.

In the modern era of big data, both computational and statistical learning accuracy are required of algorithms. Furthermore, when the objective is to learn representations for physical processes, interpretablility is crucial. In this chapter, we bridge this "Statistical-to-Computational-to-Interpretable gap" for Sylvester graphical models. We develop a simple yet powerful first-order optimization method, based on the Proximal Alternating Linearized Minimization (PALM) algorithm, for recovering the conditional dependency structure of such models. Moreover, we provide the link between the Sylvester graphical models and physical processes obeying differential equations and illustrate the link with a real-data example. The following are our principal contributions:

1. A fast algorithm that efficiently recovers the generating factors of a representation for high-dimensional multiway data, significantly improving on the Sy-Glasso algorithm described in Chapter II.

2. A comprehensive convergence analysis showing linear convergence of the objective function to its global optimum and providing insights for choices of

hyperparameters.

3. A novel application of the algorithm to an important multi-modal solar flare prediction problem from solar magnetic field sequences. For such problems, SG-PALM is physically interpretable in terms of the Poisson differential equation for solar magnetic induction fields proposed by heliophysicists.

3.2 Background and Notation

3.2.1 Notations

In this chapter, scalar, vector and matrix quantities are denoted by lowercase letters, boldface lowercase letters and boldface capital letters, respectively. For a matrix $\mathbf{A} = (\mathbf{A}_{i,j}) \in \mathbb{R}^{d\times d}$, we denote $\|\mathbf{A}\|_2, \|\mathbf{A}\|_F$ as its spectral and Frobenius norm, respectively. We define $\|\mathbf{A}\|_{1,\text{off}} := \sum_{i\neq j} |\mathbf{A}_{i,j}|$ as its off-diagonal ℓ_1 norm. For tensor algebra, we adopt the notations used by Kolda and Bader (2009). A K-th order tensor is denoted by boldface Euler script letters, e.g, $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1\times\cdots\times d_K}$. The $(i_1,\ldots,i_K)$-th element of $\boldsymbol{\mathcal{X}}$ is denoted by $\boldsymbol{\mathcal{X}}_{i_1,\ldots,i_K}$, and the vectorization of $\boldsymbol{\mathcal{X}}$ is the d-dimensional vector $\text{vec}(\boldsymbol{\mathcal{X}}) := (\boldsymbol{\mathcal{X}}_{1,1,\ldots,1}, \boldsymbol{\mathcal{X}}_{2,1,\ldots,1}, \ldots, \boldsymbol{\mathcal{X}}_{d_1,1,\ldots,1}, \ldots, \boldsymbol{\mathcal{X}}_{d_1,d_2,\ldots,d_k})^T$ with $d = \prod_{k=1}^K d_k$. 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. 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. The k-mode product of a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1\times\cdots\times d_K}$ and a matrix $\mathbf{A} \in \mathbb{R}^{J\times d_k}$, denoted as $\boldsymbol{\mathcal{X}} \times_k \mathbf{A}$, is of size $d_1 \times \cdots \times d_{k-1} \times J \times d_{k+1} \times \ldots d_K$. Its entry is defined as $(\boldsymbol{\mathcal{X}} \times_k \mathbf{A})_{i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K} := \sum_{i_k=1}^{d_k} \boldsymbol{\mathcal{X}}_{i_1,\ldots,i_K} A_{j,i_k}$. For a list of matrices $\{\mathbf{A}_k\}_{k=1}^K$ with $\mathbf{A}_k \in \mathbb{R}^{d_k\times d_k}$, 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 $\bigotimes_{k=1}^K \mathbf{A}_k = \mathbf{A}_1 \otimes \cdots \otimes \mathbf{A}_K$, and the equivalent notation for the Kronecker sum as $\bigoplus_{k=1}^K \mathbf{A}_k = \mathbf{A}_1 \oplus \cdots \oplus \mathbf{A}_K =$

$\sum_{k=1}^{K} \mathbf{I}_{[d_{k+1:K}]} \otimes \mathbf{A}_k \otimes \mathbf{I}_{[d_{1:k-1}]}$, where $\mathbf{I}_{[d_{k:\ell}]} = \mathbf{I}_{d_k} \otimes \cdots \otimes \mathbf{I}_{d_\ell}$. For the case of $K = 2$, $\mathbf{A}_1 \oplus \mathbf{A}_2 = \mathbf{I}_{d_2} \otimes \mathbf{A}_1 + \mathbf{A}_2 \otimes \mathbf{I}_{d_1}$.

3.2.2 Tensor Gaussian graphical models

A random tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ follows the tensor normal distribution with zero mean when $\text{vec}(\boldsymbol{\mathcal{X}})$ follows a normal distribution with mean $\mathbf{0} \in \mathbb{R}^d$ and precision matrix $\boldsymbol{\Omega} := \boldsymbol{\Omega}(\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_K)$, where $d = \prod_{k=1}^{K} d_k$. Here, $\boldsymbol{\Omega}(\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_K)$ is parameterized by $\boldsymbol{\Psi}_k \in \mathbb{R}^{d_k \times d_k}$ via either Kronecker product, Kronecker sum, or the Sylvester structure, and the corresponding negative log-likelihood function (assuming N independent observations $\boldsymbol{\mathcal{X}}^i, i = 1, \ldots, N$)

$$-\frac{N}{2} \log |\boldsymbol{\Omega}| + \frac{N}{2} \text{tr}(\mathbf{S}\boldsymbol{\Omega}), \tag{3.1}$$

where $\boldsymbol{\Omega} = \bigotimes_{k=1}^{K} \boldsymbol{\Psi}_k$, $\bigoplus_{k=1}^{K} \boldsymbol{\Psi}_k$, or $\left(\bigoplus_{k=1}^{K} \boldsymbol{\Psi}_k \right)^2$ for KP, KS, and Sylvester models, respectively; and $\mathbf{S} = \frac{1}{N} \sum_{i=1}^{N} \text{vec}(\boldsymbol{\mathcal{X}}^i) \, \text{vec}(\boldsymbol{\mathcal{X}}^i)^T$. For $K = 1$, this formulation reduces to the vector normal distribution with zero mean and precision matrix $\boldsymbol{\Psi}_1$.

To encourage sparsity in the high-dimensional scenario, penalized negative log-likelihood function is proposed

$$-\frac{N}{2} \log |\boldsymbol{\Omega}| + \frac{N}{2} \text{tr}(\mathbf{S}\boldsymbol{\Omega}) + \sum_{k=1}^{K} P_{\lambda_k}(\boldsymbol{\Psi}_k),$$

where $P_{\lambda_k}(\cdot)$ is a penalty function indexed by the tuning parameter λ_k and is applied elementwise to the off-diagonal elements of $\boldsymbol{\Psi}_k$. Popular choices for $P_{\lambda_k}(\cdot)$ include the lasso penalty (Tibshirani, 1996), the adaptive lasso penalty (Zou, 2006), the SCAD penalty (Fan and Li, 2001), and the MCP penalty (Zhang et al., 2010).

3.2.3 The Sylvester generating equation

The Sylvester graphical model uses the Sylvester tensor equation to define a generative process for the underlying multivariate tensor data. The Sylvester tensor equation has been studied in the context of finite-difference discretization of high-dimensional elliptical partial differential equations (Grasedyck, 2004; Kressner and Tobler, 2010). Any solution $\boldsymbol{\mathcal{X}}$ to such a PDE must have the (discretized) form:

$$\sum_{k=1}^{K} \boldsymbol{\mathcal{X}} \times_k \boldsymbol{\Psi}_k = \boldsymbol{\mathcal{T}} \iff \Big(\bigoplus_{k=1}^{K} \boldsymbol{\Psi}_k \Big) \operatorname{vec}(\boldsymbol{\mathcal{X}}) = \operatorname{vec}(\boldsymbol{\mathcal{T}}). \tag{3.2}$$

where $\boldsymbol{\mathcal{T}}$ is the driving source on the domain, and $\bigoplus_{k=1}^{K} \boldsymbol{\Psi}_k$ is a Kronecker sum of $\boldsymbol{\Psi}_k$'s representing the discretized differential operators for the PDE, e.g., Laplacian, Euler-Lagrange operators, and associated coefficients. These operators are often sparse and structured.

For example, consider a physical process characterized as a function u that satisfies:

$$\mathcal{D}u = f \quad \text{in} \quad \Omega, \quad u(\Gamma) = 0, \quad \Gamma = \partial\Omega.$$

where f is a driving process, e.g., a Wiener process (white Gaussian noise); $\mathcal{D}$ is a differential operator, e.g, Laplacian, Euler-Lagrange; Ω is the domain; and Γ is the boundary of Ω. After discretization, this is equivalent to (ignoring discretization error) the matrix equation

$$\mathbf{D}\mathbf{u} = \mathbf{f}.$$

Here, $\mathbf{D}$ is a sparse matrix since $\mathcal{D}$ is an infinitesimal operator. Additionally, $\mathbf{D}$ admits Kronecker structure as a mixture of Kronecker sums and Kronecker products.

The matrix $\mathbf{D}$ reduces to a Kronecker sum when $\mathcal{D}$ involves no mixed derivatives. For instance, consider the Poisson's equation in 2D, where $u(x, y)$ on $[0, 1]^2$ satisfies

the elliptical PDE

$$\mathcal{D}u = (\partial_x^2 + \partial_y^2)u = f.$$

The Poisson equation governs many physical processes, e.g., electromagnetic induction, heat transfer, convection, etc. A simple Euler discretization yields $\mathbf{U} = (u(i,j))_{i,j}$, where $u(i,j)$ satisfies the local equation (up to a constant discretization scale factor)

$$\begin{aligned} 2u(i,j) = u(i+1,j) + u(i-1,j) + u(i,j+1) \\ + u(i,j-1) - 4f(i,j). \end{aligned}$$

Defining $\mathbf{u} = \text{vec}(\mathbf{U})$ and $\mathbf{A}$ (a tridiagonal matrix)

$$\mathbf{A} = \begin{bmatrix} -1 & 2 & -1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \end{bmatrix},$$

then $(\mathbf{A} \oplus \mathbf{A})\mathbf{u} = \mathbf{f}$, which is the Sylvester equation ($K = 2$).

For the Poisson example, if the source $\mathbf{f}$ is a white noise random variable, i.e., its covariance matrix is proportional to the identity matrix, then the inverse covariance matrix of $\mathbf{u}$ has sparse square-root factors, since $\text{Cov}^{-1}(\mathbf{u}) = (\mathbf{A} \oplus \mathbf{A})(\mathbf{A} \oplus \mathbf{A})^T$. Other physical processes that are generated from differential equations will also have sparse inverse covariance matrices, as a result of the sparsity of general discretized differential operators. Note that similar connections between continuous state physical processes and sparse "discretized" statistical models have been established by Lindgren et al. (2011), who elucidated a link between Gaussian fields and Gaussian Markov Random Fields via stochastic partial differential equations.

The Sylvester generative (SG) model (3.2) leads to a tensor-valued random variable $\boldsymbol{\mathcal{X}}$ with a precision matrix $\boldsymbol{\Omega} = \left(\bigoplus_{k=1}^{K} \boldsymbol{\Psi}_k \right)^2$, given that $\boldsymbol{\mathcal{T}}$ is white Gaussian.

The Sylvester generating factors $\mathbf{\Psi}_k$'s can be obtained via minimization of the penalized negative log-pseudolikelihood

$$\begin{aligned}\mathcal{L}_{\boldsymbol{\lambda}}(\mathbf{\Psi}) = &- \frac{N}{2}\log|(\bigoplus_{k=1}^{K}\operatorname{diag}(\mathbf{\Psi}_k))^2| \\ &+ \frac{N}{2}\operatorname{tr}(\mathbf{S}\cdot(\bigoplus_{k=1}^{K}\mathbf{\Psi}_k)^2) + \sum_{k=1}^{K}\lambda_k\|\mathbf{\Psi}_k\|_{1,\text{off}}.\end{aligned} \tag{3.3}$$

This differs from the penalized Gaussian negative log-likelihood in the exclusion of off-diagonals of $\mathbf{\Psi}_k$'s in the log-determinant term. (3.3) is motivated and derived directly using the Sylvester equation defined in (3.2), from the perspective of solving a sparse linear system. This maximum pseudolikelihood estimation procedure has been applied to vector-variate Gaussian graphical models (see Khare et al. (2015) and references therein for discussions). It is known that inference using pseudo-likelihood is consistent and enjoys the same $\sqrt{N}$ convergence rate as the MLE in general (Varin et al., 2011). This procedure can also be more robust to model misspecification. Detailed derivations are provided in Appendix 2.1.

3.3 The SG-PALM Method

Estimation of the generating parameters $\mathbf{\Psi}_k$'s of the SG model is challenging since the sparsity penalties are applied to the square root factors of the precision matrix and the likelihood function involves a mix of Kronecker sums and Kronecker products of matrix-valued parameters. The previously proposed estimation procedure called SyGlasso (see Chapter II), recovers only the off-diagonal elements of each Sylvester factor. This is a deficiency in many applications where the factor-wise variances are desired. Moreover, the convergence rate of the cyclic coordinate-wise algorithm used in SyGlasso is unknown and the computational complexity of the algorithm is higher than other sparse Glasso-type procedures. To overcome these deficiencies, we

propose a proximal alternating linearized minimization method, called SG-PALM, for finding the minimizer of (3.3). SG-PALM is designed to exploit structures of the coupled objective function and yields simultaneous estimates for both off-diagonal and diagonal entries.

The PALM algorithm was originally proposed to solve nonconvex optimization problems with separable structures, such as those arising in nonnegative matrix factorization (Xu and Yin, 2013; Bolte et al., 2014). Its efficacy in solving convex problems has also been established, for example, in regularized linear regression problems (Shefi and Teboulle, 2016), it was proposed as an attractive alternative to iterative soft-thresholding algorithms (ISTA). For simplicity, we consider the ℓ_1-regularized case (3.3), and the general, possibly non-convex, case is described in the supplement. The SG-PALM procedure is summarized in Algorithm III.1.

For clarity of notation we write

$$\mathcal{L}_{\boldsymbol{\lambda}}(\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_K) = H(\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_K) + \sum_{k=1}^{K} G_k(\boldsymbol{\Psi}_k), \tag{3.4}$$

where $H : \mathbb{R}^{d_1 \times d_1} \times \cdots \times \mathbb{R}^{d_K \times d_K} \to \mathbb{R}$ represents the log-determinant plus trace terms in (3.3) and $G_k : \mathbb{R}^{d_k \times d_k} \to (-\infty, +\infty]$ represents the penalty term in (3.3) for each axis $k = 1, \ldots, K$. For notational simplicity we use $\boldsymbol{\Psi}$ (i.e., omitting the subscript) to denote the set $\{\boldsymbol{\Psi}_k\}_{k=1}^K$ or the K-tuple $(\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_K)$ whenever there is no risk of confusion. The gradient of the smooth function H with respect to $\boldsymbol{\Psi}_k$, $\nabla_k H(\boldsymbol{\Psi})$, is given by

$$\begin{aligned} &\operatorname{diag}\Big(\Big\{\operatorname{tr}[(\operatorname{diag}((\boldsymbol{\Psi}_k)_{ii}) + \bigoplus_{j \neq k} \operatorname{diag}(\boldsymbol{\Psi}_j))^{-1}]\Big\}_{i=1}^{d_k}\Big) \\ &\quad + \mathbf{S}_k \boldsymbol{\Psi}_k + \boldsymbol{\Psi}_k \mathbf{S}_k + 2 \sum_{j \neq k} \mathbf{S}_{j,k}. \end{aligned} \tag{3.5}$$

Here, the first "diag" maps a d_k-vector to a $d_k \times d_k$ diagonal matrix, the second one maps a scalar (i.e., $(\boldsymbol{\Psi}_k)_{ii}$) to a $(\prod_{j \neq k} d_j) \times (\prod_{j \neq k} d_j)$ diagonal matrix with the same

elements, and the third operator maps a symmetric matrix to a matrix containing only its diagonal elements. In addition, we define:

$$\begin{aligned}
\mathbf{S}_k &= \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\mathcal{X}}^i_{(k)}(\boldsymbol{\mathcal{X}}^i_{(k)})^T, \\
\mathbf{S}_{j,k} &= \frac{1}{N}\sum_{i=1}^{N} \mathbf{V}^i_{j,k}(\mathbf{V}^i_{j,k})^T, \\
\mathbf{V}^i_{j,k} &= \boldsymbol{\mathcal{X}}^i_{(k)}\Big(\mathbf{I}_{d_{1:j-1}} \otimes \boldsymbol{\Psi}_j \otimes \mathbf{I}_{d_{j:K}}\Big)^T, \quad j \neq k.
\end{aligned} \tag{3.6}$$

A key ingredient of the PALM algorithm is a proximal operator associated with the non-smooth part of the objective, i.e., G_k's. In general, the proximal operator of a proper, lower semi-continuous convex function f from a Hilbert space $\mathcal{H}$ to the extended reals $(-\infty, +\infty]$ is defined by (Parikh and Boyd, 2014)

$$\text{prox}_f(v) = \text{argmin}_{x\in\mathcal{H}} f(x) + \frac{1}{2}\|x - v\|_2^2$$

for any $v \in \mathcal{H}$. The proximal operator well-defined as the expression on the right-hand side above has a unique minimizer for any function in this class. For ℓ_1-regularized case, the proximal operator for the function G_k is given by

$$\text{prox}^{\lambda_k}_{G_k}(\boldsymbol{\Psi}_k) = \text{diag}(\boldsymbol{\Psi}_k) + \text{soft}(\boldsymbol{\Psi}_k - \text{diag}(\boldsymbol{\Psi}_k), \lambda_k), \tag{3.7}$$

where the soft-thresholding operator $\text{soft}_\lambda(x) = \text{sign}(x)\max(|x| - \lambda, 0)$ has been applied element-wise.

3.3.1 Choice of step size

In the absence of a good estimate of the blockwise Lipchitz constant, the step size of each iteration of SG-PALM is chosen using backtracking line search, which, at iteration t, starts with an initial step size η_0^t and reduces the size with a constant

Algorithm III.1: SG-PALM

Input: Data tensor $\boldsymbol{\mathcal{X}}$, mode-k Gram matrix $\mathbf{S}_k$, regularizing parameter λ_k, backtracking constant $c \in (0,1)$, initial step size η_0, initial iterate $\boldsymbol{\Psi}_k$ for each $k = 1, \ldots, K$.

while not converged **do**

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

Line search:

Let η_k^t be the largest element of $\{c^j \eta_{k,0}^t\}_{j=1,\ldots}$ such that condition (3.8) is satisfied.

Update:

$\boldsymbol{\Psi}_k^{t+1} \leftarrow \text{prox}_{G_k}^{\eta_k^t \lambda_k}\Big(\boldsymbol{\Psi}_k^t - \eta_k^t \nabla_k H(\boldsymbol{\Psi}_{i<k}^{t+1}, \boldsymbol{\Psi}_{i\geq k}^t)\Big)$.

end for

Update initial step size: Compute Barzilai-Borwein step size $\eta_0^{t+1} = \min_k \eta_{k,0}^{t+1}$, where $\eta_{k,0}^{t+1}$ is computed via (3.9).

end while

Output: Final iterates $\{\boldsymbol{\Psi}_k\}_{k=1}^K$.

factor $c \in (0,1)$ until the new iterate satisfies the sufficient descent condition:

$$H(\boldsymbol{\Psi}_{i\leq k}^{t+1}, \boldsymbol{\Psi}_{i>k}^t) \leq Q_{\eta^t}(\boldsymbol{\Psi}_{i\leq k}^{t+1}, \boldsymbol{\Psi}_{i>k}^t; \boldsymbol{\Psi}_{i<k}^{t+1}, \boldsymbol{\Psi}_{i\geq k}^t). \tag{3.8}$$

Here,

$$\begin{aligned} &Q_\eta(\boldsymbol{\Psi}_{i<k}, \boldsymbol{\Psi}_k, \boldsymbol{\Psi}_{i>k}; \boldsymbol{\Psi}_{i<k}, \boldsymbol{\Psi}_k', \boldsymbol{\Psi}_{i>k}) \\ &= H(\boldsymbol{\Psi}_{i<k}, \boldsymbol{\Psi}_k, \boldsymbol{\Psi}_{i>k}) \\ &+ \text{tr}\Big((\boldsymbol{\Psi}_k' - \boldsymbol{\Psi}_k)^T \nabla_k H(\boldsymbol{\Psi}_{i<k}, \boldsymbol{\Psi}_k, \boldsymbol{\Psi}_{i>k})\Big) \\ &+ \frac{1}{2\eta}\|\boldsymbol{\Psi}_k' - \boldsymbol{\Psi}_k\|_F^2. \end{aligned}$$

The sufficient descent condition is satisfied with any $\frac{1}{\eta} = M_k$ and $M_k \geq L_k$, for any function that has a block-wise Lipschitz gradient with constant L_k for $k = 1, \ldots, K$. In other words, so long as the function H has block-wise gradient that is Lipschitz continuous with some block Lipschitz constant $L_k > 0$ for each k, then at each iteration t, we can always find an η^t such that the inequality in (3.8) is satisfied. Indeed, we proved in Lemma 2.3.3 in the Appendix that H has the desired properties. Additionally, in the proof of Theorem 3.4.2 we also showed that the step size found

at each iteration t satisfies $\frac{1}{\eta_k^0} \leq L_k \leq \frac{1}{\eta_k^t} \leq cL_k$.

In terms of the initialization, a safe step size (i.e., very small η_0^t) often leads to slower convergence. Thus, we use the more aggressive Barzilai-Borwein (BB) step (Barzilai and Borwein, 1988) to set a starting η_0^t at each iteration (see Appendix 2.2 for justifications of the BB method). In our case, for each k, the step size is given by

$$\eta_{k,0}^t = \frac{\|\boldsymbol{\Psi}_k^{t+1} - \boldsymbol{\Psi}_k^t\|_F^2}{\mathrm{tr}(\boldsymbol{A})}, \tag{3.9}$$

where

$$\begin{aligned}\boldsymbol{A} = (\boldsymbol{\Psi}_k^{t+1} - \boldsymbol{\Psi}_k^t)^T \times \\ (\nabla_k H(\boldsymbol{\Psi}_{i\leq k}^{t+1}, \boldsymbol{\Psi}_{i>k}^t) - \nabla_k H(\boldsymbol{\Psi}_{i<k}^{t+1}, \boldsymbol{\Psi}_{i\geq k}^t)).\end{aligned}$$

3.3.2 Computational complexity

After pre-computing $\mathbf{S}_k$, the most significant computation for each iteration in the SG-PALM algorithm is the sparse matrix-matrix multiplications $\mathbf{S}_k\boldsymbol{\Psi}_k$ and $\mathbf{S}_{j,k}$ in the gradient calculation. In terms of computational complexity, the former and latter can be computed using $O(d_k^3)$ and $O(N\sum_{j\neq k} d_j m_j^2)$ operations, respectively, there $m_j = \prod_{i\neq j} d_i$. Thus, each iteration of SG-PALM can be computed using $O\Big(\sum_{k=1}^K (d_k^3 + N\sum_{j\neq k} d_j m_j^2)\Big)$ floating point operations, which is significantly lower than competing methods.

Remark III.1. *All the structured precision estimation algorithms are variants of Glasso, implemented with techniques tailored to the model assumptions for speedup. Generally speaking, the resulting complexity consists of the mode-wise complexity (d_k^3) and the cost of updating the objective: dK for TeraLasso (Greenewald et al., 2019), $N\sum_k d_k m_k^2$ for Tlasso (Lyu et al., 2019), and $N\sum_k\sum_{j\neq k} d_j m_j^2$ for SG-PALM. The mode-wise complexity of TeraLasso is dominated by matrix inversion, which is hard to scale for general problem instances. For Tlasso/KGlasso, the mode-wise complexity is the same as that of running a Glasso-type algorithm for each mode, which could*

be improved by applying state-of-the-art optimization techniques developed for vector-variate Gaussian graphical models. For SG-PALM, the mode-wise operations involve only sparse-dense matrix multiplications, which could be improved to $O(d_k \cdot \mathsf{nnz})$*, where* nnz *counts the number of non-zero elements of the sparse matrix (i.e., the estimated* $_k$ *at each iteration). This could greatly reduce the computational cost for extremely sparse* $\mathbf{\Psi}_k$*, e.g., with only* $O(d_k)$ *non-zero elements. Further, Tlasso and SG-PALM both incur a cost of* $O(Nd_k m_k^2)$ *for each mode-wise update. This can also be reduced to be* $\approx d$ *for sparse estimated* $\mathbf{\Psi}_k$*'s at each iteration. Overall, for sample-starved setting where we only have access to a handful of data samples, structured KP and KS models run similarly fast, while the Sylvester GM runs slower theoretically due to the extra and richer structures that it takes into account.*

Additionally, TG-ISTA and the Tlasso proposed both require inversion of $d_k \times d_k$ *matrices, which is not easily parallelizable and cannot easily exploit the sparsity of* $\mathbf{\Psi}_k$*'s. The cyclic coordinate-wise method used in SyGlasso does not allow for parallelization since it requires cycling through entries of each* $\mathbf{\Psi}_k$ *in specified order. In contrast, SG-PALM can be implemented in parallel to distribute the sparse matrix-matrix multiplications because at no step do the algorithms require storing all dense matrices on a single machine. Therefore, with the adaptation of communication-efficient algorithms (such as that proposed in Koanantakool et al. (2018) for vector-variate Gaussian graphical models), the scalability of the distributed SG-PALM is restricted only by the number of machines available.*

3.4 Convergence Analysis

In this section, we present the main convergence theorems. Detailed proofs are included in the supplement. Here, we study the convergence behavior in the convex cases, but similar convergence rate can be established for non-convex penalties (see supplement).

We first establish statistical convergence of a global minimizer $\hat{\mathbf{\Psi}}$ of (3.3) to its true value, denoted as $\bar{\mathbf{\Psi}}$, under the correct statistical model.

Theorem 3.4.1. *Let* $\mathcal{A}_k := \{(i,j) : (\bar{\mathbf{\Psi}}_k)_{i,j} \neq 0, i \neq j\}$ *and* $q_k := |\mathcal{A}_k|$ *for* $k = 1, \ldots, K$. *If* $N > O(\max_k q_k d_k \log d)$ *and* $d := d_N = O(N^\kappa)$ *for some* $\kappa \geq 0$, *and further, if the penalty parameter satisfies* $\lambda_k := \lambda_{N,k} = O(\sqrt{\frac{d_k \log d}{N}})$ *for all* $k = 1, \ldots, K$, *then under conditions (A1-A3) in Appendix 2.3.1, there exists a constant* $C > 0$ *such that for any* $\eta > 0$ *the following events hold with probability at least* $1 - O(\exp(-\eta \log d))$*:*

$$\sum_{k=1}^{K} \|\mathit{offdiag}(\hat{\mathbf{\Psi}}_k) - \mathit{offdiag}(\bar{\mathbf{\Psi}}_k)\|_F \leq C\sqrt{K} \max_k \sqrt{q_k}\lambda_k.$$

Here $\mathit{offdiag}(\mathbf{\Psi}_k)$ *is the the off-diagonal part of* $\mathbf{\Psi}_k$. *If further* $\min_{(i,j)\in\mathcal{A}_k} |(\bar{\mathbf{\Psi}}_k)_{i,j}| \geq 2C \max_k \sqrt{q_k}\lambda_k$ *for each* k, *then* $\mathit{sign}(\hat{\mathbf{\Psi}}_k) = \mathit{sign}(\bar{\mathbf{\Psi}}_k)$.

Theorem 3.4.1 means that under regularity conditions on the true generative model, and with appropriately chosen penalty parameters λ_k's guided by the theorem, one is guaranteed to recover the true structures of the underlying Sylvester generating parameters $\mathbf{\Psi}_k$ for $k = 1, \ldots, K$ with probability one, as the sample size and dimension grow.

We next turn to convergence of the iterates $\{\mathbf{\Psi}^t\}$ from SG-PALM to a global optimum of (3.3).

Theorem 3.4.2. *Let* $\{\mathbf{\Psi}^{(t)}\}_{t\geq 0}$ *be generated by SG-PALM. Then, SG-PALM converges in the sense that*

$$\frac{\mathcal{L}_{\boldsymbol{\lambda}}(\mathbf{\Psi}^{(t+1)}) - \min \mathcal{L}_{\boldsymbol{\lambda}}}{\mathcal{L}_{\boldsymbol{\lambda}}(\mathbf{\Psi}^{(t)}) - \min \mathcal{L}_{\boldsymbol{\lambda}}} \leq \left(\frac{\alpha^2 L_{\min}}{4Kc^2(\sum_{j=1}^{K} L_j)^2 + 4c^2 L_{\max}} + 1 \right)^{-1},$$

where α, $L_k, k = 1, \ldots, K$ *are positive constants,* $L_{\min} = \min_j L_j$, $L_{\max} = \max_j L_j$, *and* $c \in (0, 1)$ *is the backtracking constant defined in Algorithm III.1.*

Note that the term on the right hand side of the inequality above is strictly less than 1. This means that the SG-PALM algorithm converges linearly, which is a strong results for a non-strongly convex objective (i.e., $\mathcal{L}_{\boldsymbol{\lambda}}$). To the best of our knowledge, for first-order optimization methods, this rate is faster than any other Gaussian graphical models having non-strongly convex objectives (see Khare et al. (2015); Oh et al. (2014) and references therein) and comparable with those having strongly-convex objectives (see, for example, Guillot et al. (2012); Dalal and Rajaratnam (2017); Greenewald et al. (2019)). In practical large-scale applications, a fast rate is vital as it would be desired to have the iterative optimization approximation errors quickly converge to values below the statistical errors.

3.5 Experiments

Experiments in this section were performed in a system with `8-core Intel Xeon CPU E5-2687W v2 3.40GHz` equipped with `64GB RAM`. SG-PALM was implemented in `Julia v1.5`. For synthetic data analyses, we used the SyGlasso implementation in `R` with `C++` speed-up (`https://github.com/ywa136/syglasso`). For real data analyses, we used the `Tlasso` package implementation in `R` (Sun et al., 2016) and the TeraLasso implementation in `MATLAB` (`https://github.com/kgreenewald/teralasso`).

3.5.1 Synthetic data

We first validate the convergence theorems discussed in the previous section via simulation studies. Synthetic datasets were generated from true sparse Sylvester factors $\{\boldsymbol{\Psi}_k\}_{k=1}^K$ where $K = 3$ and $d_k = \{16, 32, 64\}$ for all k. Instances of the random matrices used here have uniformly random sparsity patterns with edge densities (i.e.,

the proportion of non-zero entries) ranging from $0.1\% - 30\%$ on average over all $\mathbf{\Psi}_k$'s. For each d and edge density combination, random samples of size $N = \{10, 100, 1000\}$ were tested. For comparison, the initial iterates, convergence criteria were matched between SyGlasso and SG-PALM. Highlights of the results in run times are summarized in Table III.1.

Table III.1: Run time comparisons (in seconds with N/As indicating those exceeding 24 hour) between SyGlasso and SG-PALM on synthetic datasets with different dimensions, sample sizes, and densities of the generating Sylvester factors. Note that the proposed SG-PALM has average speed-up ratios ranging from 1.5 to 10 over SyGlasso.

d	N	NZ%	**SyGlasso**		**SG-PALM**	
			iter	**sec**	**iter**	**sec**
16^3	10^1	0.11	9	4.6	11	4.5
		4.10	9	5.1	32	5.1
	10^2	0.21	8	8.8	11	5.4
		2.60	8	10.8	35	7.2
	10^3	0.26	8	82.4	12	14.3
		3.40	10	99.2	37	33.5
32^3	10^1	0.13	10	191.2	19	7.3
		7.50	17	304.8	42	10.2
	10^2	0.46	9	222.4	24	28.9
		7.00	17	395.2	41	48.5
	10^3	0.10	9	1764.8	22	226.4
		6.90	19	3789.4	41	473.9
64^3	10^1	0.65	10	583.7	42	91.3
		14.5	22	952.2	47	119.0
	10^2	0.62	9	6683.7	41	713.9
		14.4	21	15607.2	48	1450.9
	10^3	0.85		N/A	39	6984.4
		14.0		N/A	48	12968.7

Convergence behavior of SG-PALM is shown in Figure III.1 (a) for the datasets with $d_k = 32$, $N = \{10, 100\}$, and edge densities roughly around 5% and 20%, respectively. Geometric convergence rate of the function value gaps under Theorem 3.4.2 can be verified from the plot. Note an acceleration in the convergence rate (i.e., a steeper slope) near the optimum, which is suggested by the "localness" of the Kur-

dyka - Łojasiewicz (KL) property (defined in Section B.2 of the Appendix) of the objective function close to its global optimum. Further for the same datasets, in Figure III.1 (b), SG-PALM graph recovery performances is illustrated, where the Matthew's Correlation Coefficients (MCC) is plotted against run time. Here, MCC is defined by

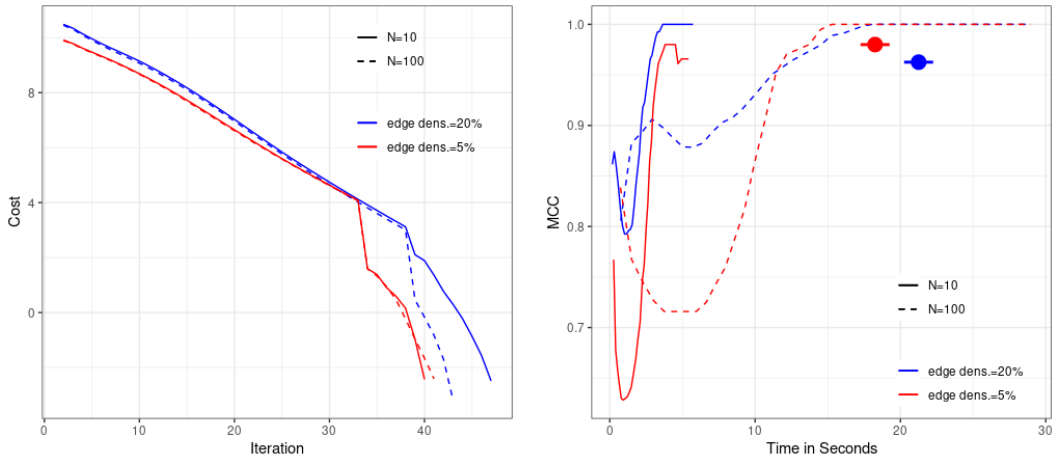
$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}},$$

where TP is the number of true positives, TN the number of true negatives, FP the number of false positives, and FN the number of false negatives of the estimated edges (i.e., non-zero elements of $\mathbf{\Psi}_k$'s). An MCC of 1 represents a perfect prediction, 0 no better than random prediction and -1 indicates total disagreement between prediction and observation. The results validate the statistical accuracy under Theorem 3.4.1. It also shows that SG-PALM outperforms SyGlasso (indicated by blue/red solid dots) within the same time budget.

3.5.2 Solar imaging data

Solar active regions are temporary centers of strong and complex magnetic field on the sun, the principal source of violent eruptions such as solar flares (van Driel-Gesztelyi and Green, 2015). While weak flares of, for example, B-class, have only limited terrestrial effect, strong flares of M- and X-class can produce tremendous amount of electromagnetic radiation, causing disturbance or damage to satellites, power grids, and communication systems. Therefore, it would be great value to be able to predict how active regions evolve before the onset of solar flares.

Although there are numerous studies that use active region images or physical parameters to predict flare activities (Leka and Barnes, 2003; Chen et al., 2019; Jiao et al., 2020b; Wang et al., 2020b; Sun et al., 2021), fewer studies have attempted to



(a) Cost gap vs. Iteration (b) MCC vs. Run time

Figure III.1: Convergence of SG-PALM algorithm under datasets with varying sample sizes (solid and dashed) generated via matrices with different sparsity (red and blue). The function value gaps on log-scale (left) verifies the geometric convergence rate in all cases and the MCC over time (right) demonstrates the algorithm's accuracy and efficiency. Note that the SG-PALM reached almost perfect recoveries (i.e., MCC of 1) within 20 seconds in all cases. In comparison, SyGlasso (big solid dots with line-range) was only able to achieve at lower MCCs for lower sample-size cases within 30 seconds.

predict the complicated preflare evolution of active regions without physical modeling (Bai et al., 2021). Furthermore, existing work tends to focus on predictions using images collected from a single space instrument. In this section, to illustrate the viability of the proposed tensor graphical models, we use multiwavelength active region observations acquired by multiple instruments: the Solar Dynamics Observatory (SDO)/Helioseismic and Magnetic Imager (HMI) and SDO/Atmospheric Imaging Assembly (AIA), to predict the evolution of two types of active regions that lead to either a weak (B-class) flare or a strong (M- or X-class) flare.

We construct a multiwavelength active region video dataset from the curated dataset generated by Galvez et al. (2019). The video data are taken in four wavelengths (94Å, 131Å, 171Å, and 193Å) by the Atmospheric Imaging Assembly (AIA, Lemen et al., 2011) plus the three prime HMI vector magnetic field components Bx, By, and Bz, both aboard the Solar Dynamics Observatory (SDO) satellite. Each video

is a 24-hour image sequence of an active region at 1-hour cadence before a strong (M- or X-class) or a weak (B-class) flare occurs in the region. We spatially interpolate the videos so that each video is represented as a $d_1 \times d_2 \times d_3 \times d_4$ tensor, where $d_1 = 13$ denotes the number of frames in the video, $d_2 = 50$ denotes the height of the frames after interpolation, $d_3 = 100$ denotes the width of the frames after interpolation, and $d_4 = 7$ represents the number of different channels/wavelength/components at which the images are recorded. To prevent information leakage, we chronologically split the active region videos into a training set (year 2011 to 2014) and a test set (year 2011 to 2014). In the training set, there are 186 active region videos that lead to a B-class flare and 48 active region videos that lead to a M/X-class flare. In the test set, the sample sizes are 93 and 24 for the B-class and the M/X-class, respectively.

To perform active region prediction, we first fit the tensor graphical models on the training set to estimate the covariance or prediction matrices for each of the two types of active region videos, and then we use the best linear predictor to predict the last frame from all previous frames for videos in the test set. The forward linear predictor is constructed in a multi-output least squares regression setting as

$$\hat{\mathbf{y}}_t = -\mathbf{\Omega}_{2,2}^{-1}\mathbf{\Omega}_{2,1}\mathbf{y}_{t-1:t-(p-1)} \tag{3.10}$$

when the precision estimate is available. Here, $t = d_1$ for predicting the last frame of a video. For notational convenience, let $p = d_1$ and $q = d_2 d_3 d_4$, then $\mathbf{y}_{t-1:t-(p-1)} = \mathbf{y}_{p-1:1} \in \mathbb{R}^{(p-1)q}$ is the stacked set of pixel values from the previous $p-1$ time instances and $\mathbf{\Omega}_{2,1} \in \mathbb{R}^{q\times(p-1)q}$ and $\mathbf{\Omega}_{2,2} \in \mathbb{R}^{q\times q}$ are submatrices of the $pq \times pq$ estimated precision matrix:

$$\hat{\mathbf{\Omega}} = \begin{pmatrix} \mathbf{\Omega}_{1,1} & \mathbf{\Omega}_{1,2} \\ \mathbf{\Omega}_{2,1} & \mathbf{\Omega}_{2,2} \end{pmatrix}.$$

The predictors were tested on the data containing flares observed from different

active regions than those in training set, so that the predictor has never "seen" the frames that it attempts to predict, corresponding to 117 observations of which 93 are B-class flares and 24 are MX-class flares. Figure III.2 shows the root mean squared error normalized by the difference between maximum and minimum pixels (NRMSE) over the testing samples, for the forcasts based on the SG-PALM estimator, TeraLasso estimator (Greenewald et al., 2019), Tlasso estimator (Lyu et al., 2019), and IndLasso estimator. Here, the TeraLasso and the Tlasso are estimation algorithms for a KS and a KP tensor precision matrix model, respectively; the IndLasso denotes an estimator obtained by applying independent and separate ℓ_1-penalized regressions to each pixel in $\boldsymbol{y}_t$. The SG-PALM estimator was implemented using a regularization parameter $\lambda_N = C_1\sqrt{\frac{\min(d_k)\log(d)}{N}}$ for all k with the constant C_1 chosen by optimizing the prediction NRMSE on the training set over a range of λ values parameterized by C_1. The TeraLasso estimator and the Tlasso estimator were implemented using $\lambda_{N,k} = C_2\sqrt{\frac{\log(d)}{N\prod_{i\neq k} d_i}}$ and $\lambda_{N,k} = C_3\sqrt{\frac{\log(d_k)}{Nd}}$ for $k = 1, 2, 3$, respectively, with C_2, C_3 optimized in a similar manner. Each sparse regression in the IndLasso estimator was implemented and tuned independently with regularization parameters chosen from a grid via cross-validation.

We observe that SG-PALM outperforms all three other methods, indicated by NRMSEs across pixels. Figure III.3 depicts examples of predicted images, comparing with the ground truth. The SG-PALM estimates produced most realistic image predictions that capture the spatially varying structures and closely approximate the pixel values (i.e., maintaining contrast ratios). The latter is important as the flares are being classified into weak (B-class) and strong (MX-class) categories based on the brightness of the images, and stronger flares are more likely to lead to catastrophic events, such as those damaging spacecrafts. Lastly, we compare run times of the SG-PALM algorithm for estimating the precision matrix from the solar flare data with SyGlasso. Table B.1 in Appendix 2.5 illustrates that the SG-PALM algorithm

converges faster in wallclock time. Note that in this real dataset, which is potentially non-Gaussian, the convergence behavior of the algorithms is different compare to synthetic examples. Nonetheless, SG-PALM enjoys an order of magnitude speed-up over SyGlasso.

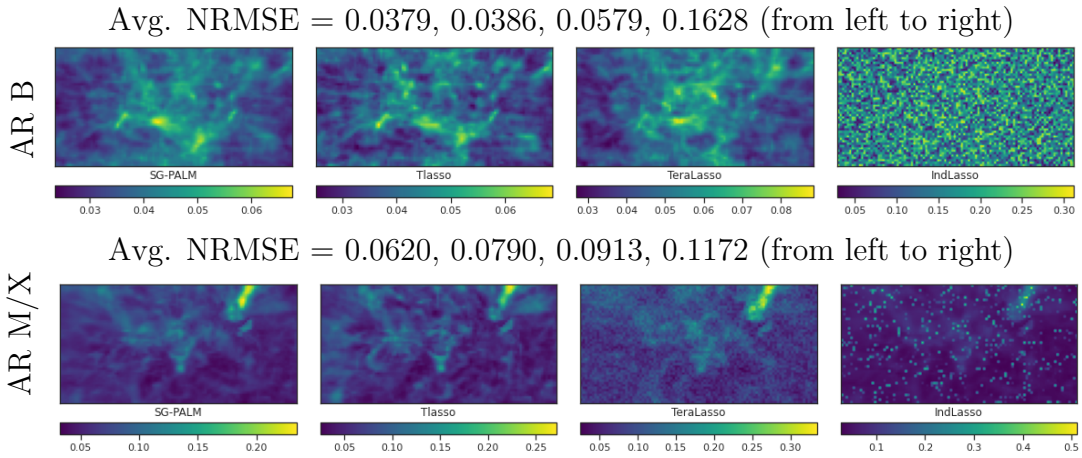


Figure III.2: Comparison of the SG-PALM, Tlasso, TeraLasso, IndLasso performances measured by NRMSE in predicting the last frame of 13-frame video sequences leading to B- and MX-class solar flares. The NRMSEs are computed by averaging across testing samples and AIA channels for each pixel. 2D images of NRMSEs are shown to indicate that certain areas on the images (usually associated with the most abrupt changes of the magnetic field/solar atmosphere) are harder to predict than the rest. SG-PALM achieves the best overall NRMSEs across pixels. B flares are generally easier to predict due to both a larger number of samples in the training set and smoother transitions from frame to frame within a video (see the supplemental material for details).

3.6 Conclusion

We proposed SG-PALM, a proximal alternating linearized minimization method for solving a pseudo-likelihood based sparse tensor-variate Gaussian precision matrix estimation problem. Geometric rate of convergence of the proposed algorithm is established building upon recent advances in the theory of PALM-type algorithms. We demonstrated that SG-PALM outperforms the coordinate-wise minimization method in general, and in ultra-high dimensional settings SG-PALM can be faster by at least

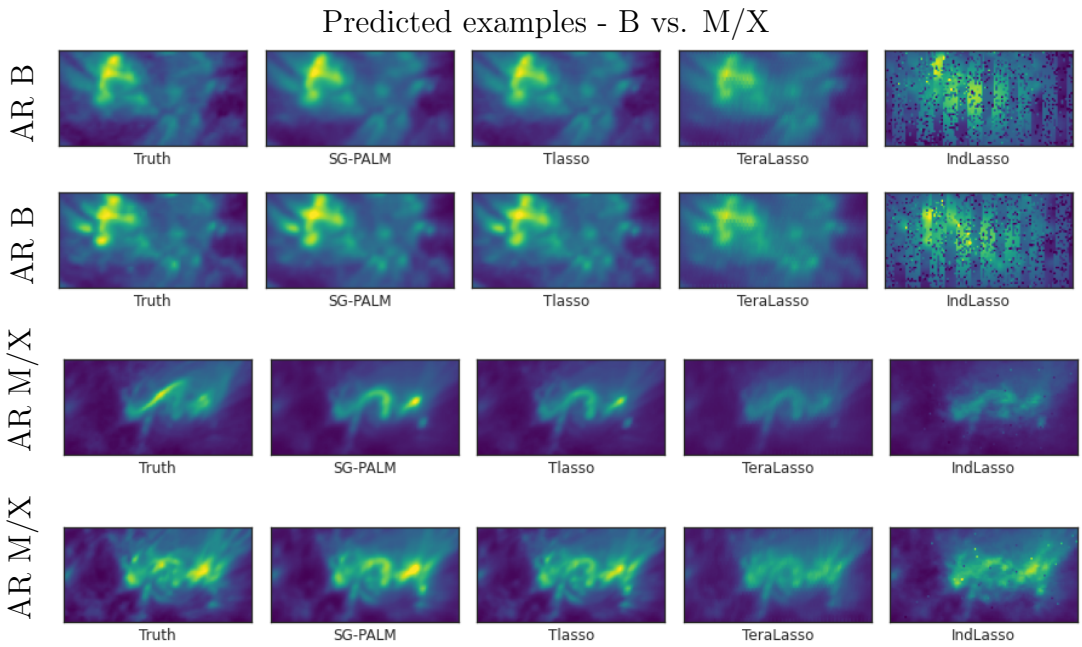


Figure III.3: Examples of one-hour ahead prediction of the first two AIA channels of last frames of 13-frame videos, leading to B- (first two rows) and MX-class (last two rows) flares, produced by the SG-PALM, Tlasso, TeraLasso, IndLasso algorithms, comparing to the real image (far left column). Note that in general linear forward predictors tend to underestimate the contrast ratio of the images. The proposed SG-PALM produced the best-quality images in terms of both the spatial structures and contrast ratios. See the supplemental material for examples of predicted images from the HMI instrument.

an order of magnitude. A link between the Sylvester generating equation underlying the graphical model and certain physical processes was established. This connection was illustrated on a novel astrophysics application, where multi-instrument imaging datasets characterizing solar flare events were used. The proposed methodology was able to robustly forward predict both the patterns and intensities of the solar atmosphere, yielding potential insights to the underlying physical processes that govern the flaring events.

Future directions include additional downstream tasks involving solar flare predictions using the estimated precision matrix, such as classification of strong/weak flares. Furthermore, the statistical convergence rate outlined in Theorem 3.4.1 might not be optimal. We have observed that, for example, from the simulation study in

Section 3.5, where we see in Figure III.1(b) that the estimator achieves perfect graph recovery accuracy even when $N = 10$, which is better than the sample complexity implied by the theorem. We are actively working towards obtaining a tighter upper bound on the statistical error.

CHAPTER IV

Multiway Ensemble Kalman Filter

In this chapter, we develop methods of forecasting multiway times-series generated by dynamical systems. These methods can be used to study the emergence of sparsity and multiway structures in second-order statistical characterizations of dynamical processes governed by partial differential equations (PDEs). We consider several state-of-the-art multiway covariance and inverse covariance (precision) matrix estimators and examine their pros and cons in terms of accuracy and interpretability in the context of physics-driven forecasting when incorporated into the ensemble Kalman filter (EnKF). In particular, we show that multiway data generated from the Poisson, the convection-diffusion, and the Kuramoto–Sivashinsky types of PDEs can be accurately tracked via EnKF when integrated with appropriate covariance and precision matrix estimators.

4.1 Introduction

There has recently been a resurgence of interest in integrating machine learning with physics-based modeling. Much of the recent work has focused on black-box models such as deep neural networks (Takeishi et al., 2017; Long et al., 2018; Zhang et al., 2018; Vlachas et al., 2018; Reichstein et al., 2019; Wang et al., 2020a). However, seeking shallower models that capture mechanism in a physically interpretable man-

ner has been a recurring theme in both machine learning and physics (Weinan et al., 2020). The Kalman filter is a well-known technique to track a linear dynamical system over time by assimilating real-world observations into physical knowledge. Many variants based on the extended and ensemble Kalman filters have been proposed to deal with non-linear systems. However, these systems are often high dimensional and forecasting each ensemble member forward through the system is computationally expensive. Moreover, in the high dimensional and low sample regime ($N \ll d$), the sample covariance matrix of the forecast ensemble is extremely noisy. Previous methods for dealing with these sampling errors can be dived into the "stochastic filters" and the "deterministic filters". The former often involve manually "tuning" of the sample covariance with variance inflation and localization (Hamill et al., 2001; Houtekamer and Mitchell, 2001; Ott et al., 2004; Wang et al., 2007; Anderson, 2007, 2009; Li et al., 2009; Bishop and Hodyss, 2009a,b; Campbell et al., 2010; Greybush et al., 2011; Miyoshi, 2011). However, these schemes require carefully choosing the inflation factor and using expert knowledge to determine local areas of interest that are used in assimilation. Additionally, they work with perturbed observations that introduce further sampling errors due to the lack of orthogonality between the perturbation noise and the ensembles. This has led to the development of deterministic versions of the EnKF such as the square root and transform filters (Bishop et al., 2001; Evensen, 2004; Whitaker and Hamill, 2002; Tippett et al., 2003; Hunt et al., 2007; Godinez and Moulton, 2012; Nerger et al., 2012; Tödter and Ahrens, 2015), which do not perturb the observations and are designed to avoid these additional sampling errors. Lawson and Hansen (2004) studies the differences between different approaches (stochastic vs. deterministic) of EnKF and the implications of those differences in various regimes, and claims that the stochastic filters can better withstand regimes with nonlinear error growth.

Most similar to our proposed work is Hou et al. (2021), which suggests to imple-

ment EnKF with a sparse inverse covariance estimator to handle the high-dimensional regime. However, we note that many real-world processes are complex and generating heterogeneous multiway/tensor-variate data. For example, weather satellites measure spatio-temporal climate variables such as temperature, wind velocity, sea level, pressure, etc. Due to the non-homogeneous nature of these data, estimation of the second-order information that encodes (conditional) dependency structure within the data is of great importance. Assuming the data are drawn from a tensor normal distribution, a straightforward way to estimate this structure is to vectorize the tensor and estimate the underlying Gaussian graphical model associated with the vector, as suggested by Hou et al. (2021). Such an approach ignores the tensor structure and requires estimating a rather high dimensional precision matrix, often with insufficient sample size. In many scientific applications the sample size can be as small as one when only a single tensor-valued measurement is available. In this chapter, we introduce a high-dimensional statistical approach that naturally integrates physics and machine learning through Kronecker-structured Gaussian graphical models. The learned representation can then be incorporated into a high dimensional predictive model using the ensemble Kalman filtering framework.

4.2 Background

We consider a noisy, non-linear dynamical model $f(\cdot)$ that evolves some unobserved states $\mathbf{x}_t \in \mathbb{R}^d$ through time. A noisy version of the states, $\mathbf{x}_t \in \mathbb{R}^r_t$, is observed via a transformation of $\mathbf{x}_t$ by a function $h(\cdot)$. Both the state/process noise $\mathbf{v}_t$ and the observation noise $\mathbf{w}_t$ are assumed to be independent of the states. Further, we assume both noises are zero-mean Gaussians with known diagonal covariance

matrices $\mathbf{Q}_t$ and $\mathbf{R}_t$. Specifically,

$$\begin{aligned} \mathbf{x}_t &= f(\mathbf{x}_{t-1}) + \mathbf{v}_t, \\ \mathbf{y}_t &= h(\mathbf{x}_t) + \mathbf{w}_t. \end{aligned} \tag{4.1}$$

In this work, we further restrict both noise variables $\mathbf{v}$ & $\mathbf{w}$ and the observational process are time-invariant, i.e., $\mathbf{v}_t = \mathbf{v}$, $\mathbf{w}_t = \mathbf{w}$, and $r_t = r$, although the methods developed here work in time-variant scenarios. In geophysical problems such as weather prediction, the state and observation dimensions are often enormous (i.e., $d \geq 10^7$ and $r_t \geq 10^5$). Therefore, as with localization methods, we make an assumption about the correlation structure of the state vector in order to handle the high dimensionality of the state. Specifically, only a small number of pairs of state variables are assumed to have non-zero conditional correlation, i.e., $\text{cov}(x_i, x_j | \mathbf{x}_{-(i,j)}) \neq 0$ where $\mathbf{x}_{-(i,j)}$ represents all state variables except x_i and x_j. For an illustrating example, consider a one-dimensional spatial field with three locations x_1 , x_2, and x_3 where x_1 and x_3 are both connected to x_2, but not each other. In this case, it is natural to model x_1 and x_3 as uncorrelated conditioned on x_2 although they are not necessarily marginally uncorrelated, that is, $\text{cov}(x_1, x_3 | x_2) = 0$ but $\text{cov}(x_1, x_3) \neq 0$. Similar conditional independence assumptions have been used in the study of Markov random fields (MRFs), which find applications in, for example, image processing to generate textures as they can be used to generate flexible and stochastic image models (Kindermann, 1980). A spacial case is the Gaussian MRFs, which is most widely used in spatial statistics (Rue and Held, 2005). For Gaussian states, the assumption that the set of non-zero conditional correlations is sparse is equivalent to the assumption that the inverse correlation matrix of the model state is sparse with few non- zero off-diagonal entries (Lauritzen, 1996).

4.2.1 Ensemble Kalman filter

The ensemble Kalman filter (EnKF) is particularly effective when the dynamical system is complicated and non-linear, which is often the case in physical systems (Evensen, 1994; Burgers et al., 1998). In these cases, analytic propagation of the entire Gaussian systems as in the classic Kalman filter (KF) algorithm fails (Evensen, 2003). The EnKF can be viewed as an approximate version of the KF, in which the state distribution is represented by a sample or "ensemble" from the distribution. This ensemble is then propagated forward through time and updated when new data become available.

The forecast covariance matrix is replaced by its sample estimate obtained from the forecast ensemble. However, such systems are often high-dimensional and the EnKF operates in the regime where the number of ensemble members, N, is much less than the size of the state, d, suggesting that the sample covariance matrix is singular and may introduce spurious correlations (Greybush et al., 2011). In this case, regularized inverse covariance models will be especially attractive. Hou et al. (2021) introduced a sparsity-penalized EnKF, which replace the sample covariance with an estimator of the forecasting covariance whose inverse is sparsity regularized. Here we propose incorporating the multiway covariance / inverse covariance models into the penalized EnKF framework of Hou et al. (2021).

4.2.2 Multiway representations for diffusion processes

Multiway representations are particularly useful when modeling data generated from physical processes as many of these processes obey partial differential equations of the form

$$\begin{aligned} \mathcal{D}u &= f \quad \text{in } \Omega, \\ u &= g \quad \text{on } \partial\Omega, \end{aligned} \tag{4.2}$$

where u is the unknown physical process, f is the driving process (e.g., white Gaussian noise), g is the function value of u on the boundary, $\mathcal{D}$ is some differential operator (e.g, a Laplacian or an Euler-Lagrange operator), and Ω is the domain. After finite difference discretization over the domain Ω, the model is equivalent to (ignoring discretization error) the matrix equation

$$\mathbf{Du} = \mathbf{f}.$$

Here, $\mathbf{D}$ is a sparse matrix since $\mathcal{D}$ is a differential operator. Additionally, as shown below, $\mathbf{D}$ admits the Kronecker structure as a mixture of Kronecker sums and Kronecker products.

The matrix $\mathbf{D}$ reduces to a Kronecker sum when $\mathcal{D}$ involves no mixed derivatives. As an example, we consider the Poisson equation, an elliptical PDE that governs many physical processes including electromagnetic induction, heat transfer, and convection. On a rectangular region $\Omega = (0, d_1) \times (0, d_2)$ in the 2D Cartesian plane, the Poisson equation with homogeneous Dirichlet boundary condition is expressed as

$$\begin{aligned} \mathcal{D}u = (\partial_x^2 + \partial_y^2)u = f \quad &\text{in } \Omega, \\ u = 0 \quad &\text{on } \partial\Omega \end{aligned} \tag{4.3}$$

where $f : \Omega \to \mathbb{R}$ is the given source function and $u : \Omega \to \mathbb{R}$ is the unknown process of interest. Using the finite difference method with a square mesh grid with unit spacing, the unknown and the source can be expressed as d_1-by-d_2 matrices, $\mathbf{U}$ and $\mathbf{F}$, respectively, that are related to each other via

$$U_{i+1,j} + U_{i-1,j} + U_{i,j+1} + U_{i,j-1} - 4U_{i,j} = F_{i,j} \tag{4.4}$$

for any interior grid point (i, j). Defining n-by-n square matrix

$$\mathbf{A}_n = \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix},$$

the above relation can be expressed as the (vectorized) Sylvester equation with $K = 2$:

$$(\mathbf{A}_{d_1} \oplus \mathbf{A}_{d_2})\mathbf{u} = \mathbf{f}, \tag{4.5}$$

where $\mathbf{u} = \text{vec}(\mathbf{U})$, $\mathbf{f} = \text{vec}(\mathbf{F})$. Note that $\mathbf{A}$ is tridiagonal. In the case where $\mathbf{f}$ is white noise with variance σ^2, the inverse covariance matrix of $\mathbf{u}$ has the form $\text{cov}^{-1}(\mathbf{u}) = \sigma^{-2}(\mathbf{A}_{d_1} \oplus \mathbf{A}_{d_2})^T(\mathbf{A}_{d_1} \oplus \mathbf{A}_{d_2})$ and hence sparse.

More generally, any physical process generated from Equation (4.2) also has sparse inverse covariance matrices due to the sparsity of general discretized differential operators. Note that similar connections between continuous state physical processes and sparse "discretized" statistical models have been established by Lindgren et al. (2011), who elucidated a link between Gaussian fields and Gauss Markov Random Fields via stochastic partial differential equations.

4.2.3 Kronecker-structured covariance models

Classic regularized estimators such as the graphical lasso (Glasso, Friedman et al., 2008) for the (inverse) covariance induced by Equation (4.5) may fail because: 1) both d_1 and d_2 may be large (and as a result, $d = d_1 d_2$ is large) for large spatial fields/domains; 2) ignoring the Kronecker structure may lead to (statistical) inefficiency of the method, i.e., the estimator not converging to the estimand; 3) ignoring the generative process in Equation (4.5) will result in learned structures that are not

easily (physically) interpretable.

To address these issues in learning second-order representations for multiway (tensor) data, (sparse) Kronecker product (KP) or Kronecker sum (KS) decomposition of $\mathbf{\Sigma}$ or $\mathbf{\Omega}$ are often employed. Statistical models and corresponding learning algorithms can be derived using generative models or matrix approximations. The former include: KGlasso/Tlasso (Tsiligkaridis et al., 2013; Lyu et al., 2019) for estimating $\mathbf{\Omega} = \mathbf{A} \otimes \mathbf{B}$, using a autoregressive representation $\mathbf{AXB} = \mathbf{Z}$ for data $\mathbf{X}$ when $\mathbf{Z}$ is white noise. Another generative model is SyGlasso/SG-PALM (Chapter II,III) that models the precision matrix as $\mathbf{\Omega} = (\mathbf{A} \oplus \mathbf{B})^2$, which corresponds to assuming the data $\mathbf{X}$ obeys a Sylvester equation $\mathbf{XA} + \mathbf{BX} = \mathbf{Z}$. Matrix approximation methods include: KPCA (Tsiligkaridis and Hero, 2013; Greenewald and Hero, 2015) that approximates the covariance matrix as $\mathbf{\Sigma} = \sum_{i=1}^{l} \mathbf{A}_i \otimes \mathbf{B}_i$, i.e., low separation rank l. Another matrix approximation method is the TeraLasso (Greenewald et al., 2019) that models the precision matrix as $\mathbf{\Omega} = \mathbf{A} \oplus \mathbf{B}$. TeraLasso is equivalent to approximation of the conditional dependency graph (encoded by the precision matrix) with a Cartesian product of smaller graphs [1].

All of KGlasso/Tlasso, TeraLasso, and SyGlasso/SG-PALM can be formulated using a penalized Gaussian likelihood approach. Here, we give a brief review of penalized Gaussian graphical models for multiway, tensor-valued data. A random tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ follows the tensor normal distribution with zero mean when $\text{vec}(\boldsymbol{\mathcal{X}})$ follows a normal distribution with mean $\mathbf{0} \in \mathbb{R}^d$ and precision matrix $\mathbf{\Omega} := \mathbf{\Omega}(\mathbf{\Psi}_1, \ldots, \mathbf{\Psi}_K)$, where $d = \prod_{k=1}^{K} d_k$. Here, $\mathbf{\Omega}(\mathbf{\Psi}_1, \ldots, \mathbf{\Psi}_K)$ is parameterized by $\mathbf{\Psi}_k \in \mathbb{R}^{d_k \times d_k}$ via either Kronecker product, Kronecker sum, or the Sylvester structure, and the corresponding negative log-likelihood function (assuming N independent observations $\boldsymbol{\mathcal{X}}^i, i = 1, \ldots, N$)

$$-\frac{N}{2} \log |\mathbf{\Omega}| + \frac{N}{2} \text{tr}(\mathbf{S\Omega}), \tag{4.6}$$

[1] Note that Tlasso, TeraLasso, Syglasso/SG-PALM are generalizable to precision matrices of the form $\bigotimes_{k=1}^{K} \mathbf{\Psi}_k$, $\bigoplus_{k=1}^{K} \mathbf{\Psi}_k$, and $(\bigoplus_{k=1}^{K} \mathbf{\Psi}_k)^2$, respectively, for $K \geq 2$.

where $\boldsymbol{\Omega} = \bigotimes_{k=1}^{K} \boldsymbol{\Psi}_k$, $\bigoplus_{k=1}^{K} \boldsymbol{\Psi}_k$, or $\left(\bigoplus_{k=1}^{K} \boldsymbol{\Psi}_k\right)^2$ for KP, KS, and Sylvester models, respectively; and $\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N} \operatorname{vec}(\boldsymbol{\mathcal{X}}^i)\operatorname{vec}(\boldsymbol{\mathcal{X}}^i)^T$. For $K = 1$, this formulation reduces to the vector normal distribution with zero mean and precision matrix $\boldsymbol{\Psi}_1$.

To encourage sparsity in the high-dimensional scenario, penalized negative log-likelihood function is proposed

$$-\frac{N}{2}\log|\boldsymbol{\Omega}| + \frac{N}{2}\operatorname{tr}(\mathbf{S}\boldsymbol{\Omega}) + \sum_{k=1}^{K} P_{\lambda_k}(\boldsymbol{\Psi}_k),$$

where $P_{\lambda_k}(\cdot)$ is a penalty function indexed by the tuning parameter λ_k and is applied elementwise to the off-diagonal elements of $\boldsymbol{\Psi}_k$. Popular choices for $P_{\lambda_k}(\cdot)$ include the lasso penalty (Tibshirani, 1996), the adaptive lasso penalty (Zou, 2006), the SCAD penalty (Fan and Li, 2001), and the MCP penalty (Zhang et al., 2010).

To further reduce computational complexity and improve robustness, the Sylvester models in SyGlassoi/SG-PALM consider the penalized negative log-pseudolikelihood

$$\begin{aligned}\mathcal{L}_\lambda(\boldsymbol{\Psi}) = &-\frac{N}{2}\log\Big|\Big(\bigoplus_{k=1}^{K}\operatorname{diag}(\boldsymbol{\Psi}_k)\Big)^2\Big| \\ &+ \frac{N}{2}\operatorname{tr}(\mathbf{S}\boldsymbol{\Omega}) + \sum_{k=1}^{K} P_{\lambda_k}(\boldsymbol{\Psi}_k).\end{aligned} \tag{4.7}$$

This differs from the true penalized Gaussian negative log-likelihood in the exclusion of off-diagonals of $\boldsymbol{\Psi}_k$'s in the log-determinant term. It is motivated and derived directly using the Sylvester equation, from the perspective of solving a sparse linear system (see Chapters III and III for details). This maximum pseudolikelihood estimation procedure has been applied to vector-variate Gaussian graphical models (see Khare et al. (2015) and references therein).

Lastly, the matrix approximation approach of Tsiligkaridis and Hero (2013) to multiway covariance estimation is based on the representation $\boldsymbol{\Sigma} = \sum_{i=1}^{l} \mathbf{A}_i \otimes \mathbf{B}_i$. The representation is universal: any square matrix can be represented as a sum of

l Kronecker products for sufficiently large $l \leq \min\{d_1^2, d_2^2\}$, as shown in Van Loan and Pitsianis (1993). The Kronecker components can be obtained via a penalized optimization approach for estimating a rank l Kronecker product decomposition of the sample covariance $\mathbf{S}$, i.e.,

$$\min_{\{\mathbf{A}_i, \mathbf{B}_i\}} \left\| \mathbf{S} - \sum_{i=1}^{l} \mathbf{A}_i \otimes \mathbf{B}_i \right\|_F^2 + \lambda \left\| \sum_{i=1}^{l} \mathbf{A}_i \otimes \mathbf{B}_i \right\|_*$$

for a user-supplied regularization parameter $\lambda > 0$. The solution to this penalized optimization is specified by the first l principal components of the singular value decomposition (SVD) of $\mathcal{R}(\mathbf{S})$ where l is determined by λ through a soft-thresholding of the SVD spectrum. In analogy to the ordinary PCA algorithm, the soft-thresholding SVD solution to this optimization problem was called Kronecker PCA (KPCA) in Greenewald and Hero (2014).

4.3 Penalized Multiway Ensemble Kalman Filter

The proposed multiway ensemble Kalman filter, whose pseudo code is shown in Algorithm IV.1, modifies the EnKF by using a forecast (inverse) covariance estimator $\widehat{\mathbf{\Sigma}}_t^f = (\widehat{\mathbf{\Omega}}_t^f)^{-1}$ obtained from one of the Kronecker-structured methods. From this, the correspondingly modified Kalman gain matrix is given by

$$\widehat{\mathbf{K}}_t = \widehat{\mathbf{\Sigma}}_t^f \mathbf{H}^T (\mathbf{H} \widehat{\mathbf{\Sigma}}_t^f \mathbf{H}^T + \mathbf{R})^{-1} = ((\widehat{\mathbf{\Omega}}_t^f)^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{R}^{-1}. \tag{4.8}$$

As the observations are assimilated into the EnKF through the ensemble update, which depends linearly on the Kalman gain matrix (as outlined in Algorithm IV.1), an accurate estimate of the true $\mathbf{K}_t$ ensures that data is properly incorporated into the forecast ensemble. Hou et al. (Theorem 1, 2021) argues that the estimator $\widehat{\mathbf{K}}_t$ via a Glasso covariance estimator is asymptotically consistent with the true Kalman gain

matrix, given conditions on the state ensembles and the regularization parameters in the covariance estimation procedure. This nice property is a result of convergence of the Glasso-type sparse covariance estimator. Here, the Kronecker-structured estimators outlined in the previous section all enjoy faster rates of convergence, under appropriate regularity conditions. In Table IV.1, we summarize the theoretical guarantees on the (inverse) covariance estimators, and hence the estimator $\widehat{\mathbf{K}}_t$ under different modeling assumptions, i.e., KP, KS, Sylvester. All estimators have a similar $\sqrt{\log d/N}$ factor. However, comparing to Glasso that has an additional $\sqrt{d+s}$ factor, the Kronecker-structured estimators have additional factors that depend on the smaller d_k's (assuming the number of tensor modes remain constant), that is, $\sqrt{l\sum_k d_k^2}$ for matrix approximation based estimators [2], $\sqrt{\sum_k d_k}$ for KP based inverse covariance estimators, $\sqrt{(d+s)/\min_k m_k}$ for KS based inverse covariance estimators, and $\max_k \sqrt{s_k d_k}$ for the Sylvester estimators. These indicate improved theoretical accuracy on estimating the state (inverse) covariance and the Kalman gain matrix.

Algorithm IV.1: Multiway Ensemble Kalman Filter

Input: Initial ensemble $\widehat{\mathbf{x}}_0^{(1)}, \ldots, \widehat{\mathbf{x}}_0^{(N)}$, observations at each time $\mathbf{y}_t$, measurement operator $\mathbf{H}$, state and observation noise covariance matrices $\mathbf{Q}$ and $\mathbf{R}$

for $t = 1, \ldots, T$ **do**

for $i = 1, \ldots, N$ **do**

Forecast Step: Evolve each ensemble member forward in time via $\widetilde{\mathbf{x}}_t^{(i)} = f(\widehat{\mathbf{x}}_{t-1}^{(i)}) + \mathbf{w}^{(i)}$, with $\mathbf{w}^{(i)} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{Q})$

Multiway Covariance Estimation: Estimate the (inverse) covariance via and compute the Kalman gain matrix $\widehat{\mathbf{K}}_t$ via (4.8)

Update Step: Update the ensemble with the observations by computing $\widehat{\mathbf{x}}_t^{(i)} = \widetilde{\mathbf{x}}_t^{(i)} + \widehat{\mathbf{K}}_t(\mathbf{y}_t + \mathbf{v}_t^{(i)} - \mathbf{H}\tilde{\mathbf{x}}_t^{(i)})$, where $\mathbf{v}_t^{(i)} \sim \mathcal{N}_r(\mathbf{0}, \mathbf{R})$

end for

end for

Output: Final ensemble $\widehat{\mathbf{x}}_T^{(1)}, \ldots, \widehat{\mathbf{x}}_T^{(N)}$.

In terms of computational complexity, all the structured precision estimation algorithms are variants of Glasso, implemented with techniques tailored to the model

[2] Here, l indicates the separation rank.

assumptions for speedup. Generally speaking, the resulting complexity consists of the mode-wise complexity (d_k^3) and the cost of updating the objective: dK for TeraLasso, Nd for KGlasso, and $N\sum_k \sum_{j\neq k} d_j m_j^2$ for SG-PALM. The mode-wise complexity of TeraLasso is dominated by matrix inversion, which is hard to scale for general problem instances. For KGlasso, the mode-wise complexity is the same as that of running a Glasso-type algorithm for each mode, which could be improved by applying state-of-the-art optimization techniques developed for vector-variate Gaussian graphical models such as Hsieh et al. (2013). For SG-PALM, the mode-wise operations involve only sparse-dense matrix multiplications, which could be improved to $O(d_k \cdot \mathsf{nnz})$, where nnz counts the number of non-zero elements of the sparse matrix (i.e., the estimated $\mathbf{\Psi}_k$ at each iteration). This could greatly reduce the computational cost for extremely sparse $\mathbf{\Psi}_k$, e.g., with only $O(d_k)$ non-zero elements. Further, KGlasso and SG-PALM both incur a cost of the type $O(Nd_k m_k^2)$ for each mode-wise update. This can also be reduced to be $\approx d$ for sparse estimated $\mathbf{\Psi}_k$'s at each iteration. Overall, for sample-starved setting where we only have access to a handful of data samples, structured KP and KS models run similarly fast, while the Sylvester GM runs slower theoretically due to the extra and richer structures that it takes into account. The matrix approximation based estimation procedure, KPCA, is in general computationally more expensive than other KP, KS, and Sylvester based methods. This is mostly due to the absence of the sparsity structure in the latter model, as well as as SVD step involved during the its estimation algorithm. There exist faster randomized methods for truncated SVD (Halko et al., 2011). Thus, it still scales well for moderately high-dimensional applications. In Table IV.2 wall-clock runtimes of EnKF integrated with theses (inverse) covariance estimation algorithms are compared under various settings. The table confirms the aforementioned theoretical computational complexities.

Table IV.1: Comparison of theoretical guarantees on sample complexity (statistical error) and computational complexity of various precision / covariance estimators. Here, $M = \max\{d_1, d_2, N\}$, $m_k = \prod_{i \neq k} d_i$ is the co-dimension of the k-th mode, $d = \prod_{k=1}^{K} d_i$, and s_k characterizes the sparsity of each of the inverse covariance Kronecker factors $s_k = |\{(i,j) : i \neq j, [\mathbf{\Psi}_k]_{i,j} \neq 0\}|$, s is the sparsity of the full inverse covariance $s = |\{(i,j) : i \neq j, \mathbf{\Omega}_{i,j} \neq 0\}|$ and $s = \sum_{k=1}^{K} m_k s_k$ if $\mathbf{\Omega}$ satisfies the Kronecker sum model.

Model	Algorithm	Statistical Error	Computational Complexity
Sparse-Precision	Glasso (Friedman et al., 2008)	$O_P\Big(\sqrt{\frac{(d+s)\log d}{N}}\Big)$	$O(d^3)$
KP-Covariance	Robust KPCA (Greenewald and Hero, 2015)	$O_P\Big(\sqrt{\frac{l(d_1^2+d_2^2+\log M)}{N}}\Big)$	$O(ld^2)$
KP-Precision	KGlasso (Tsiligkaridis et al., 2013)	$O_P\Big(\sqrt{\frac{(d_1+d_2)\log M}{N}}\Big)$	$O(d_1^3 + d_2^3 + Nd)$
KS-Precision	TeraLasso (Greenewald et al., 2019)	$O_P\Big(\sqrt{K+1} \cdot \sqrt{\frac{(d+s)\log d}{N \min_k m_k}}\Big)$	$O(dK + \sum_{k=1}^{K} d_k^3)$
Sylvester GM	SG-PALM (Wang and Hero, 2021b)	$O_P\Big(\sqrt{K} \cdot \max_k \sqrt{\frac{s_k d_k \log d}{N}}\Big)$	$O\Big(\sum_{k=1}^{K}(d_k^3 + N\sum_{j \neq k} d_j m_j^2)\Big)$

Table IV.2: Runtime (in seconds) of 20 time steps of EnKF tracking using various (inverse) covariance estimation algorithms. Comparisons under various problem sizes (i.e., different d and N) and two observation types (i.e., fully observed or partially observed) are shown. Note the sparse multiway precision models (SG-PALM, KGlasso, TeraLasso) are comparably fast and are all faster than Glasso (for large problems) and KPCA.

d	N	obs.	**Glasso**	**SG-PALM**	**TeraLasso**	**KGlasso**	**KronPCA**
			sec	**sec**	**sec**	**sec**	**sec**
32^2	25	full	14.52(0.20)	18.03(0.15)	17.06(0.35)	18.60(0.11)	80.88(0.20)
		part	10.95(0.01)	13.94(0.05)	11.02(0.03)	11.88(0.10)	53.89(0.88)
	50	full	24.88(0.21)	31.02(0.08)	27.53(0.50)	33.25(0.10)	92.76(0.55)
		part	18.95(0.09)	25.37(0.53)	18.63(0.02)	28.63(0.11)	66.12(0.56)
	100	full	48.23(0.05)	58.63(0.13)	49.62(0.28)	67.53(0.30)	131.43(1.05)
		part	36.41(1.05)	41.80(0.25)	36.80(0.55)	53.77(0.14)	78.74(1.13)
64^2	25	full	288.13(2.45)	207.71(1.09)	195.92(0.58)	217.71(1.99)	2562.19(2.69)
		part	281.96(2.04)	191.56(1.21)	185.19(0.88)	193.67(3.09)	2593.48(3.89)
	50	full	489.82(0.98)	353.10(1.81)	222.09(1.98)	370.11(2.00)	4535.72(2.19)
		part	422.94(0.72)	287.34(1.90)	277.78(2.09)	290.05(0.56)	3890.22(1.96)
	100	full	734.72(2.01)	529.65(1.10)	499.60(0.72)	555.17(0.57)	6522.67(4.01)
		part	507.53(0.91)	344.81(1.90)	333.24(2.89)	348.61(3.90)	4668.26(2.67)

4.4 Numerical Experiments

We describe three dynamic models that extend the spatial Poisson equation described in Section 4.2.2 to incorporate temporal dynamics, and the resulting multiway (inverse) covariance structure. These models will be used in our numerical experiments to generate data to demonstrate the performance of the proposed multiway ensemble Kalman filter algorithm.

Poisson-AR(1) Process. The first extension, which we call the Poisson-AR(1) process, imposes an autoregressive temporal model of order 1 on the source function f in the Poisson equation (4.3). Specifically, we say a sequence of discretized spatial observations $\{\mathbf{U}^k \in \mathbb{R}^{d_1 \times d_2}\}_k$ indexed by time step $k = 1, \cdots, T$ is from a Poisson-AR(1) process if

$$
\begin{aligned}
&(\mathbf{A}_{d_1} \oplus \mathbf{A}_{d_2}) \operatorname{vec}(\mathbf{U}^k) = \operatorname{vec}(\mathbf{Z}^k), \\
&\operatorname{vec}(\mathbf{Z}^k) = a \operatorname{vec}(\mathbf{Z}^{k-1}) + \operatorname{vec}(\mathbf{W}^k), \quad |a| < 1,
\end{aligned}
$$

where $\{\mathbf{W}^k \in \mathbb{R}^{d_1 \times d_2}\}_k$ is spatial white noise, i.e., $W^k_{i,j} \sim \mathcal{N}(0, \sigma_w^2)$, i.i.d.

Convection-diffusion Process. The second time-varying extension of the Poisson PDE model (4.3) is based on the convection-diffusion (C-D) process (Chandrasekhar, 1943)

$$
\frac{\partial u}{\partial t} = \theta \sum_{i=1}^{2} \frac{\partial^2 u}{\partial x_i^2} - \epsilon \sum_{i=1}^{2} \frac{\partial u}{\partial x_i}. \tag{4.9}
$$

Here, $\theta > 0$ is the diffusivity; and $\epsilon \in \mathbb{R}$ is the convection velocity of the quantity along each coordinate. Note that for simplicity of discussion here, we assume these coefficients do not change with space and time (see, Stocker (2011), for example, for a detailed discussion). These equations are closely related to the Navier-Stokes equation commonly used in stochastic modeling for weather and climate prediction (Chan-

drasekhar, 1943; Stocker, 2011). Coupled with Maxwell's equations, these equations can be used to model magneto-hydrodynamics (Roberts, 2006), which characterize solar activities including flares.

A solution of Equation (4.9) can be approximated similarly as in the Poisson equation case, through a finite difference approach. Denote the discrete spatial samples of $u(\mathbf{x}, t)$ at time t_k as a matrix $\mathbf{U}^k \in \mathbb{R}^{d_1 \times d_2}$. We obtain a discretized update propagating $u(\mathbf{x}, t)$ in space and time, which locally satisfies

$$\begin{aligned} \frac{U_{i,j}^k - U_{i,j}^{k-1}}{\Delta t} = \theta & \left(\frac{U_{i+1,j}^k + U_{i-1,j}^k + U_{i,j+1}^k + U_{i,j-1}^k - 4U_{i,j}^k}{h^2} \right) \\ & - \epsilon \left(\frac{U_{i+1,j}^k - U_{i-1,j}^k + U_{i,j+1}^k - U_{i,j-1}^k}{2h} \right), \end{aligned} \tag{4.10}$$

where $\Delta t = t_{k+1} - t_k$ is the time step and h is the mesh step (spatial grid spacing). Then, the temporal update of $\mathbf{U}^k$ can be shown to obey the Sylvester matrix update equation (Thomas, 2013) $\mathbf{A}_{d_1}\mathbf{U}^k + \mathbf{U}^k\mathbf{A}_{d_2}^T = \mathbf{U}^{k-1}$, or equivalently,

$$(\mathbf{A}_{d_2} \oplus \mathbf{A}_{d_1}) \operatorname{vec}(\mathbf{U}^k) = \operatorname{vec}(\mathbf{U}^{k-1}), \tag{4.11}$$

where $\mathbf{A}_{d_1} = \mathbf{A}_{d_1}(\theta, \epsilon, h, \Delta t)$ and $\mathbf{A}_{d_2} = \mathbf{A}_{d_2}(\theta, \epsilon, h, \Delta t)$ are symmetric tridiagonal matrices whose entries depend on θ, ϵ, Δt and h (Grasedyck, 2004).

Kuramoto-Sivashinsky Process. The third extension of a spatial diffusion process is the Kuramoto-Sivashinsky (K-S) equation, which is a class of non-linear fourth-order PDEs known to exhibit chaotic behaviors (Hyman and Nicolaenko, 1986). Specifically, the K-S equation in a 2D spatial domain can be written as

$$u_t + \Delta u + \Delta^2 u + \frac{1}{2}|\nabla u|^2 = 0,$$

or equivalently,

$$\frac{\partial u}{\partial t} + \sum_{i=1}^{2} \frac{\partial^2 u}{\partial x_i^2} + \sum_{i=1}^{2} \frac{\partial^4 u}{\partial x_i^4} + \frac{\partial^4 u}{\partial x_1^2 \partial x_2^2} + \sum_{i=1}^{2} \Big(\frac{\partial u}{\partial x_i} \Big)^2. \tag{4.12}$$

Here, although we can similarly apply finite difference approximation to the differential operators, the equation is non-linear and simple linear algebraic update like in the Poisson-AR and convection-diffusion cases is not available.

For numerical illustrations, we consider a 2D spatio-temporal process of dimension 64×64 where only half of the entries are observed, which leads to a measurement matrix $\mathbf{H} \in \{0, 1\}^{2048 \times 4096}$. We generated the true states and the corresponding observations according to Poisson-AR(1), convection-diffusion, and Kuramoto-Sivashinsky dynamics for $T = 50$ time steps. Several realizations of the true state variables are shown in Figure C.1 of Appendix C. At each time step, we generated an ensemble of size $N = 15$ and estimated the state covariance / inverse covariance using several sparse (multiway) inverse covariance estimation methods, including Glasso (Friedman et al., 2008), KPCA (Greenewald and Hero, 2015), KGlasso (Tsiligkaridis et al., 2013), TeraLasso (Greenewald et al., 2019), SG-PALM (Wang and Hero, 2021b).

Figure IV.1 shows evolution of the computed root mean squared errors (RMSEs) for the estimated states under the Poisson-AR (left panel) and the convection-diffusion (right panel) processes across all ensemble members. It is noted that SG-PALM, which corresponds to the statistical method that models the inverse covariance as a squared Kronecker sum, performs the best under the Poisson-AR generating process. In Figure IV.2 (a) we show the true and estimated (inverse) covariance matrices obtained at the last time step – at each time step the multiway EnKF involves estimation of a sparse Kronecker sum squared inverse covariance matrix induced by the Poisson-AR process. Hence, the SG-PALM method operates under the correct model assumption in this situation. On the other hand, the KPCA method outperforms other methods

as time progresses. This is due to the fact that the inverse covariance structure under the convection-diffusion dynamics model is dense due to the smoothing nature of the Kalman filtering algorithm. But, its steady-state covariance has low-dimensional structures as shown in Figure IV.2 (b). The KPCA in this case was able to approximate this structure reasonably well as it does not impose any sparsity on the precision matrix. Remarks IV.1 and IV.3 below further discuss this emergence of dense precision matrix for the marginal spatial process. Appendix C illustrates situations where the joint spatio-temporal precision matrix is sparse. Note in this case, the EnKF with Glasso converges slower than the multiway methods.

Remark IV.1. *Although the state variable following the convection-diffusion dynamics evolves via a Sylvester equation, similar to the Poisson-AR case, the state (inverse) covariance matrix at time step t_k admits different structures. Specifically, the state precision matrix $\mathbf{\Omega}^k = \mathrm{cov}^{-1}(\mathrm{vec}(\mathbf{U}^k)) \in \mathbb{R}^{d_1 d_2 \times d_1 d_2}$ evolves as $\mathbf{\Omega}^k = (\mathbf{A}_{d_1} \oplus \mathbf{A}_{d_2})\mathbf{\Omega}^{k-1}(\mathbf{A}_{d_1} \oplus \mathbf{A}_{d_2}) + \sigma_w^{-2}\mathbf{I}$ (see Katzfuss et al. (2016), for example). This matrix is not necessarily sparse for finite k but, assuming that the eigenvalues of the matrix $\mathbf{A}_{d_1} \oplus \mathbf{A}_{d_2}$ are in $(-1, 1)$, the limiting precision matrix $\mathbf{\Omega}^\infty = \lim_{k\to\infty} \mathbf{\Omega}^k$ is $\mathbf{\Omega}^\infty = (\mathbf{A}_{d_1} \oplus \mathbf{A}_{d_2})\mathbf{\Omega}^\infty(\mathbf{A}_{d_1} \oplus \mathbf{A}_{d_2}) + \sigma^{-2}\mathbf{I}$. The $\mathbf{\Omega}^\infty$ matrix is sparse because $\mathbf{A}_{d_1}$ and $\mathbf{A}_{d_2}$ are both tridiagonal.*

Tacking the Poisson-AR and convection-diffusion dynamics with EnKF both involve sparse (on either the covariance or its inverse) and tractable linear updates. The Kuramoto-Sivashinsky dynamical model will similarly involve sparse updates if finite difference approximations are employed for solving the PDE because the discretized differential operators will always be sparse. But, the non-linear nature of the problem makes the update intractable. Moreover, the KS equation is known to generate chaotic behaviors, making it a more realistic benchmark model for real-world systems. Here, two of the best performers under the Possion-AR and convection-diffusion dynamics (SG-PALM and KPCA) are compared against the ensemble transform Kalman

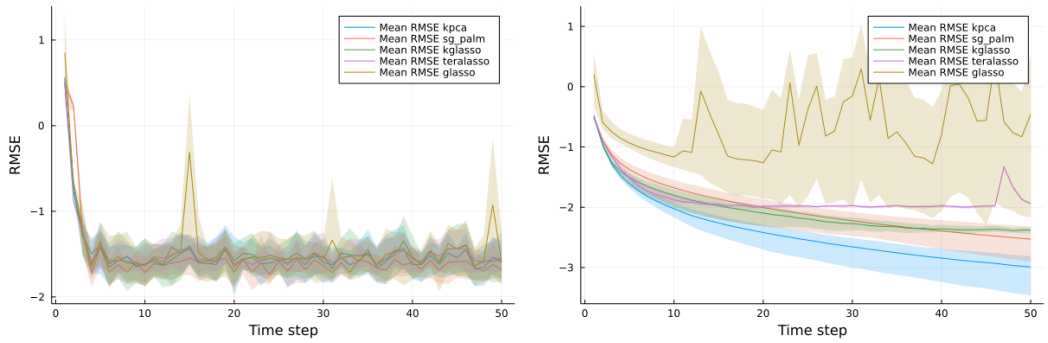
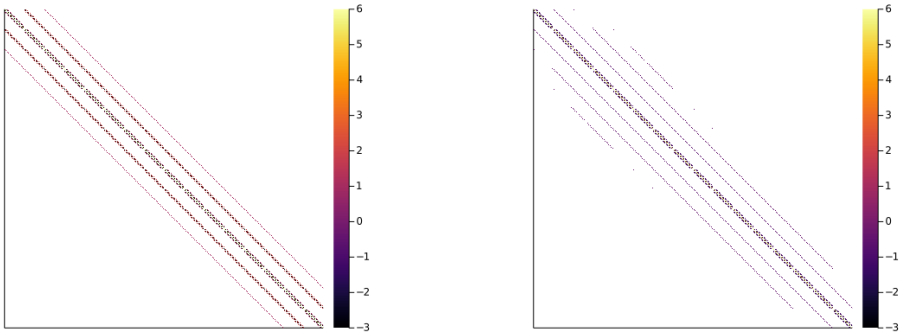


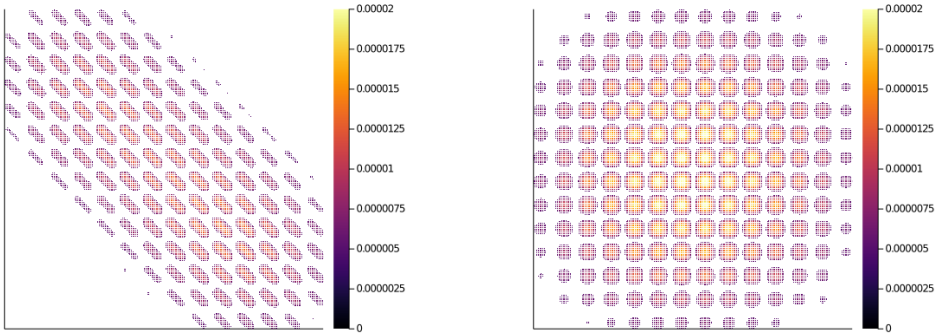
Figure IV.1: RMSEs of the estimated states via EnKF over 50 time steps using different (inverse) covariance estimators. The 95% posterior interval for RMSEs over all ensemble members are shown here with the posterior mean highlighted using solid lines. Here, each state is of dimension 64×64 and is generated via either a convection-diffusion (right) or Poisson-AR(1) equation (left). The best performers in terms of mean RMSE over all ensemble members are KPCA for convection-diffusion and SG-PALM for Poisson-AR(1).

filter (ETKF) and its localized version, a method known to work well for tracking high-dimensional highly non-linear systems with limited ensemble size. It has been successfully applied for data assimilation of, for example, the solar photospheric magnetic flux, which are fundamental drivers for simulations of the corona and solar wind (Hickmann et al., 2015). Figure IV.3 (a) shows that the proposed multiway EnKF outperforms the (local) ETKF. The KPCA based estimator outperforms the SG-PALM based estimator as time progresses, likely due to a similar reason discussed previously – the inverse covariance structure becomes denser and denser, making the sparse models less appealing. The local ETKF performs similarly well as the non-local version but facilitates parallel estimation schemes where the "local patches" of the state variable can be updated and evolved simultaneously. Figure IV.3 (b) visualizes the true and estimated KS states by SG-PALM multiway EnKF at several timestamps. It shows that the proposed method can correct the noisy observations (with missing values) and recovers the true states reasonably well.

Remark IV.2. *The Sylvester matrix equations (and hence the sparse Kronecker structures) arise when the finite-difference discretization is performed on a rectan-*



(a) Poisson-AR inverse covariance structure (left) and the estimate obtained by SG-PALM (right) at the last time step.



(b) Convection-diffusion covariance structure (left) and the estimate obtained by KPCA at the last time step.

Figure IV.2: Covariance/precision structures for Poisson-AR and convection-diffusion dynamics and their estimates. Here, white/blank entries indicate zeros in the (inverse) covariance matrix. For Poisson-AR dynamics the Sylvester graphical model approximately matches the true structure of the precision matrix. For convection-diffusion dynamics the covariance instead of the precision matrix is structured and sparse.

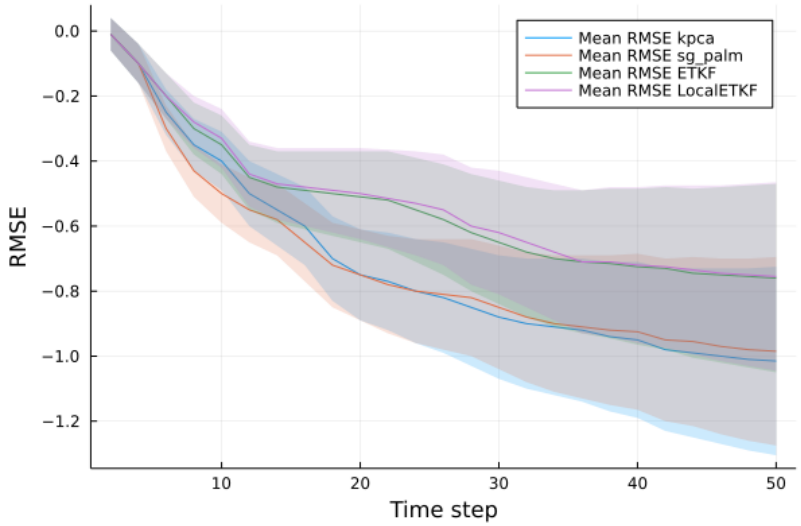
gular grid. The relations (4.4) *and* (4.11) *might not hold for finite-difference on, for example, spherical coordinates, as well as approximations to the equations using other types of methods, such as finite volume, finite element, and spectral methods.*

Remark IV.3. *Although the precision matrix of the state ensemble becomes dense as the temporal update progresses, making sparse Kronecker-structured methods less appealing as illustrated in Figure IV.1 and IV.2, if we consider "temporal blocks" of states then the precision matrix remains sparse. Appendix C includes detailed derivations of the blocked versions of the Poisson-AR and convection-diffusion dynamics, and illustrates the performances of the Kronecker-structured models under these sce-*

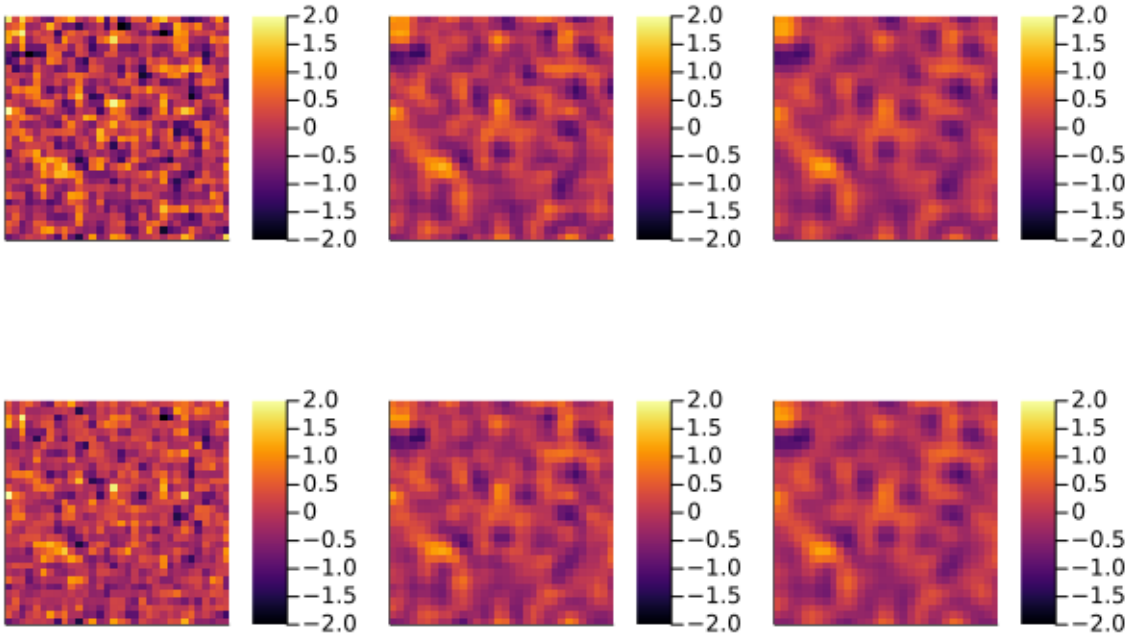
narios.

4.5 Conclusions

Spatiotemporal PDEs are prominent techniques for modeling real-world physical systems. In this chapter, we introduced a multiway ensemble Kalman filtering framework that integrates the powerful ensemble Kalman filters with state-of-the-art Kronecker-structured covariance/precision models. The resulting framework allows one to track simulated complex, potentially chaotic systems. One such system in the real world arises in space physics, where solar flares and coronal mass ejections are associated with rapid changes in field connectivity and are powered by partial dissipation of electrical currents in the solar atmosphere (Schrijver et al., 2008). The nonlinear force-free field model is often used to describe the solar coronal magnetic field (DeRosa et al., 2015; Wheatland and Gilchrist, 2013) and can be derived from the convection-diffusion process described in this work. Additionally, global maps of the solar photospheric magnetic flux are fundamental drivers for simulations of the corona and solar wind. However, observations of the solar photosphere are only made intermittently over approximately half of the solar surface. Hickmann et al. (2015) introduced the Air Force Data Assimilative Photospheric Flux Transport model that uses localized ensemble transform Kalman filtering to adjust a set of photospheric simulations to agree with the available observations. In future work, we plan to incorporate our proposed multiway EnKF framework for tracking these solar physical systems.



(a) Root mean squared errors in log 10 scale for state estimated by EnKF variants. The solid lines and the shaded areas indicate the posterior mean and the 95% posterior interval over all ensemble estimates.



(b) Comparison between true states (top) and estimated states by EnKF with the SG-PALM inverse covariance estimator (bottom) at several time stamps.

Figure IV.3: Visualizations of the performances by various EnKF methods for tracking the Kuramoto-Sivashinsky system. The proposed multiway EnKF outperforms the ETKF and its localized version.

CHAPTER V

A Geometry-driven Framework for Dynamic Topic Modeling

A simple and scalable framework for longitudinal analysis of text data is developed that combines latent topic models with computational geometric methods. Dimensionality reduction tools from computational geometry are applied to learn the intrinsic manifold on which the latent, temporal topics reside. Then shortest path distances on the manifold are used to link together these topics. The proposed framework permits visualization of the low-dimensional embedding, which provides clear interpretation of the complex, high-dimensional trajectories that may exist among latent topics. Practical application of the proposed framework is demonstrated through its ability to 1) capture and effectively visualize natural progression of latent COVID-19–related topics learned from Twitter data; 2) learn latent topics correspond to human-labeled data and "generate" novel latent topics from TalkLife – a peer support network focused on mental health. Interpretability of the trajectories and the learned topics is achieved by comparing to real-world events and expert knowledge (e.g., labeled data). The analysis demonstrates that the proposed framework is able to 1) capture granular-level impact of COVID-19 on public discussions; and 2) learn mental health focused topic clusters that resemble human-level expert knowledge.

5.1 Introduction

The continued digitization of public discourse in news feeds, books, scientific reports, social media, blogs, microblogs, and web pages creates opportunities to discover meaningful patterns and trends of public opinion. Methods of probabilistic topic modeling have been used to extract such patterns using a suite of algorithms that aim to automatically discover and annotate large collections of documents with thematic labels (Blei, 2012). Topic modeling algorithms are computational methods that manipulate word frequencies in document corpora to discover the themes that run through them, quantify how those themes are connected to each other, and how they change over time.

5.1.1 Probabilistic topic models and computational geometry

A probabilistic topic model that has seen success in many applications is the latent Dirichlet allocation (LDA) model (Blei et al., 2003), which uses a latent topic model to extract thematic information from document corpora to infer an underlying generative process that explains hidden relationships among documents. Many real-world document corpora, however, have complex structure and include temporal information that is ignored by traditional LDA models. For example, discussions of COVID-19 on Twitter between February and May 2020 involve the emergence, evolution, and extinction of multiple topics over time. Moreover, data generated from emerging social media platforms, such as Twitter, Reddit, TalkLife, etc. are short bursts composed in micro-text (Ellen (2011)), which traditional LDA models struggle to model effectively. Additionally, side information is commonly available such as document-level labels/tags or word-level features. For example, a significant proportion of the news articles on Reuters is labeled with multiple human-provided tags (Ramage et al., 2009). Effectively incorporating these additional information is key to reliability and interpretability of many machine learning algorithms, including

LDA and topic models.

Extensions of the standard LDA have been proposed to learn latent topics in the context of complex structure and temporal information. An early modeling strategy is to assume a temporally Markovian relationship where the state of the process at time $t+1$ is independent of past history given the state at time t. Blei and Lafferty (2006) proposed the dynamic topic model (DTM) for modeling time-varying topics, where the topical-alignment over time is captured by a Kalman filter procedure. Further improvements have been in various directions, including: (1) relaxation of the Markov assumption, as discussed by Wang and McCallum (2006), who introduced a non-Markov continuous-time model called the topics-over-time (TOT) model, capturing temporal changes in the occurrence of the topics themselves, and (2) circumvent of time discretization, as proposed by Wang et al. (2008) that improved the DTM using a continuous time variant, called cDTM, formulated on Brownian motion to model the latent topics in a longitudinal collection of documents. These approaches rely on spatiotemporally coupled stochastic processes for modeling the evolution of topics over time. Such integrated models employ a global joint parameterization of time evolution and word co-occurrence, producing a unified generative probabilistic model for both temporal and topical dimensions.

However, global parameterized DTMs have several deficiencies that motivate the model proposed in this article. The main issue is that global parameterization can increase the computational complexity of parametric inference. Wang et al. (2008) and Blei and Lafferty (2006) argued that applying Gibbs sampling to perform inference on DTMs is more difficult than on static models, principally due to the nonconjugacy of the Gaussian and multinomial distributions. As an alternative, they proposed the use of inexact variational methods, in particular, variational Kalman filtering and variational wavelet regression, for inference. These approximate inference procedures face two issues: 1) they usually involve assumptions on the correlation structures among

latent variables, for example, mean-field, which undermines uncertainty quantification; 2) the resulting optimization problems are usually nonconvex, which means that the approximate posterior distribution found might only be locally optimal–trapping the topic parameters in a neighborhood of a local optima. An additional issue is that posterior inference via variational approximation usually relies on batch algorithms that need to scan the full data set before each update of the model. This increases the computational burden, especially for long time sequences, and parallel computing cannot be easily exploited (Bhadury et al., 2016). Such issues can lead to numerical instability and lack of interpretability of the model predictions. Furthermore, incorporating side information, such as document-level labels, word-level features (e.g., word volumes), imposes additional challenges to dynamic topic modeling. Hong et al. (2011) proposed a variant of DTM for tracking topic trends, by incorporating word volumes and assuming these volumes are generated by the latent topics through a linear model. Similarly, Park et al. (2015) introduced a supervised DTM (sDTM), where a time-series of numerical values are assumed to be generated by the topic assignment distributions via a normal linear model. Both of these variants require the aforementioned Kalman filtering and variational approximation procedures for inference, in addition to the extra modeling assumption on the side information.

Rather than jointly modeling word co-occurrence and the temporal dynamics, there exist alternatives that adopt simpler analysis strategies that motivate our proposed approach. Most of these approaches to nonglobal modeling involve fitting a local time-unaware topic model to predivided discrete time slices of data, and then examining the topic distributions in each time-slice in order to assemble topic trends that connect related topics (Griffiths and Steyvers, 2004; Wang et al., 2005; Malik et al., 2013; Cui et al., 2011). A difficulty with these approaches is that aligning the topics from each time slice can be challenging, even though several strategies have been proposed. Malik et al. (2013) proposed a framework to connect every pair of top-

ics from adjacent time slices whose similarity, measured by the cosine metric, exceeds a certain threshold. Cui et al. (2011) used a semiparametric clustering algorithm to identify similar topics at adjacent time slices. However, these approaches suffer from an inherent inflexibility in modeling diverse dynamical structures that exist in a potentially large collection of temporal topic sequences. Such methods are developed to model and visualize specific, and relatively rare, types of temporal dynamics and are often not able to capture all types of variations, for example, anomalies, bifurcations, emergence, convergence, and divergence.

We propose a flexible and scalable computational geometry framework that remedies the above mentioned issues and complements the existing methods in the dynamic topic modeling toolbox. Specifically, in this article a time-evolving topic model is introduced that uses a local LDA-type model for discrete time slices of collections of documents, and a geometric proximity model to align the topics from time to time. In contrast to global parametric dynamic latent variable approaches to summarizing time-evolving unstructured texts, our framework offers a wrapper for a suite of tools. The proposed wrapper framework has the flexibility to allow any particular topic model to be applied locally to each time slice of documents. This allows any side information to be included via supervised/semi-supervised variants of LDAs (e.g., Ramage et al. (2009); Mcauliffe and Blei (2007); Petterson et al. (2010); Zhu et al. (2012); Lu et al. (2011)). It then implements a fast and scalable shortest path algorithm to stitch together the locally learned LDA topics into an integrated collection of temporal topic trends.

To facilitate visualization and interpretation of the learned topic trends, an emphasis of this article, the proposed framework also implements a recent geometric embedding method called PHATE (Potential of Heat-diffusion for Affinity-based Trajectory Embedding) that projects the high-dimensional word distributions representing latent topics to lower dimensional coordinates. The PHATE embedding has been shown to

preserve the intrinsic geometry of high-dimensional time-varying data (Moon et al., 2019), which provides a clear and intuitive visualization of any progressive structure that exists among the topics. We note that similar computational geometric representations of data have been used in unsupervised, semisupervised, and supervised learning, both as principal learning models and as supplementary regularizers of other models. In manifold learning, geometric affinity (or distance) between data points drives dimensionality reduction (Tenenbaum et al., 2000; Donoho and Grimes, 2003) and dimensionality estimation methods (Costa and Hero, 2006). Several deep learning architectures, like the deep k-nearest neighbors (Papernot and McDaniel, 2018, DkNN), use interpoint distances and the kNN classifier to induce interlayer representational continuity and robustness against adversarial attacks. Semisupervised classification approaches adopt geometric measures over reproducing kernel Hilbert space (RKHS) to associate unlabeled data with labeled data in geometry-regularized empirical loss frameworks (Belkin and Niyogi, 2004). Geometry is the driver for many missing data models, for example, synthetic minority oversampling technique (Chawla et al., 2002, SMOTE) and more generally, nearest neighbor interpolations.

We point out that dimensionality reduction is the basis for latent semantic analysis (LSA) in computational linguistics. In particular, Doxas et al. (2010) had a similar objective to ours, to explore temporal evolution of discourse, but in long text with labeled corpora. The authors constructed semantic spaces for various corpora, and then calculated the intrinsic dimensionality of the paragraph trajectories through these corpora. The work focuses on investigating the intrinsic dimension of the trajectories and they used LSA to construct representations of the texts. However, they did not address the topic alignment or trajectory clustering problems for which our PHATE and Hellinger shortest path framework is designed.

5.1.2 Application to Twitter data

Enabled by our proposed longitudinal dynamic topic model, we leverage recent activity on social media to understand the impact of the COVID-19 pandemic and, in particular, its impact on social discourse. The utilization of novel data sources is vital, as the current data landscape for understanding the pandemic remains imperfect. For example, public databases maintained by Johns Hopkins University (`https://bit.ly/2UqFSuA`) and *The New York Times* (`https://bit.ly/2vUHfrK`) provide incoming county-level information of confirmed cases and deaths. Unfortunately, these data streams are of limited utility due to limited testing capacity and selection bias (Dempsey, 2020). The public health community requires auxiliary sources of information to improve national and local health policy decisions. A critical question is whether there are complementary data streams that may be leveraged to better understand the COVID-19 pandemic in the United States. Social media platforms, such as Twitter, Reddit, Facebook, and so on, are examples of such data streams. These platforms generate high resolution spatiotemporal data sets that concern public opinions on various societal issues, including health care, government decisions, and politics, all of which could be highly relevant to understanding the impact of COVID-19.

Although use of these novel data streams create new challenges due to limitations such as high noise level, high volume, and selection bias, many recent efforts have explored social media data as a complementary source to traditional health care data and applied topic models to understand public concerns toward COVID-19 (Doogan et al., 2020; Stokes et al., 2020; Boon-Itt and Skunkan, 2020; Xue et al., 2020; Jang et al., 2020), as well as related socioeconomic issues (Su et al., 2021; Sha et al., 2020; Liu et al., 2020). Here, we extract information from Twitter, a particularly popular social media platform, and focus on studying its spatiotemporal behaviors that are believed to be affected by COVID-19. We use subsamples of tweets generated from

February 15, 2020, to May 15, 2020, a period over which a large volume of COVID-19-related tweets occurred. An extended analysis of data collected from May to August 2020 can be found on `https://github.com/ywa136/twitter-covid-topics`. We apply our temporal topic modeling framework to discover sets of COVID-19-related latent topics that impact public discourse.

5.1.3 Application to TalkLife data

In addition to its lasting physical health effects, a side effect of the COVID-19 is a noticeable and disproportionate increase in the global burden of depressive and anxiety disorders worldwide. Santomauro et al. (2021) showed that the pandemic led to a dramatic rose in the overall number of cases of mental disorders, with an additional 53.2 million and 76.2 million cases of anxiety and major depressive disorders (MDD), respectively. Even before the COVID-19 pandemic, mental health disorders posed a significant burden worldwide. However, access to mental healthcare resources remain poor worldwide. Online social media platform, especially peer-to-peer support platforms attempt to alleviate this fundamental gap by enabling those who struggle with mental illness to provide and receive social support from their peers.

Recent work found that social media big data combined with NLP and machine learning techniques can help address public health, especially mental health, research questions (Conway and O'Connor, 2016; De Choudhury, 2013; Gkotsis et al., 2016; De Choudhury et al., 2016; Kim et al., 2021; Amir et al., 2019). As another practical application of this work, we use data from TalkLife (`https://www.talklife.com`), the largest online peer-to-peer support platform for mental health support. Several work (Sharma et al., 2020b,a, 2021) have demonstrated the usefulness of the TalkLife data as a machine learning dataset for training models that help improve understanding of the mental health issues. Here, we obtain posts from the platform in the year of 2019 and apply our temporal topic modeling framework to extract mental health

related discussions. A distinguishing feature of the TalkLife data is that labels for the posts are created by human experts for further investigation of the contents and to aid potential early prevention and intervention of mental health issues. Extending the Twitter analysis, we develop a weakly-supervised temporal topic model and show that our framework is able to capture latent topics that correlate well with a set of labels created by human experts.

5.1.4 Key contributions and outline of the chapter

We highlight key contributions of this article:

- A modular framework that provides a wrapper for a suite of tools for interpretation and visualization of temporal topic models.
- A new approach for aligning independently learned topic models over time based on computational geometry.
- A scheme for visualizing and understanding temporal structures of the aligned topics via manifold learning.

The remainder of the article is organized as follows: Section 5.2 introduces the methods and tools that have been applied in our analysis framework. Section 5.3 and Section 5.4 present numerical results and visualizations with several case studies. Section 5.5 gives some concluding remarks.

5.2 Methods

In this section, we discuss the building blocks for the proposed framework: Section 5.2.1 briefly describes the LDA model and its variants for dealing with micro-text; Section 5.2.2 introduces two key components of the framework for propagating and associating topics over time; and Section 5.2.3 reviews and applies a dimension reduction technique to visualize the temporal trajectories of the evolving topics.

5.2.1 LDA for micro-text documents

Since the literature in probabilistic topic models and their dynamic variants is enormous (see Blei (2012) for a survey), we focus our discussion on the LDA (Blei et al., 2003), which is the building block for all other algorithms targeting similar applications. A graphical model representing its generating process is presented in Appendix 4.1. The idea of LDA is: from a collection of documents (each composed of set of words $w_{d,n}$), one is able to infer the per-word topic assignment $z_{d,n}$, the per-document topic proportions θ_d, and the per-corpus topic distributions β_k, through a joint posterior distribution $p(\theta, z, \beta|w)$. Numerous inference algorithms are developed to handle data at scale, for example, variational methods (Blei et al., 2003; Teh et al., 2008; Hoffman et al., 2013; Mimno et al., 2012; Srivastava and Sutton, 2017), expectation propagation (Minka and Lafferty, 2002), collapsed Gibbs sampling (Griffiths and Steyvers, 2002), distributed sampling (Newman et al., 2008; Ahmed et al., 2013), and spectral methods (Arora et al., 2012; Anandkumar et al., 2014). The posterior expectations can then be used to perform the task at hand: information retrieval, document similarity determination, exploration, and so on.

The standard LDA, however, may not work well with micro-text like tweets. In particular, each tweet usually concentrates on a single topic, and it is not reasonable to consider one tweet as a document in the traditional sense as there is limited data (e.g., word co-occurrences) from which the latent topics can be learned. To overcome this "data sparsity" issue, efforts have been made along on three major directions (Qiang et al., 2020): 1) methods predicated on the assumption that each text (e.g., tweet) is sampled from only one latent topic; 2) methods utilizing global (i.e., the whole corpus) word co-occurrences structures; 3) methods based on aggregation/pooling of texts into 'pseudo-documents' prior to topic inference.

In this article, we apply the Twitter LDA model (T-LDA, Zhao et al., 2011), for modeling topics at each time slice. T-LDA can be categorized along the directions 1)

and 3) mentioned above. But we note that the proposed framework works with any topic model that outputs word distributions representing learned latent topics. We selected T-LDA since it has been widely used in many related applications, including aspect mining (Yang et al., 2016), user modeling (Qiu et al., 2013), and bursty topic detection (Diao et al., 2012). The generative model underlying T-LDA assumes that there are K topics in the Tweets, each represented by a word distribution, denoted as β_k for topic k and β_B for background words. Let θ_u denote the topic assignment distribution for user u. Let π denote a Bernoulli distribution that governs the choice between background words and topic words. The generating process for a tweet is as follows: a user first chooses a topic based on its user-specific topic assignment distribution. Then the user chooses a bag of words one-by-one based on the chosen topic or the background model. The generation process is summarized in Algorithm V.1, and a plate notation comparison between the T-LDA and standard LDA is included in Appendix 4.1. Similarly to a standard LDA algorithm, parameters in each multinomial distribution are governed by symmetric Dirichlet priors. The model inference can be performed using collapsed Gibbs sampling (code available at `https://github.com/minghui/Twitter-LDA`). Due to space limitations we leave out derivation details and sampling formulas. More details on the implementation can be found in Appendix 4.1.

Weak Supervision with Word-level Prior Knowledge: To encourage topic models to learn latent topics that correlate directly with word-level side information, we augment them with a weakly supervised signal in the form of seed words. Rather than fully guiding the model with labels in a supervised version, as in for example, Mcauliffe and Blei (2007) and Ramage et al. (2009), we use a set of seed words to define an asymmetric prior on the word-topic distributions. The reasons for modeling choice are twofold: 1) In one of the applications we concern, we have access to micro-text

Algorithm V.1: Generating process for T-LDA

Input: Constants η, γ
  Draw $\beta_B \sim \text{Dir}(\eta)$, $\pi \sim \text{Dir}(\gamma)$
  for topic $k = 1, \ldots, K$ **do**
    Draw $\beta_k \sim \text{Dir}(\eta)$
  end for
  for user $u = 1, \ldots, U$ **do**
    Draw $\theta_u \sim \text{Dir}(\alpha)$
    for Tweet $s = 1, \ldots, S_u$ **do**
      Draw $z_{u,s} \sim \text{Multi}(\theta_u)$
      for word $n = 1, \ldots, N_{u,s}$ **do**
        Draw $y_{u,s,n} \sim \text{Multi}(\pi)$
        if $y_{u,s,n} = 0$ **then**
          Draw $w_{u,s,n} \sim \text{Multi}(\beta_B)$
        else
          Draw $w_{u,s,n} \sim \text{Multi}(\beta_{z_{u,s}})$
        end if
      end for
    end for
  end for

data from TalkLife and the corresponding labels for each post. However, the labeling is noisy (labels could be wrong/imperfect), limited (not every post is labeled), and have overlaps (a post could be tagged with multiple related labels). Directly applying supervised LDAs that assume perfect labeling may not be appropriate. 2) We hope to learn novel latent topics that have not yet been discovered and/or have been missed by domain experts.

Using seed words as a form of word-level side information has been considered by a few researchers (Lu et al., 2011; Zhu et al., 2009; Wang et al., 2010), although their goal was normally aspect ratings and multi-aspect sentence labeling. In this work, we characterize our prior knowledge (seed words) for the original T-LDA model using a conjugate Dirichlet prior to the multinomial word-topic distributions. We define a combined conjugate prior for each word n in the vocabulary V as $\beta_k \sim \text{Dir}\Big(\{\eta + w_n\}_{n \in V}\Big)$ for each topic k, where w_n can be interpreted as an equivalent sample size, i.e., the impact of our asymmetric prior is equivalent to adding w_n pseudo

counts to the sufficient statistics of the topic to which word n belongs. In practice, w_n can be obtained empirically as proportional to the volume of a word n in the corpus. We set $w_n = 0$ when we do not have prior knowledge of a word.

5.2.2 Time evolution of topics and shortest paths

Instead of explicitly building the temporal structures into the model as in the globally parameterized DMT and its variants, we propose a two-stage approach: 1) construct a new corpus at each time point via subsampling the documents and independently fitting a topic model to each new corpus; 2) link each of these time points together via shortest distance paths through topics.

Temporal smoothing by subsampling A subsample of tweets is constructed at each time point by conditional sampling of all the tweets with a sampling distribution that is inversely proportional to the temporal proximity of the tweet. This produces subsamples that are local mixtures of tweets at nearby time points, accomplishing a degree of temporal smoothing prior to topic analysis.

To clarify the subsampling procedure, we give a simple example. Assume that the corpus is composed of five tweets per day over a 5-day period. We write d_t as the set of five tweets on day t. On the first day ($t = 1$), exponential weights are computed $w_1 = \{1.000, 0.7500, 0.5625, 0.4219, 0.3164\}$ and normalized by their sum, defining the sampling distribution used to construct the subsample. That is, at time 1, the subsample consists of 100% of all the tweets from day 1, 75% of the tweets from day 2, $75\%^2 = 56.25\%$ of the tweets from day 3 and so on. This subsample is denoted $C_1 = \{s_1, \ldots, s_5\}$. To construct the subsample on day 2, we condition on the subsample C_1 on day 1 and we construct exponential sampling weights for day 2 of the form $w_2 = \{0.7500, 1.000, 0.7500, 0.5625, 0.4219\}$. We combine the subsample C_1 and weights w_2 to construct the subsample on day 2, denoted C_2,

by either randomly removing extra tweets if the sampling weights on a particular day i decreased or randomly adding more tweets from $d_i - s_i$ if the weights on a particular day i increased. This algorithm ensures a (tuneable) degree of smoothness over the subsamples generated at each time in the sense that subsamples which are close in time are likely to contain similar tweets. The procedure is illustrated in Figure V.1. Specifically, using notations developed above, $C_1 = \{s_1, \ldots, s_5\}$ and $C_2 = \{s_1 - \text{Tweet 3}, s_2 + \text{Tweet 7}, s_3 + \text{Tweet 15}, s_4 + \text{Tweet 20}, s_5 + \text{Tweet 24}\}$, where Tweet 3, Tweet 7, Tweet 15, Tweet 20, and Tweet 24 were randomly chosen.

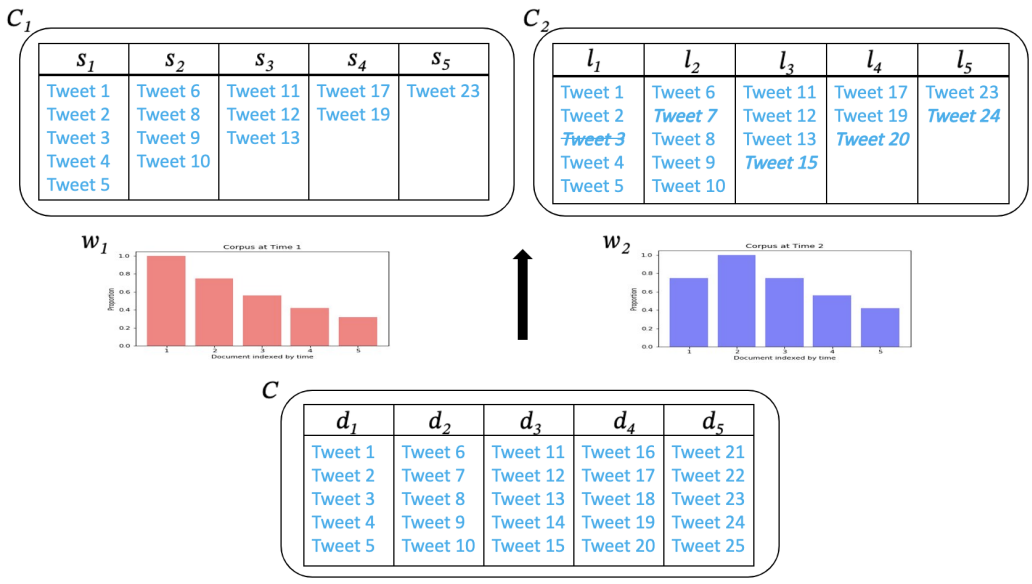


Figure V.1: Conditional subsampling procedure using a hypothetical corpus composed of five documents each containing five tweets. For example, $C = \{d_1, \ldots, d_5\}$, d_1 aggregates tweets from day 1, d_2 aggregates tweets from day 2, and so on. The subsampling weights for each document are shown in the bar plots and are exponentially decaying with a factor of 0.75, centered at day 1 (left, w_1) and day 2 (right, w_2), respectively. Each newly generated corpus is a proportionally weighted random sample and a realization of these samples are shown in the tables (C_1 and C_2). Note that the two corpora differ only by those highlighted and italicized tweets.

After constructing the temporally smoothed corpus, any topic modeling algorithm, such as LDA or its (weakly) supervised variants, can be independently applied to each of the extracted corpora, using either the same set of parameters across all LDA runs, or seeding the next LDA with estimated parameter values (e.g., the posterior mean of

the Markov chain Monte Carlo samples) from the current LDA, as described in Song et al. (2005).

Our proposed temporal smoothing technique is similar to smoothing approaches introduced in spatiotemporal statistics. For example, in time-series analysis the idea of exponential smoothing was proposed in the late 1950s (Brown, 1959; Holt, 2004; Winters, 1960), and has motivated some of the most successful forecasting methods. Forecasts produced using exponential smoothing methods are weighted averages of past observations, with the weights decaying exponentially as the observations become older. As another example, kriging (Krige, 1951) or Gaussian process regression is a widely used method of interpolation in geostatistics. The basic idea of kriging is to predict the value of a function at a given point by computing a weighted average of the known values of the function in the neighboring points (Cressie, 2015). Lastly, in nonparametric regression analysis, the locally estimated scatterplot smoothing (LOESS) is a widely used method that combines multiple regression models in a k-nearest neighbor based framework. Particularly, at each point in a data set a low-degree polynomial is fitted to a subset of the data, with explanatory variable values near the point whose response is being estimated. The subsets used for each of these polynomial fits are determined by a nearest neighbor algorithm. A user-specified 'bandwidth' or smoothing parameter determines how much of the data is used to fit each local polynomial. This smoothing parameter is the fraction of the total number of data points that are used in each local fit.

Dissimilarity between topic word distributions After applying local LDA to each of the time localized subsamples of the smoothed corpus, we stitch together the local LDA results. The alignment of topics with different time stamps is accomplished by creating a weighted graph connecting all pairs of topics where the edge weights are a measure of topic similarity, to be described below. Assume that each local model

generates K topics, resulting in a total of $K \times T$ topics across T time points. A weighted adjacency matrix is constructed from the similarities between $\binom{K \times T}{2}$ topic pairs. The similarities between topics will allow the alignment algorithm to relate topics together across time and enable us to track topic evolution.

As each topic is characterized by the LDA word distribution, any metric that measures dissimilarity between discrete distributions could be used to construct a similarity measure. It is well known that the Euclidean distance is not well adapted to measuring dissimilarity between probability distributions (Amari, 2012). As an alternative, we propose using the Hellinger metric on the space of distributions, which we justify as follows. The LDA word distribution is conditionally multinomial and it lies on a statistical manifold called an *information geometry*, that is endowed with a natural distance metric, called the Fisher-Rao Riemannian metric. Unlike the Euclidean metric, this Riemannian metric characterizes intrinsic minimal (geodesic) distances between multinomial distributions and it depends on the Fisher information matrix $[\mathcal{I}(\theta)]$, θ is the multinomial probability vector. Carter et al. (2009) showed that this metric can be well approximated by the Hellinger distance between multinomial distributions. The Hellinger distance between discrete probability distributions $P = (p_1, \ldots, p_N)$ and $Q = (q_1, \ldots, q_N)$ is defined as

$$H(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{n=1}^{N} (\sqrt{p_n} - \sqrt{q_n})^2}, \quad 0 \leq H(\cdot, \cdot) \leq 1.$$

The major advantages of the Hellinger distance are threefold: 1) it defines a true metric for probability distributions, as compared to, for example, the Kullback-Leibler divergence; 2) it is computationally simple, as compared to the Wasserstein distance; 3) and it is a special case of the f-divergence, which enjoys many geometric properties and has been used in many statistical applications. For example, Liese (2012) showed that f-divergence can be viewed as the integrated Bayes risk in hypothesis testing

where the integral is with respect to a distribution on the prior; Nguyen et al. (2009) linked f-divergence to the achievable accuracy in binary classification problems; Jager and Wellner (2007) used a subclass of f-divergences for goodness of fit testing; Rao (1995) demonstrated the advantages of the Hellinger metric for graphical representations of contingency table data; Srivastava and Klassen (2016) adopted the Hellinger distance to measure distances between functional and shape data; Shemyakin (2014) showed the connection of the Hellinger distance to Hellinger information, which is useful in nonregular statistical models when Fisher information is not available; and finally, Servidea and Meng (2006) derived an identity between the Hellinger derivative and the Fisher information that is useful for studying the interplay between statistical physics and statistical computation.

We note that in previous work on aligning topics the $L2$ or cosine distance is commonly applied (Chuang et al., 2013, 2015; Yuan et al., 2018). As discussed above, these distances are practically and theoretically deficient for aligning distributions. A simulation study is presented in Appendix 4.5 that compares use of these distances to the Hellinger distance, showing that the latter better preserves topic trend coherence.

Nearest neighbor graphs and shortest paths We use the topic graph with Hellinger weights to identify natural progressions from one topic to another over time. We use Dijkstra shortest paths through a nearest neighbor subgraph to identify these progressions. These paths can be interpreted as trajectories of public discourse on the topics identified. This is of interest because we want to understand how conversations around a topic evolves over time. Shortest path analysis allows us to do this with minimal assumptions on the data. In particular, we do not assume or further encourage temporal smoothness in the data beyond the temporally smoothed corpora described in Section 5.2.2.

Due to the noisy nature of social media data and the wide range of topics, we pay

special attention to local neighborhoods of data points. Hence, instead of working with a fully connected graph induced by the full $N \times N$ Hellinger distance matrix of pairwise distances between topics, we build a k-nearest neighbor graph from it. Natural evolution of a topic over time can then be inferred by finding a shortest path of topics on the weighted k-nearest graph, where Hellinger distances represent edge weights. Here, a shortest path is a path between two vertices (i.e., two topics) in a weighted graph such that the total sum of edges weights is minimum, and can be computed efficiently using, for example, Dijkstra's algorithm. The approach of using neighborhood graphs for estimating the intrinsic geometry of a data manifold is justifiable both empirically and theoretically. In manifold learning similar ideas are used to reconstruct lower dimensional geometry from data. For example, the isometric feature mapping (Tenenbaum et al., 2000, ISOMAP) extends metric multidimensional scaling (MDS) by replacing the matrix of Euclidean distances in MDS with the matrix of shortest path distances between pairs of vertices in the Euclidean k nearest neighbor graph. Using such embedding, ISOMAP is able determine lower dimensional structure in high-dimensional data and capture perceptually natural but highly nonlinear "morphs" of the corresponding high-dimensional observations (see figure 4 in Tenenbaum et al. (2000)). Such shortest path analysis is supported by substantial theory (Bernstein et al., 2000; Costa and Hero, 2006; Hwang et al., 2016). Under the assumption that the data points are random realizations on a compact and smooth Riemannian manifold, as the number of data points grows, the shortest paths over the k nearest neighborhood graph converge to the true geodesic distance along the manifold.

In the context of our topic alignment application, this theory suggests that the analogous Hellinger shortest paths should be able to achieve alignment if the empirical LDA word distributions can themselves be interpreted as random draws from an underlying distribution that varies continuously and smoothly over time along a

statistical manifold. To illustrate, Figure V.2 demonstrates how a COVID-19-related topic learned from the corpus on February 15, 2020 (far left), evolves to a COVID-19-related health care–focused topic learned from the corpus on May 15, 2020 (far right). The top row in the figure was constructed by computing the shortest Hellinger distance path on a 10-nearest neighbor graph, whereas the bottom row was constructed using the full graph. As expected, the shortest path on the neighborhood graph captures perceptually natural but highly nonlinear 'morphs' of the corresponding high-dimensional word distributions by transforming them approximately along geodesic paths. On the other hand, the shortest path on the full graph connects the two observations through a sequence of apparently unrelated and nonintuitive topics. In Appendix 4.6, we compare the proposed Hellinger shortest path topic alignment method with TopicFlow (Malik et al., 2013), a common method for topic alignment that uses local matching and Euclidean distances.

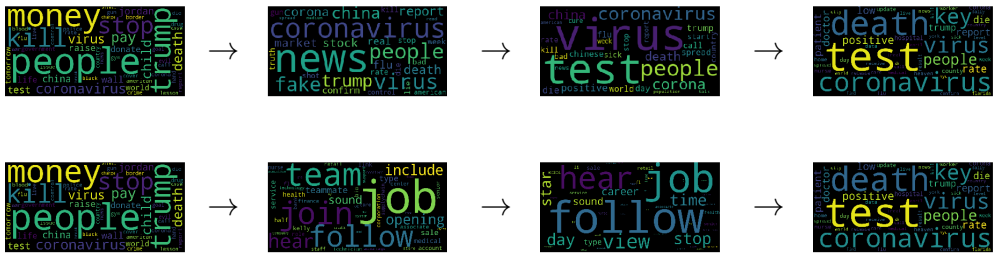


Figure V.2: Evolution along the Hellinger shortest paths of a COVID-19 topic on February 15, 2020, to a COVID-19 topic on May 15, 2020. The paths are computed on a 10-nearest neighbor graph (top) and a fully connected graph (bottom). Each word cloud image represents a topic at a particular time, showing the word distribution encoded by font size (only the top 30 words in each topic are shown). The middle two word clouds represent two intermediate topics on the respective paths and illustrate the benefit of using the k nearest neighbor graph. The middle two topics on the top row seem naturally connected to the beginning and the end topics, in contrast to the bottom row.

The choice of k for the neighborhood graph affects the approximation to the Hellinger geodesic path: choosing a k that is too large creates short circuits in the graph, resulting in a noisy path like the bottom row of Figure V.2; choosing a k that

is too small results in a graph that is disconnected for which there might not exist a path between two points of interest. The problem of selecting an optimal value of k remains open, although several computational data-driven approaches have been proposed for ISOMAP (Tenenbaum et al., 2000; Samko et al., 2006; Gao and Liang, 2011). Here we use $k = 10$, which exceeds the connectivity threshold, to induce the most natural approximation to the true geodesic path between topics of interest. In Appendix 4.8 we establish that our results are robust to perturbations around this value of k.

We also note that the Hellinger shortest paths may differ in length, which is the number of topics that they connect over time. This variation is due to the occasional time skips in the path that occur when the shortest path algorithm does not find an adequate match between topics at successive time points. Such skipping can occur when a topic thread wanes temporarily, merges with another thread, or dies. In Appendix 4.2 we provide statistics on the occurrences of skips for a subset of paths.

5.2.3 Interpretation and visualization of topic trends via low-dimensional embedding

In LDA each latent topic is represented by a vector that lies on a simplex that constitutes a discrete probability distribution over words. This vector could be very high dimensional depending on the size of the vocabulary. Dimensionality reduction methods are useful for visualization, exploration, and interpretation of such high-dimensional data, as they enable extraction of critical information in the data while discarding noise. Many popular methods are available for visualizing high dimensional data, such as principle component analysis (PCA), MDS, uniform manifold approximation and projection (McInnes et al., 2018, UMAP), and t-distributed stochastic neighbor embedding (van der Maaten and Hinton, 2008, t-SNE). These methods use spectral decompositions of the pairwise distance matrix to embed the data into lower

dimension. PHATE (Moon et al., 2019), on the other hand, is designed to visualize high-dimensional time-varying data. As demonstrated by the authors, it is capable of uncovering hidden low-dimensional embedded temporal progression and branching structure.

Here we embed the estimated LDA word distributions into lower dimensions using a novel application of PHATE to the Hellinger distance matrix. For details on our implementation, see Appendix 4.4. Here, using simulated data, we demonstrate the power of the proposed PHATE-Hellinger embedding for visualization of temporal evolution patterns as compared to other embedding methods. Specifically, we simulate 10 trajectories of 100-dimensional probability vectors using the model

$$X_t^j | X_{t-1}^j \sim \mathcal{N}_{100}(X_{t-1}^j, \sigma_j^2 I)$$

and

$$P_{t,i}^j = \frac{\exp\big(X_{t,i}^j\big)}{\sum_{i=1}^p \exp\big(X_{t,i}^j\big)}, \qquad i = 1, \ldots, 100$$

for $j = 1, \ldots, 10$ and $t = 0, \ldots, 99$. Each trajectory starts at the same point $X_0 \in \mathbb{R}^{100}$ and differs from realization to realization depending on σ_j. We project all 1000 vectors onto a hypersphere by computing the element-wise square root of each probability vector and using the mapping $P_{t,1} + \cdots + P_{t,100} = 1 \Leftrightarrow (\sqrt{P_{t,1}})^2 + \cdots + (\sqrt{P_{t,100}})^2 = 1^2$. Figure V.3 presents the 2D embeddings of this synthetic dataset using PCA on the Euclidean distance matrix, and t-SNE, UMAP, and PHATE on the Hellinger distance matrix. Observe that, among all methods, only PHATE correctly captures the temporal progressions as distinct trajectories originating from a common initial point X_0. Additional simulation studies comparing PCA, t-SNE, UMAP, and PHATE with and without Hellinger distance are included in the Appendix 4.5. In particular, the benefit of using the Hellinger distance instead of the Euclidean distance is demonstrated.

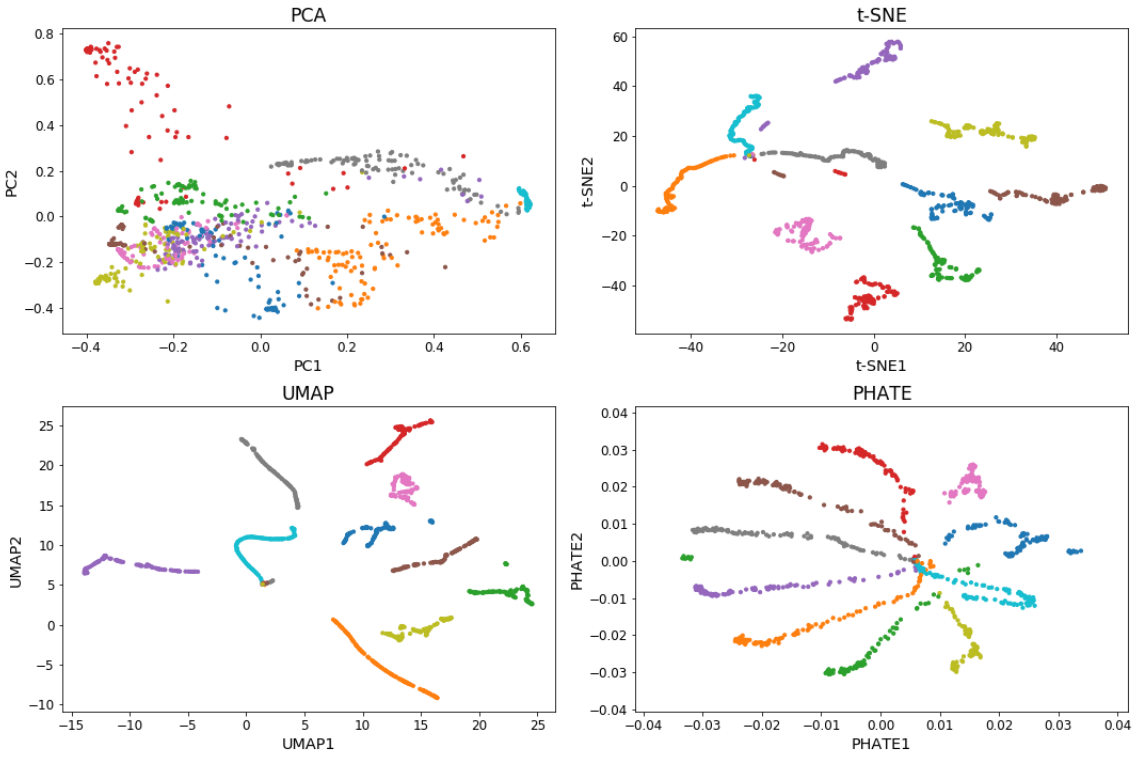


Figure V.3: Comparison of principle component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), uniform manifold approximation and projection (UMAP), and potential of heat-diffusion for affinity-based transition embedding (PHATE) for dimensionality reduction. The methods are applied to 2D embedding of simulated 10 trajectories (identified by color) of 100-dimensional probability vectors, all originating from a common initial point. Except for PCA, all these methods are applied to the matrix of Hellinger distances. Only PHATE correctly captures the temporal progressions as distinct trajectories originating from a common initial point.

5.3 Twitter Data Analysis

The entire pipeline for our analysis is described in Algorithm V.2. The implementation requires setting several hyperparameters. For Twitter data, the temporal smoothing parameter was selected as $\gamma = 0.75$, which corresponds to smoothing approximately one month of tweets into the current time point, in inverse proportion to temporal proximity; the parameter for the number of topics was set to $K = 50$ at every time point; the number of neighbors was set to $k = 10$ for the neighborhood graph to compute shortest paths. In Appendix 4.8 we show relative insensitivity of

our framework to the choice of these hyperparameters. Although not explored here, one could also vary K over time, for example, selected by minimizing perplexities or Bayesian information criteria (BIC) scores at each time (see Appendix 4.8 for a further discussion).

Algorithm V.2: Longitudinal analysis of micro-text data.

Input: Raw micro-text data

1: Preprocess data and organize tweets into a temporally smoothed corpus as described in Section 5.2.2, with smoothing parameter γ.

2: Apply topic models described in Section 5.2.1 independently to each one of T corpus with K topics. This results in TK word distributions.

3: Compute pairwise Hellinger distances for word distributions.

4: Compute:

a: k-nearest neighbor graph with from the TK-by-TK Hellinger matrix and find the shortest path of interest on the neighborhood graph using the Djikstra algorithm.

b: PHATE embedding of high-dimensional word distributions in 2D and 3D.

Output: Shortest paths and PHATE coordinates.

5.3.1 Data preparation

We downloaded data via the Twitter Decahose Stream API (`https://developer.twitter.com/en/docs/Tweets/sample-realtime/overview/decahose`). The Decahose includes a random sample of $\sim 10\%$ of the tweets from each day, resulting in a sample of $300-500$ millions of tweets per day. Among all tweets that are sampled, between $\sim 0.1\%$ and 0.5% (see Appendix 4.7 for details) of them contain geographic location information, called geotags, that localize the tweet to within a neighborhood of the user's location when the tweet was generated. Note that Twitter's precise location service that uses GPS information has been turned off by default (`https://twitter.com/TwitterSupport/status/1141039841993355264`). We consider here the more common Twitter "Place" object that consists of 4 longitude-latitude coordinates that define the general area from which the user is posting the tweet (`https://developer.`

twitter.com/en/docs/tutorials/filtering-Tweets-by-location). Here, we focus on a time period from February 15, 2020, to May 15, 2020, where we expect there to be a large volume of tweets that are COVID-19 related. Figure D.8a in Appendix 4.7 shows the number of tweets for each day in the study period. The following filtering was used:

- **U.S. geographic area**: Tweets that are geotagged and originated in the United States as indicated by the Twitter location service.
- **English language Tweets**: Tweets from users who selected English as their default language.
- **Non-retweets**: Tweets that contain original content from the users and are not a retweet of other tweets.

The following text preprocessing steps were undertaken: 1) we remove stop words (e.g., *in, on, and*, etc., which do not carry semantic meaning); 2) we keep only common forms of words (lemmatization); 3) we remove words that occurred less than 5 times in a document. As a result, the average vocabulary length per timestamp was been reduced from around 300000 to 3000. Further, the union of the unique words from each timestamp has been used as the common vocabulary with word frequencies zeroed out on days where those words do not occur.

5.3.2 Hellinger-PHATE embedding for all topics

Figure V.4 shows the 2D Hellinger-PHATE embedding of 4500 word distributions. We labeled the points on the plots with different colors, sizes, and styles for visualization and interpretation of various time points, tweet volumes, and shortest paths. The full labeling scheme is included in Appendix 4.9. Figure V.4 also shows (as insets) two zoom-ins onto selected COVID-19 topics. We observe several interesting trajectory patterns in the PHATE embeddings. For example, the "STAY HOME

(executive order)" cluster (bottom inset) is organized along a straight line, where the points are more dense at the beginning as well as at the end of the line while sparser in between. The COVID and COVID NEWS clusters (top inset) behave like a splitting between two branches of a tree, and the COVID NEWS (presidential election) path in those clusters exhibits a 'hook' or a 'U' shape. Within the COVID NEWS cluster, the presidential election path also splits and diverges from other points in the same cluster. The following two subsections will focus on these two clusters and paths therein to illustrate the advantages of the proposed framework. Additional visualizations for the SANITIZING (wash hands) and STAY HOME (executive order) paths are included in Appendix 4.10.

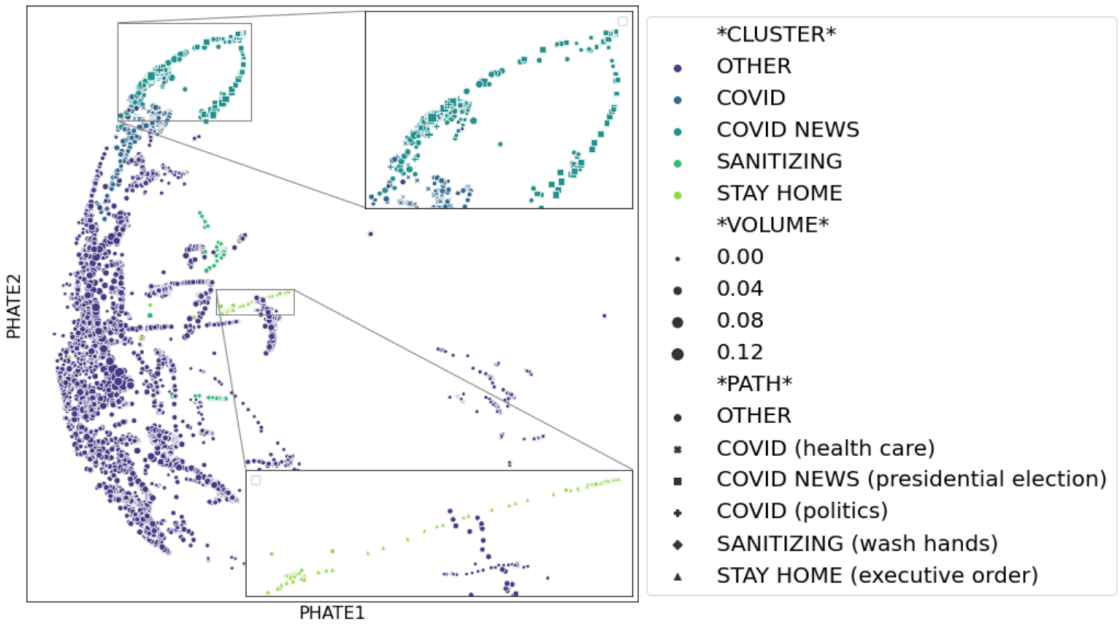


Figure V.4: Potential of heat-diffusion for affinity-based transition embedding (PHATE) for all word distributions. Here the two bounding boxes and insets highlight two of the COVID-19-related topic clusters/paths (COVID/COVID NEWS and STAY HOME). The colors, sizes, and styles signify various clusters, tweet volumes, and shortest paths, as given in the dictionary in Appendix 4.9. Note that the embedding captures some important clustering/trajectory structures, for example, branching, splitting, merging, and so on.

5.3.3 Case study I: presidential election topic path

Here, we focus on a cluster of topics that is implicitly COVID-related but can be well understood from associated real-world events. We call this the presidential election topical path. The subset of topics lying on this shortest path is illustrated in Figure V.5. Here continuous color scales are used to illustrate temporal evolution, which exhibits a smooth transition from the beginning to the end points on the path. The PHATE embedding exhibits three subclusters on the path: 1) an early March cluster that groups topics related to Super Tuesday; 2) an April cluster that groups topics related to or triggered by the "Bernie Sanders dropped out of the presidential race" event; 3) an early to mid-May cluster that groups topics converging to more general COVID-related political topics.

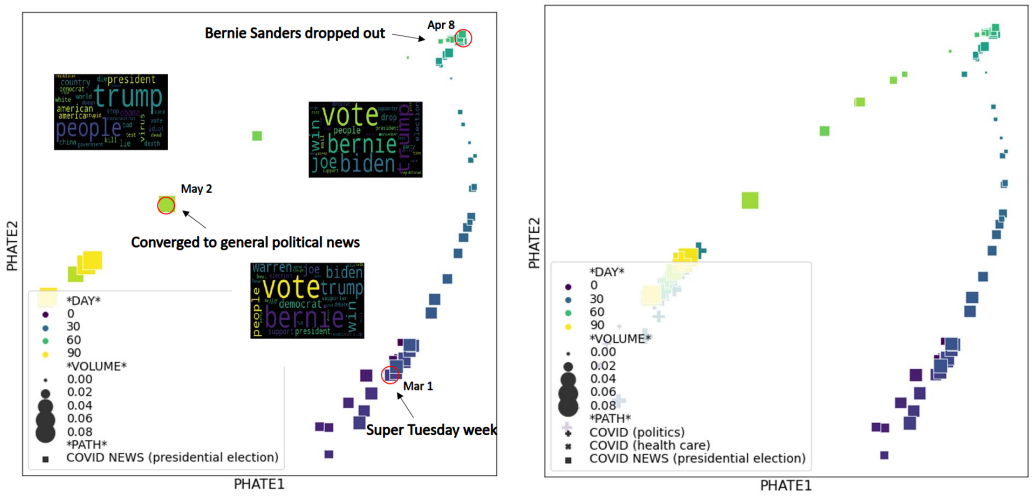


Figure V.5: Potential of heat-diffusion for affinity-based transition embedding (PHATE) for subsets of topics in the COVID NEWS cluster (right) and the presidential election path (left) within the cluster. Colors and sizes highlight time and tweet volumes, respectively. Here three word clouds containing top 30 words in corresponding topics are shown for the time points highlighted by red circles, showing important real-word events that are annotated. Note the plot at the bottom shows (near lower left) the merge and split of different paths (labeled by filled squares, crosses, and pluses) within the same cluster.

Additionally, in terms of tweet volume generated by COVID NEWS topic, there exists again a U-shaped trend: starting at a high level in mid-February the tweet

volumes dropped down after the Super-Tuesday week and started to rise near the time when Bernie Sanders dropped out and eventually peaked in mid-May. We believe that this modulation of the presidential election path can be explained by the COVID-19 pandemic in the United States, which accelerated through March when many states issued stay-at-home orders. This then triggered public discourse around COVID-19, increasing the volume of COVID-19-related topics. However, starting in May, as many stay-at-home orders were lifted, more mainstream political news topics reentered the discourse.

Following we present results of spatial analysis, showing county-level tweet volume in California, illustrating that the Hellinger-distance shortest path combined with PHATE is able to capture more granular-level variations in both space and time. In Figure V.6 we plot smoothed choropleth maps for the same three topics that were highlighted in Figure V.5, where the color changes with respect to tweet proportions (the estimated tweet volumes generated from the given topics normalized by the total tweet volumes for the given days for each county). Here raw tweet proportions have been smoothed using a simple Markov random field (MRF) smoother (Wood, 2017), which regularizes neighboring counties (i.e., regions with contiguous boundaries, that is, sharing one or more boundary point) to have similar tweet proportions. This smoothing procedure is used to identify hot spots, or areas whose tweet volumes have a high likelihood of differing over neighboring locations. The regularization removes some of the variance one would normally see in a choropleth, and gives a bird's eye view of the entire state. For visualization of similar choropleth maps for other states, as well as a comparison of the maps between states, we include interactive maps at `https://wayneyw.shinyapps.io/mrf_smooth_map_app/`.

From the top row of Figure V.6 we observe two 'presidential election' hot spots in counties near the Bay Area and in counties near Los Angeles. The local trend in tweet volume for California is similar to the global trend overall in the United States,

as indicated by Figure V.5 above as well as Figure D.14 in Appendix 4.11.

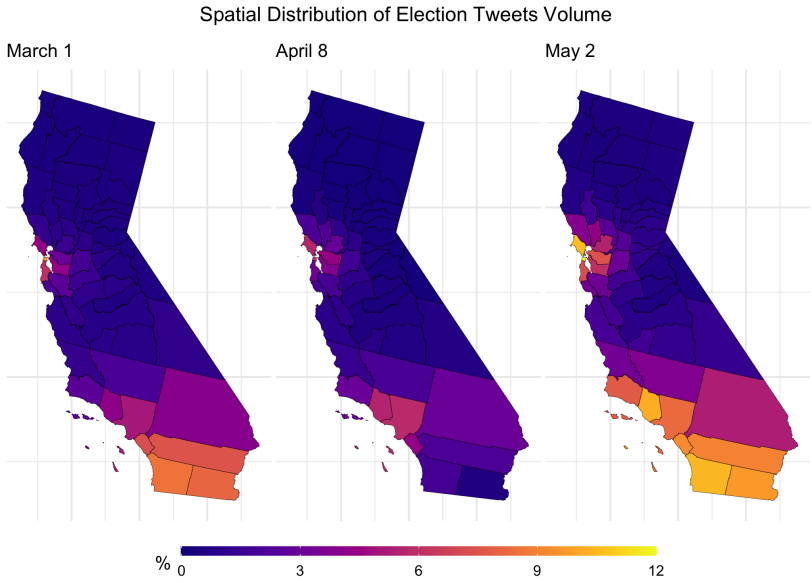


Figure V.6: County-level maps for California. It shows the spatial distribution of proportional tweet volumes for the three time points on the COVID NEWS (presidential election) path.

5.3.4 Case study II: general COVID-19 topic path

In this case study we focus on an explicit COVID-19 topic cluster and shortest paths therein. Figure V.7 shows the PHATE embedding for subsets of topics in the COVID cluster. The embedding identifies two paths that together exhibit splitting behavior, which can be considered as types of structures built into PHATE a priori. In this case, two similar discussions around COVID-19 split into a path that focused on health care, for example, testing, deaths, hospital, and so on, and a path that focused on politics, for example, government, Trump, president, and so on, respectively. The split of the two paths into two different sets of topics is revealed by naive clustering algorithms, such as hierarchical clustering. We emphasize here that such bifurcation behavior would be difficult to model explicitly, for example, using a time-varying global LDA-type model, but appears naturally in the PHATE embedding of the shortest paths using Hellinger distance.

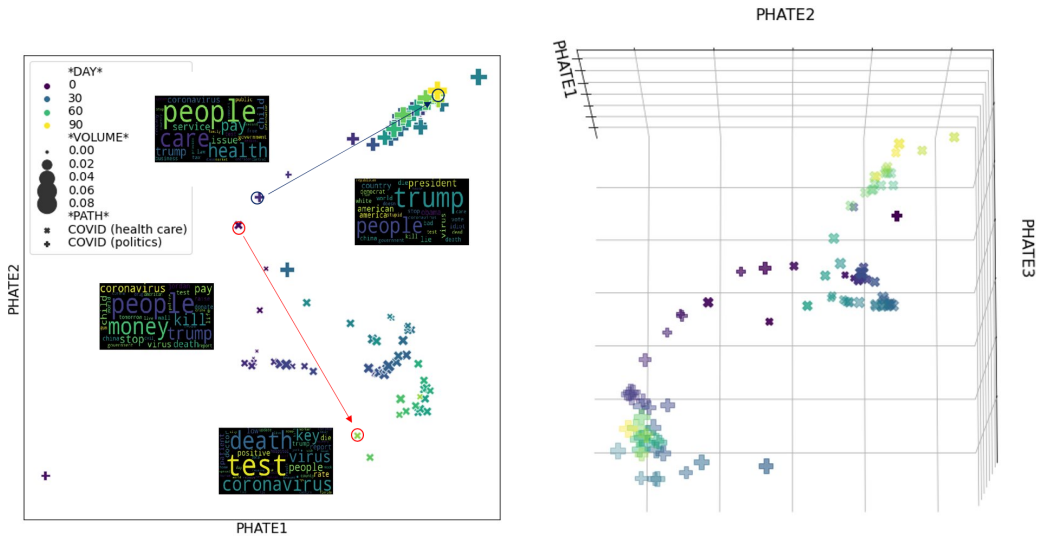


Figure V.7: Potential of heat-diffusion for affinity-based transition embedding (PHATE) for subsets of topics in the COVID cluster. The plots demonstrate a 2D (left) and a 3D (right) embedding of two different paths (i.e., health care and politics). Colors and sizes highlight time and tweet volumes, respectively. Here four word clouds containing top 30 words in corresponding topics are shown for the time points (with arrows connecting the beginning and the end topics on the same path) highlighted by red (health care) and black (politics) circles. Note the plots show divergent behavior of public discourse around COVID-19, where two similar discussions diverge to different discussions (indicated by the word clouds). The 3D embedding illustrates nonlinear paths, that is, spirals and loops, for this topic.

The two separated paths can be more clearly observed in the 3D view, where a 'spiral' structure in the path labeled by filled circles is revealed. This spiral as well as the 'loop' presented in Figure V.5 capture sharp transitions of discussions within a topic path, in contrast to more linear structures such as those exhibited in the SANITIZING (wash hands) and the STAY HOME (executive order) clusters, where the discussion is stable over time. In particular, the health care trajectory transitioned from a discussion on general concerns about the coronavirus to testing-focused discussions on a similar topic; the discussions along the presidential election trajectory transitioned from politicians in the presidential race to more general politics. On the other hand, as illustrated in Appendix 4.10, for more linear 'wash hands' and 'executive order' trajectories, discussions along the paths are quite stable in terms

of the most relevant words. We conjecture, more formally, that linear paths geometrically constitute a one-dimensional subspace over which a single multinomial word distribution propagates over time, unaffected by nearby clusters. This represents stability in the discussions of the topic. Nonlinear paths like spirals, on the other hand, likely constitute a nonlinear subspace where the multinomial word distribution changes smoothly over time, affected by proximity to other clusters.

For county-level spatial analysis, three examples of events can be visualized in Figure V.8 (following the list of relevant events found at `https://en.wikipedia.org/wiki/COVID-19_pandemic_in_California`):

- Spatial distribution of COVID tweet proportions on March 1, where the Bay Area is identified as a relative hot spot in the state. Around late February and early March, counties near the Bay Area were first hit by the coronavirus pandemic. For example, cases were reported in Alameda and Solano Counties on that day; a case was reported in Marin County, who was a passenger on the Grand Princess cruise.

- On March 11, the first death due to coronavirus was reported in LA County, and Ventura County reported their first case on the day before. These 'light up' the two counties on the map as a hot spot.

- On March 20 to March 21, Los Angeles County, which is nationally the second-largest municipal health system, announced that it could no longer contain the virus and changed their guidelines for COVID-19 testing to not test symptomatic patients if a positive result would not change their treatment. Note that the Bay Area hot spot before started to 'fade away' in terms of Tweets volume proportions.

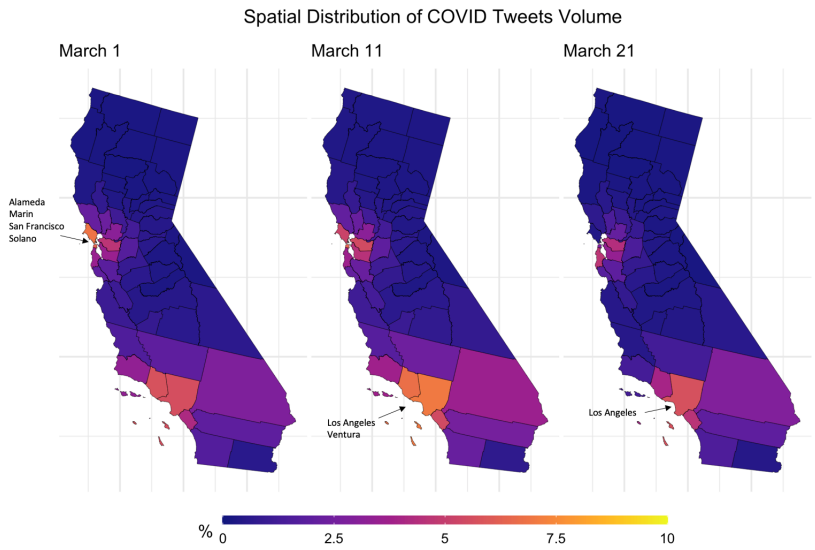


Figure V.8: County-level maps for California. It shows the spatial distribution of proportional tweet volumes for three time points on the COVID (health care) path. Note that counties' names are given for spatial hot spots (in terms of tweet volume).

5.4 TalkLife Data Analysis

For TalkLife data, a similar procedure detailed in Algorithm V.2 is applied with a weakly-supervised T-LDA using seed words. Here, we set $\gamma = 0.5$ which corresponds to one month of posts for the current time point; the parameter for the number of topics was set to $K = 15$ at every time point; the number of neighbors was set to $k = 10$ for the neighborhood graph to compute shortest paths.

In contrast to the Twitter analysis, we do not have access to a time period of high-volume data and a diverse set of real-world events (e.g., COVID-19 pandemic, presidential election) that can be compared against each other. Although this results in a lack of ground truth for our dynamical analysis, we do have access to a set of labels generated by human experts. This motivates us to compare the learned topics with these labels in the following subsections.

5.4.1 Data preparation

We use data provided by the TalkLife platform (`https://www.talklife.com/research`). All posts from the year 2019 are extracted for further processing. Since the volume of daily posts is much lesser compared to that of Twitter data, we combined posts on a weekly basis and consider "week" as our time unit for the following analyses. Volumes are these weekly posts are comparable to the volume of the daily tweets (see Figure D.8b). Furthermore, similar text preprocessing steps as in the Twitter analysis were undertaken, and the union of the unique words from each timestamp has been used as the common vocabulary with word frequencies zeroed out on weeks where those words do not occur.

A notable feature of the TalkLife data is the human-generated labels for the posts. There are 33 labels in total, and each post is tagged by ≥ 0 labels that describe the underlying/suspected mental health issues embedded in the post. Note that there are cases where more than one label is tagged to a post. In Table D.6 of Appendix 4.12 basic information including percentage volume of each label is shown. Additionally, the labels are used for constructing word-level features that are fed into the weakly supervised LDA model. Specifically, top words (measured by percentage volume) from posts associated with the labels are selected as "seed words" that guide the LDA discovery of latent topics. Using the notation from Section 5.2.1, for each seed word $n \in V$, an incremental weight of w_n that is proportional to the volume of the corresponding word has been employed in the Dirichlet prior to the multinomial word distribution. Table D.7 of Appendix 4.12 provides details of the seed word selection and the associated prior weights being used in the weakly-supervised T-LDA model.

5.4.2 Clustering of labels

An issue with 33 labels is that they overlap with each other in terms of the bag-of-words representations of the corresponding posts. For example, the posts with

labels "NauseaWithEatingDisorderSuspected" and "NauseaSuspected" respectively may look similar from the perspective of bag-of-words. As our goal of using the labels is to compare them against the learned topics from an LDA model that is based on the bag-of-words approach, we further cluster the labels into meta-labels in order to reduce noise and redundancy.

In particular, we extract the associated posts for each given label and construct a word distribution where the weights are computed as the percentage volumes. When using a potentially large set of features, one might expect that the true underlying clusters present in the data differ only with respect to a small fraction of the features, and will be missed if one clusters the observations using the full set of features (Witten and Tibshirani, 2010). Here, since we are training a clustering model on a dataset with 33 samples where each sample is a 3623-dimensional vector ($3623 \gg 33$), i.e., a discrete distribution over the vocabulary, we sparsify each feature vector by zeroing out the weights of words that belong to the intersection of the 33 samples. This procedure will de-emphasize words such as "feel", "sad", "upset", etc. that are common to most mental health-related posts in any clustering algorithm. Moreover, as noted by several researchers (Gopal and Yang, 2014; Batmanghelich et al., 2016; Meng et al., 2019), clustering text data with Euclidean metrics such as used by the kmeans algorithm or Gaussian mixture models are not appropriate as the data is usually normalized (e.g., term frequency, word counts) and lies on a unit-sphere manifold induced by the Hellinger metric over distribution pairs. Here, taking into consideration the above-mentioned issues, we employ a von Mises-Fisher (vMF) mixture model (Banerjee et al., 2005; Gopal and Yang, 2014) to cluster the label using their sparse representations of word distributions.

We construct 10 meta-clusters where the number of components is chosen via a Bayesian information criteria (BIC). Figure V.9 visualizes the top words from the merged words distributions. Table D.8 and Figure D.16 of Appendix 4.13 shows the

labels that belong to each of the cluster and a similar top-word visualization for the 33 samples before clustering, respectively.

Figure V.9: Top words from the merged/clustered words distributions.

5.4.3 Learned topics vs. label topics

In this study we focus on comparing the learned latent topics against the label topics (i.e., meta-clusters). At each timestamp, we compute the dot products between the learned topics with each of the label topic, that is, we compute the dot products between the corresponding weights over the vocabulary. Figure V.10 depicts topics that have the largest (i.e., most similar), smaller (i.e., not quite similar), and the smallest (i.e., least similar) dot products compared with a meta label topic that includes posts labeled by "NauseaSuspected" and/or "NauseaWithEatingDisorderSuspected". The topics with smaller dot products are clearly unrelated to eating disorder. This comparison is potentially helpful for discovering novel mental health related discussions/topics that have not yet been assigned a proper label by experts. Specifically, the learned topics that consistently result in small dot products with all label topics may indicate such novel discovery. For example, the last topic in the middle row of Figure V.10 may indicate a topic related to sleeping disorder (e.g., insomnia), which has not been included in the label set.

Furthermore, as one increases the level of supervision by increasing the prior weights on the seed words, we expect to see an increased similarity between the learned topics and a given label topic. This is confirmed in Figure V.11 that compares two

weighting schemes (with different seed words weights) with an unsupervised T-LDA.

Figure V.10: Examples of the most similar (top row), not quite similar (middle row), and the least similar (bottom row) learned topics compared to the label topic under labels "NauseaSuspected" and/or "NauseaWithEatingDisorderSuspected" at various timestamps. The top row clearly resembles the discussion expected from expert knowledge.

5.4.4 Case study: anxiety and suicide topic paths

In this case study we focus on two critical mental health issues and the corresponding latent topic paths learned using our method: anxiety and suicide ideation. Figure V.12 depicts the PHATE embedding for these two topic paths. The embedding identifies two paths that together exhibit converging behavior, which can be considered as types of structures built into PHATE a priori (similar to the splitting structure in case study II of Twitter analysis). Here, two dissimilar discussions on TalkLife around anxiety and suicide, respectively, merge into discussions centered on life-worthlessness and suicidal ideation. The convergence of the two paths is further revealed by clustering of the true labels – Figure D.8 in the appendix shows that the labels "AnxietyPanicFearSuspected" and "SuicidalIdeationAndBehaviorSuspected", as well as other related labels such as "AgitationOrIrritationSuspected", "SelfHarm-

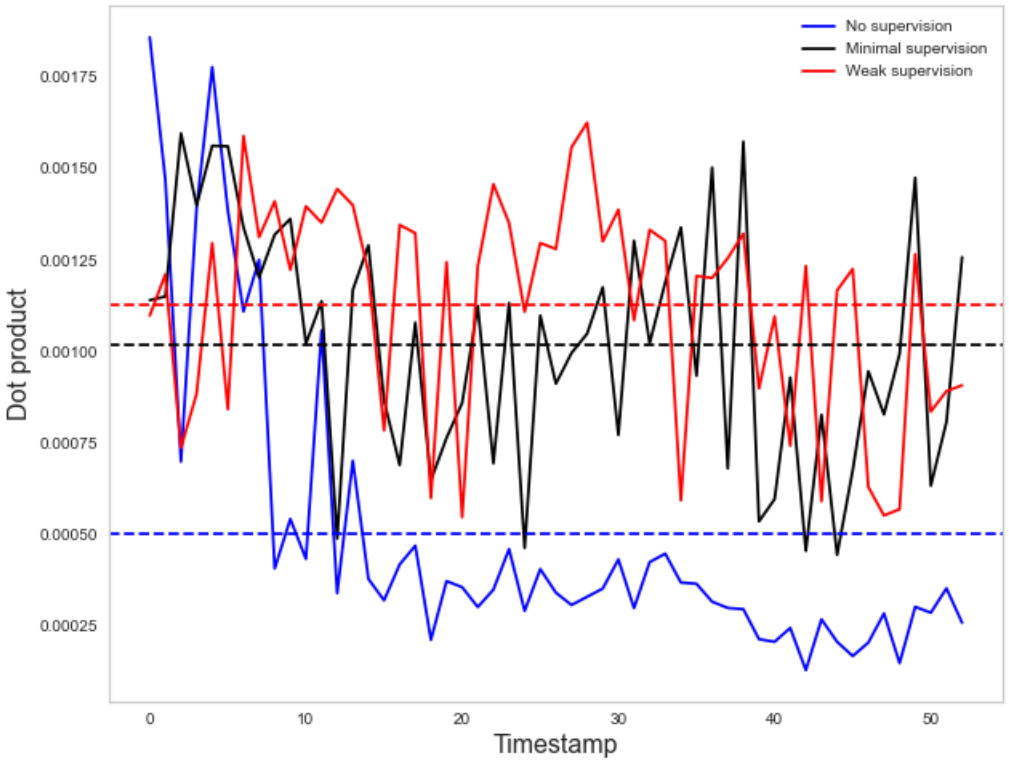


Figure V.11: The dot product scores between the learned topics and the label topic under labels "NauseaSuspected" and/or "NauseaWithEatingDisorderSuspected" across timestamps (52 weeks in 2019), where the horizontal dotted line indicate the average over those timestamps. Here, scores computed from topics learned with no supervision, minimal supervision, and weak supervision are compared – more supervision results in more similar topics compared with the labels.

RelapseSuspected" all belong to the same cluster.

Moreover, in terms of the shape of the embedding, we again observe similar curved and spiral structures that occurred in Figure V.7 and Figure V.5 in the Twitter analysis. For example, we believe the curvy structure appeared on the suicidal path (solid circle) is due to seasonal effect in suicide rates. In particular, a study by the Annenberg Public Policy Center (Rozansky, 2020) found that in 2018 the month with the lowest average daily suicide rate was December with the next-lowest rates in November and January (e.g, winter months). In the same year, the highest rates

were in June, July, and August (summer months). Here, the "inflection point" in the suicidal path occurred around the beginning of summer with an increase in post volumes. The drop in suicidal rate during the winter months is signified by the convergence of the path with the anxiety path. On the anxiety path (solid cross), there is a similar inflection that occurred around early May, which indicates an increase in the suicidal rates and convergence of the path with the suicidal path. From a mental health point of view, a preexisting anxiety issue is a risk factor for the subsequent onset of suicidal ideation and attempts. This is consistent with published analysis (Sareen et al., 2005). Further, another change of direction occurred on the anxiety path towards the winter, which coincides with the decrease in suicidal events. After this second transition, we observe on the PHATE plot that the topics in the anxiety path are diverging from the suicidal path and converging with the earlier topics on anxiety.

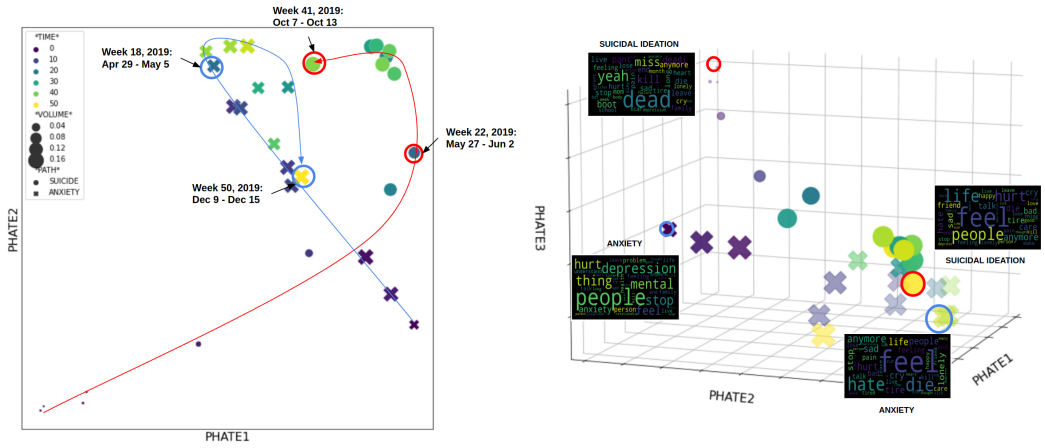


Figure V.12: Potential of heat-diffusion for affinity-based transition embedding (PHATE) for two different topic paths. The plots demonstrate a 2D (left) and a 3D (right) embedding of two different paths – anxiety and suicidal ideation/attempts. Colors and sizes highlight time (52 weeks in 2019) and posts volumes, respectively. Here four word clouds containing top 30 words in corresponding topics are shown for the time points highlighted by red (suicide) and blue (anxiety) circles on the 3D plot. On the 2D plot, arrows are drawn connecting the beginning and the end topics on the same path with circles emphasizing several key time points. Note the plots show convergent behavior of these two temporal topic paths, where two dissimilar discussions converge to similar discussions (indicated by the word clouds). The 3D embedding further confirms this converging behavior.

5.5 Conclusion

We proposed a framework for longitudinal analysis of text data, combining tools from graph algorithms, statistics, and computational geometry. The proposed procedure works by linking together marginal topic-word distributions discovered by a 'regularized' LDA model designed for micro-text, via Hellinger distances and shortest paths on neighborhood graphs. The resulting chain of topics can then be visualized by PHATE dimensionality reduction, which preserves the progressive nature of the input data. With this framework, we discovered and interpreted how certain conversations split and merged under the impact of the COVID-19 pandemic, which can be validated by associating with real-world events. Granular-level spatial analyses showed that our framework is able to capture both global (in the United States) and local variations of COVID-19-related discussions. We further extended the framework to incorporate side information via weak supervision in the form of seed words. With TalkLife data, this extension has been shown to be able to capture latent topics that coincide with expert knowledge. Finally, we believe that social media data could be used to supplement traditional health care or census data to provide fresh insights into the impact of events, like the pandemic, on society, as well as to aid study of mental health issues.

5.5.1 Limitations

There are several limitations of our analysis that deserve additional attention. First, as with most statistical algorithms, there are user tuning parameters that must be selected. There are three tuning parameters that the user must provide: 1) the numbers of nearest neighbors k in the k nearest neighbor graph; 2) the data smoothing parameter γ; and 3) the number of topics K for the T-LDA algorithm. We have shown that our results are robust to perturbations about the parameters we chose, but there may be better choices. These include comprehensive cross-validation methods which,

with sufficient computational resources, can be used to reliably select parameters that minimize a loss function. Such methods have been proposed for selecting k. For selection of K, a promising option is the hierarchical Dirichlet process (HDP), a nonparametric Bayesian model for the number of topics that could vary over time and model birth and death of topics (Teh et al., 2006). Our wrapper framework could easily incorporate an HDP in place of the T-LDA model, but at the expense of increased computation. Two more challenging limitations are those of selection bias and model bias.

Selection biases The use of Twitter data for studying public discourse may be subject to selection bias as users of Twitter may not be representative of the U.S. population. Additionally, users of Twitter may be engaged in different types of public discourses around COVID-19 than users of other social media platforms, for example, Facebook and Reddit, which have different user demographics and privacy policies. Different types of subsampling of Tweets may create their own biases. For example, subsampling based on retweet status, geotag information, country, and time range (e.g., Feb 15 to May 15) are all subject to selection biases. Our subsampling procedure may leave out some important information. For example, we did not consider any retweets, which may contain information on how popular a particular topic might be. Retweets could possibly shed light on a particular topic, which can be measured, for example, by the longitudinal distribution of retweet frequencies for the topic. However, we could not perform a retweet analysis on our geotagged tweets since Twitter does not allow retweets to be geotagged. We also leave out tweets that are generated from U.S. users who are outside of the United States.

Model biases The LDA algorithm we have applied to topic modeling summarizes unstructured texts by themes or topics using a *bag-of-words* approach. This particular approach is computationally scalable but it ignores the relative order of words.

For example, a topic about 'vaccines are not available' can be very close to a topic on 'vaccines are available, but not to me.' The issue may be alleviated by using more sophisticated representations, for example, bigrams or latent semantic analysis. This would result in higher computational burden–the length of unique phrases would increase exponentially as the word order dimension. Other approaches that attempt to model the semantic meaning of topics, such as deep neural networks, could also be used. Additionally, our construction of the smoothed corpora assumes temporal similarities between tweets generated at adjacent time points. Similar types of smoothing assumptions are common in other areas of spatiotemporal statistics as described in Section 5.2.2. The manifold hypothesis is also essential in our model for the shortest path algorithms to recover the intrinsic similarities between topics over time.

CHAPTER VI

Conclusion and Future Work

6.1 Summary

This thesis focused on statistical methods concerning data with spatio-temporal structure. In particular, it touched on two separated research areas:

- Tensor-variate Gaussian graphical models and its connections to and applications in spatio-temporal physical processes
- Temporal topic modeling in both unsupervised and weakly-supervised settings with applications to analyses of public opinions and mental health

These areas are connected in that in each of them there exists structured and graphical representations of the data, and it is imperative to utilize these interpretable representations to achieve scalable and valid inference. This dissertation advanced the state-of-the-art by introducing a new class of Gaussian graphical model for tensor-valued data, designing new estimation algorithms with fast convergence, introducing a new framework for filtering and data assimilation, and finally describing a new approach to temporal topic modeling. There are multiple fruitful extensions of these methodology that warrant further investigation, which we discuss next.

6.2 Future Work

Physical interpretability. While the Kronecker products expansion used in Kronecker PCA captures dense structures in the covariance matrix of data generated from more complex spatio-temporal physical processes as illustrated in Chapter IV, it lacks physical interpretability. In contrast to the case of Sylvester graphical model and Poisson-AR(1) processes, it is not obvious whether the sum of Kronecker products structure corresponds to any true physical models. Recent development in quantum informatics (Chu and Lin, 2021) has demonstrated a link between estimation of the density matrix for entangled quantum states and the structured tensor approximation via $\sum_{i=1}^{l} \mathbf{A}_i \otimes \mathbf{B}_i$. Further characterizing and extending these connections to other classes of discretized PDEs is an interesting future direction. Furthermore, in both the blocked Poisson-AR(1) and convection-diffusion examples, a mixed Kronecker sum and Kronecker product structure emerges that can be related to the state inverse covariance of a dynamical system.

Heavy-tailed multiway covariance/precision models. Most existing work on multiway covariance and inverse covariance models focus on modeling Gaussian variables. It would be interesting to explore whether the pseudo-likelihood framework we adopted for SyGlasso and SG-PALM can be extended to non-Gaussian heavy-tailed models, e.g., using copula's or elliptically contoured distributions. This could have important practical applications, in particular to solar flare and active region prediction problems presented in Chapter III. The images that characterize the active regions generally include a small number of pixels of extreme high-intensity. These pixels might not be captured by a Gaussian-like distribution. Recently, there have been advances (Wei and Minsker, 2017; Ke et al., 2019) in covariance estimation for heavy-tailed, non-Gaussian vector-variate data. Multiway (inverse) covariance estimation is an open problem. Furthermore, robust Kronecker structured covariance / correlation

models where robust estimator of the correlation matrix with sparse Kronecker structure has been recently studied for high-dimensional matrix-variate data (Niu et al., 2020). But, there are still open problems such as theoretical guarantees (comparable to those of traditional methods) and efficient computational algorithms that warrant future development.

Kronecker-structured autoencoders. Low-rank covariance models have close connections with variational autoencoders (VAEs). Dai et al. (2018) studied the relationship between (robust) PCA and VAEs. Since the Kronecker product for matrices is a generalization of the outer product for vectors, KPCA can be considered as a generalization of a the low-rank approximation method of PCA. It is thus natural to exploit similar relationships between KPCA and VAEs. In this case VAE may be considered as a nonlinear/non-Gaussian extension to KPCA for low separation rank covariance models. Additionally, recent advances in efficient training of the VAE-type neural network architecture (e.g., using stochastic gradient descent) could improve the computational complexity of KPCA that is currently limited by an expensive singular value decomposition (Tsiligkaridis and Hero, 2013; Greenewald and Hero, 2015).

Model selection for Kronecker-structured models. Each of the KP, KS, or Sylvester structure has its pros and cons and is appealing only under appropriate data generating processes. It is still unclear, for a given data problem with unknown underlying generative process, how to choose among various Kronecker-structured models. This problem has been attracting attentions only very recently – Guggenberger et al. (2022) developed a procedure for testing for a covariance matrix to have Kronecker product structure. However, the method proposed relies on an expensive rank test procedure (Kleibergen and Paap, 2006) that is not scalable to modern big-data applications. Moreover, it is still an open problem to develop similar tests for Kronecker sum and Sylvester structures in either the covariance or its inverse.

Theoretical analysis of geometry-driven dynamic topic models. The "non-parametric" geometry-driven framework proposed in Chapter V shows promising results in recovering perceptually natural temporal dynamics that may exist among data. We demonstrated the "closeness" of the recovered chain of topics to a series of real events happened around the same time period. However, from a theoretical point of view, it is desirable to understand whether the estimated topic chain approximates well the truth. Just as statisticians have studied when least-squares regression can estimate the "true" regression model, it is natural and important for us to study the ability of the computational geometric algorithms to estimate the "true" topic path in a stochastic topic model.

Researchers have explored the performance of nonparametric algorithms that are based on heuristics or insights on the underlying problems under certain statistical/stochastic models. For example, Rohe et al. (2011) showed the consistency of the spectral clustering algorithms in identifying clusters in network data under a true network generated from the Stochastic Blockmodel (Holland et al., 1983). Bickel and Chen (2009) proved that, also under the Stochastic Blockmodel, a nonparametric community detection algorithm called the Newman–Girvan modularity (Newman and Girvan, 2004) are asymptotically consistent estimators of block partitions.

Akin to these work of studying the performance of nonparametric methods on parametric tasks of estimating quantities in statistical models, we propose to study the consistency of the geometry-driven topic modeling algorithm in identifying the true topic path, under topics generated by the DTM model proposed in Blei and Lafferty (2006). More specifically, under DTM, the generative process at a time stamp t is

1. Draw $\beta_{t,k}|\beta_{t-1,k} \sim \mathcal{N}(\beta_{t-1,k}, \sigma^2\mathbf{I}), \forall k$

2. Draw $\alpha_t|\alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2\mathbf{I})$

3. For each document:

   (a) Draw $\eta_{t,d} \sim \mathcal{N}(\alpha_t, a^2\mathbf{I})$

   (b) For each word:

      i. Draw topic $Z_{t,d,n} \sim \text{Multi}(\pi(\eta_{t,d}))$

      ii. Draw $W_{t,d,n} \sim \text{Multi}(\pi(\beta_{t,Z_{t,d,n}}))$,

where $\pi(x)$ is a mapping from the natural parameterization x to the mean parameterization. Here, define

α_t as the per-document topic distribution at time t.

$\beta_{t,j}$ as the word distribution of topic k at time t.

$\eta_{t,d}$ as the topic distribution for document d at time t.

$z_{t,d,n}$ as the topic for the nth word in document d in time t.

$w_{t,d,n}$ as the word.

A plausible direction in proving performance of the nonparametric topic modeling approach proposed in Chapter V under DTM is to characterize the distances between the recovered topic paths to the true paths, i.e., $\beta_{t,k}$'s, for $t = 1, \ldots$ and show that these distances vanish as both the length of the documents and the number of latent topics grow to infinity.

APPENDIX A

Appendix of Chapter II

In this Appendix,

Section 1.1 provides the detailed derivation of the updates for Algorithm II.1;

Section 1.2 provides the proofs of theorems stated in Section 2.3;

Section 1.3 provides details on the simulated data in Section 2.4.

1.1 Derivation of the Nodewise Tensor Lasso Estimator

1.1.1 Off-Diagonal updates

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

$$(T_{i_k j_k}(\boldsymbol{\Psi}_k))^{\text{off}}_{i_k j_k} = \frac{S_{\frac{\lambda_k}{N}}\left(F_{\boldsymbol{\mathcal{X}},\{\boldsymbol{\Psi}_k\}_{k=1}^K}\right)}{(\frac{1}{N}\boldsymbol{\mathcal{X}}_{(k)}\boldsymbol{\mathcal{X}}_{(k)}^T)_{i_k i_k} + (\frac{1}{N}\boldsymbol{\mathcal{X}}_{(k)}\boldsymbol{\mathcal{X}}_{(k)}^T)_{j_k j_k}}, \tag{1.1}$$

where

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

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

1.1.2 Diagonal updates

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

$$(T(\boldsymbol{\mathcal{W}}))_{i_{[1:K]}} = \frac{-\Big(\boldsymbol{\mathcal{X}}_{(N)}^T\boldsymbol{\mathcal{Y}}_{(N)}\Big)_{i_{[1:K]}} + \sqrt{\Big(\boldsymbol{\mathcal{X}}_{(N)}^T\boldsymbol{\mathcal{Y}}_{(N)}\Big)_{i_{[1:K]}}^2 + 4\Big(\boldsymbol{\mathcal{X}}_{(N)}\boldsymbol{\mathcal{X}}_{(N)}^T\Big)_{i_{[1:K]}}}}{2\Big(\boldsymbol{\mathcal{X}}_{(N)}\boldsymbol{\mathcal{X}}_{(N)}^T\Big)_{i_{[1:K]}}}. \tag{1.2}$$

Here we define $\boldsymbol{\mathcal{Y}} := \sum_{k=1}^K \Big(\boldsymbol{\mathcal{X}} \times_k \boldsymbol{\Psi}_k^{\text{off}}\Big)$. Equations (1.1) and (1.2) give necessary ingredients for designing a coordinate descent approach to minimizing the objective function in (3.3). The optimization procedure is summarized in Algorithm II.1.

1.1.3 Derivation of updates

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

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

where

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

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

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

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 (1.1) then follows.

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

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

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 $\text{diag}\Big(\boldsymbol{\mathcal{X}}_{(N)}^T \boldsymbol{\mathcal{X}}_{(N)} \Big) = \Big(\Big(\boldsymbol{\mathcal{X}}_{(N)}^T \boldsymbol{\mathcal{X}}_{(N)} \Big)_{i_{[1:K]}}, \forall i_{[1:K]} \Big)$.

1.2 Proofs of Main Theorems

We first list some properties of the loss function.

Lemma 1.2.1. *The following is true for the loss function:*

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

$$\Lambda_{\min}^L \leq \lambda_{\min}(\bar{L}''_{\mathcal{S}_k, \mathcal{S}_k}(\bar{\boldsymbol{\beta}})) \leq \lambda_{\max}(\bar{L}''_{\mathcal{S}_k, \mathcal{S}_k}(\bar{\boldsymbol{\beta}})) \leq \Lambda_{\max}^L$$

(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* $1 \leq i_k < j_k \leq m_k$

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

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

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

where $\mathcal{A}_k^{ij} := \mathcal{A}_k / \{(i,j)\}$.

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

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

proof of Lemma A.2.1. 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 (2.5). Then $\bar{L}''_{\mathcal{S}_k,\mathcal{S}_k}(\bar{\boldsymbol{\beta}})$ is equivalent

to

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

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

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

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

□

Lemma 1.2.2. *Suppose conditions (A1-A2) hold, then for any $\eta > 0$, there exist constant $c_{0,\eta}, c_{1,\eta}, c_{2,\eta}, c_{3,\eta}$, such that for any $u \in \mathbb{R}^{q_k}$ the following events hold with probability at least $1 - O(\exp(-\eta \log p))$ for sufficiently large N:*

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

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

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

(iv) $|L''_{N,\mathcal{A}_k\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}})u - \bar{L}''_{\mathcal{A}_k\mathcal{A}_k}(\bar{\boldsymbol{\beta}})u| \leq c_{3,\eta}\|u\|_2^2 q_k\sqrt{\frac{\log p}{N}}$

proof of Lemma A.2.2. (i) By Cauchy-Schwartz inequality,

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

Then note that

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

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

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

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,

$$\begin{aligned}
&P(\|L'_{N,\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}})\|_2 \leq c_{0,\eta}\sqrt{q_k\frac{\log p}{N}}) \\
&\geq P(\sqrt{q_k}\max_{i\in\mathcal{A}_k}|L'_{N,i}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}})| \leq c_{0,\eta}\sqrt{q_k\frac{\log p}{N}}) \geq 1 - O(\exp(-\eta\log p)).
\end{aligned}$$

(iii) By Cauchy-Schwartz,

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

Then by Bernstein inequality,

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

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

Lemma A.2.3. and A.2.4. are used later to prove Theorem 1.

Lemma 1.2.3. *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* (2.8) *within the disc:*

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

with probability at least $1 - O(\exp(-\eta \log p))$ *for sufficiently large* N*.*

proof of Lemma A.2.3. Let $\alpha_N = \max_k \sqrt{q_k}\lambda_{N,k}$. 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 $u^k_{\mathcal{A}_k^c} = 0$, $\|u^k\|_2 = C_k$, 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

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

Thus,

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

Next,

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

Here the first equality is due to the second order expansion of the loss function and the inequality is due to Lemma A.2.2 For sufficiently large N, by assumption that $\lambda_{N,k}\sqrt{N/\log p} \to \infty$ if $m_k \to \infty$ and $\sqrt{\log p/N} = o(1)$, the second term in the last line above is $o(\alpha_N \sqrt{q_k}\lambda_{N,k}) = o(\alpha_N^2)$; the last term is $o(\alpha_N^2)$. Therefore, for sufficiently large N

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

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

By Lemma A.2.1., $(u_{\mathcal{A}_k}^k)^T \bar{L}''_{N,\mathcal{A}_k\mathcal{A}_k}(\bar{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}}) u_{\mathcal{A}_k}^k \geq \Lambda_{\min}^L \|u_{\mathcal{A}_k}^k\|_2^2 = \Lambda_{\min}^L (C_k)^2$, for each

k. 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

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

with probability at least $1-O(\exp(-\eta\log p))$, which means any solution to the problem defined in (2.8) is within the disc $\{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}\|_2 \le \alpha_N\|u\|_2 \le \alpha_N\sqrt{K}\max_k C_k\}$ with probability at least $1-O(\exp(-\eta\log p))$.

□

Lemma 1.2.4. *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 $1-O(\exp(-\eta\log p))$: if for any $\boldsymbol{\beta}\in S=\{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}\|_2 \ge C_2(\bar{\boldsymbol{\beta}})\sqrt{K}\max_k\sqrt{q_k}\lambda_{N,k}, \boldsymbol{\beta}_{\mathcal{A}_N^c}=0\}$, then $\|L'_{N,\mathcal{A}_N}(\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}},\boldsymbol{\mathcal{X}})\|_2 > \sqrt{K}\max_k\sqrt{q_k}\lambda_{N,k}$.*

proof of Lemma A.2.4. Let $\alpha_N=\max_k\sqrt{q_k}\lambda_{N,k}$. For $\boldsymbol{\beta}\in S$, we have $\boldsymbol{\beta}=\bar{\boldsymbol{\beta}}+\alpha_N u$, with $u_{(\mathcal{A})^c}$ and $\|u\|_2\ge C_2(\bar{\boldsymbol{\beta}})$. Note that by Taylor expansion of $L'_{N,\mathcal{A}}(\bar{\boldsymbol{\mathcal{W}}},\boldsymbol{\beta},\boldsymbol{\mathcal{X}})$ at $\bar{\boldsymbol{\beta}}$

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

By triangle inequality and similar proof strategies as in Lemma A.2.3., for sufficiently large N

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

with probability at least $1 - O(\exp(-\eta \log p))$. By Lemma A.2.1., $\|\bar{L}''_{N,\mathcal{A}\mathcal{A}}(\bar{\boldsymbol{\beta}})u_{\mathcal{A}}\|_2 \geq \Lambda^L_{\min}(\bar{\boldsymbol{\beta}})\|u_{\mathcal{A}}\|_2$. 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 (2.8), it satisfies $\|L'_{N,\mathcal{A}_k}(\boldsymbol{\mathcal{W}}, \hat{\boldsymbol{\beta}}, \boldsymbol{\mathcal{X}})\|_\infty \leq \lambda_{N,k}$. Thus,

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

Then by Lemmas A.2.4., for any $\eta > 0$, for N sufficiently large, all solutions of (2.8) are inside the disc $\{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}\|_2 \leq C_2(\bar{\boldsymbol{\beta}}) \max_k \sqrt{q_k}\lambda_{N,k}, \boldsymbol{\beta}_{\mathcal{A}_N^c} = 0\}$ with probability at least $1 - O(\exp(-\eta \log p))$. If we further assume that $\min_{(i,j)\in\mathcal{A}_k} |\bar{\boldsymbol{\beta}}_{i,j}| \geq 2C(\bar{\boldsymbol{\beta}}) \max_k \sqrt{q_k}\lambda_{N,k}$ for each k, then

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

□

proof of Theorem 2. Let $\mathcal{E}_{N,k} = \{\mathrm{sign}(\hat{\boldsymbol{\beta}}^{\mathcal{A}_k}_{i_k j_k}) = \mathrm{sign}(\bar{\boldsymbol{\beta}}^{\mathcal{A}_k}_{i_k j_k})\}$. Then by Theorem 1, $P_{\bar{\boldsymbol{\mathcal{W}}},\bar{\boldsymbol{\beta}}}(\mathcal{E}_{N,k}) \geq 1 - O(\exp(-\eta \log p))$ for large N. On $\mathcal{E}_{N,k}$, by the KKT condition and

the Taylor's expansion of $L'_{N,\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}}, \hat{\boldsymbol{\beta}}^{\mathcal{A}_k}, \boldsymbol{\mathcal{X}})$ at $\bar{\boldsymbol{\beta}}^{\mathcal{A}_k}$

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

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

$$
\begin{aligned}
&v_{N,k} = \\
&-\lambda_{N,k}[\bar{L}''_{\mathcal{A}_k\mathcal{A}_k}]^{-1}\text{sign}(\bar{\boldsymbol{\beta}}^{\mathcal{A}_k}) - [\bar{L}''_{\mathcal{A}_k\mathcal{A}_k}]^{-1}[L'_{N,\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}^{\mathcal{A}_k}, \boldsymbol{\mathcal{X}}) + D_{N,\mathcal{A}_k\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}^{\mathcal{A}_k})v_{N,k}],
\end{aligned} \tag{1.3}
$$

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

$$
L'_{N,ij}(\bar{\boldsymbol{\mathcal{W}}}, \hat{\boldsymbol{\beta}}^{\mathcal{A}_k}, \boldsymbol{\mathcal{X}}) = L'_{N,ij}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}^{\mathcal{A}_k}, \boldsymbol{\mathcal{X}}) + L''_{N,ij,\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}^{\mathcal{A}_k}, \boldsymbol{\mathcal{X}})v_{N,k}. \tag{1.4}
$$

Then combining (1.3) and (1.4) we get

$$
\begin{aligned}
&L'_{N,ij}(\bar{\boldsymbol{\mathcal{W}}}, \hat{\boldsymbol{\beta}}^{\mathcal{A}_k}, \boldsymbol{\mathcal{X}}) \\
&= -\lambda_{N,k}\bar{L}''_{ij,\mathcal{A}_k}(\bar{\boldsymbol{\beta}}^{\mathcal{A}_k})[\bar{L}''_{\mathcal{A}_k\mathcal{A}_k}]^{-1}\text{sign}(\bar{\boldsymbol{\beta}}^{\mathcal{A}_k}) - \bar{L}''_{ij,\mathcal{A}_k}(\bar{\boldsymbol{\beta}}^{\mathcal{A}_k})[\bar{L}''_{\mathcal{A}_k\mathcal{A}_k}]^{-1}L'_{N,\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}^{\mathcal{A}_k}, \boldsymbol{\mathcal{X}}) \\
&+ [D_{N,ij,\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}^{\mathcal{A}_k}) - \bar{L}''_{ij,\mathcal{A}_k}(\bar{\boldsymbol{\beta}}^{\mathcal{A}_k})[\bar{L}''_{\mathcal{A}_k\mathcal{A}_k}]^{-1}D_{N,\mathcal{A}_k\mathcal{A}_k}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}^{\mathcal{A}_k})]v_{N,k} \\
&+ L'_{N,ij}(\bar{\boldsymbol{\mathcal{W}}}, \bar{\boldsymbol{\beta}}^{\mathcal{A}_k}, \boldsymbol{\mathcal{X}}).
\end{aligned} \tag{1.5}
$$

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

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

Thus, following straightforwardly (with the modification that we are considering each $\mathcal{A}_k$ instead of $\mathcal{A}$) from the proofs of Theorem 2 of Peng et al. (2009), the remaining terms in (1.5) can be shown to be all $o(\lambda_{N,k})$, and $\max_{(i,j)\in\mathcal{A}_k^c} |L'_{N,ij}(\bar{\boldsymbol{\mathcal{W}}}, \hat{\boldsymbol{\beta}}^{\mathcal{A}_k}, \boldsymbol{\mathcal{X}})| < \lambda_{N,k}$ with probability at least $1 - O(\exp(-\eta \log p))$ 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 $\mathcal{A}_k$ and hence, no wrong edge will be included in $\mathcal{A} = \cup_k \mathcal{A}_k$. □

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

1.3 Simulated Precision Matrix

1. **AR1(ρ)**: 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 $(\boldsymbol{\Psi}_k)_{ij} = 1$. 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 $\mathbf{A}_{ii} \to \mathbf{A}_{ii} + \psi$, $\mathbf{A}_{jj} \to \mathbf{A}_{jj} + \psi$.

APPENDIX B

Appendix of Chapter III

In this Appendix,

Section 2.1 provides detailed derivation of the log-pseudolikelihood function.

Section 2.2 provides justifications for the Barzilai-Borwein step sizes implemented in Algorithm III.1.

Section 2.3 provides detailed proofs of Theorems 3.4.1 and 3.4.2.

Section 2.4 discusses extensions of Algorithm III.1 and its convergence properties to non-convex cases.

Section 2.5 provides additional details of the solar flare experiments.

2.1 Derivation of the Log-Pseudolikelihood

By rewriting the Sylvester tensor equation defined in (3.2) element-wise, we first observe that

$$\begin{aligned} &\left(\sum_{k=1}^{K}(\boldsymbol{\Psi}_k)_{i_k,i_k}\right)\boldsymbol{\mathcal{X}}_{i_{[1:K]}} \\ &= -\sum_{k=1}^{K}\sum_{j_k\neq i_k}(\boldsymbol{\Psi}_k)_{i_k,j_k}\boldsymbol{\mathcal{X}}_{i_{[1:k]},j_k,i_{[k+1:K]}} + \boldsymbol{\mathcal{T}}_{i_{[1:K]}}. \end{aligned} \tag{2.1}$$

Note that the left-hand side of (2.1) involves only the summation of the diagonals of the $\mathbf{\Psi}_k$'s and the right-hand side is composed of columns of $\mathbf{\Psi}_k$'s that exclude the diagonal terms. Equation (2.1) 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 $\mathbf{\Psi}_k$'s act as regression coefficients. The formulation in (2.1) naturally leads to a pseudolikelihood-based estimation procedure (Besag, 1977) for estimating $\mathbf{\Omega}$ (see also Khare et al. (2015) for how this procedure applied to vector-variate Gaussian graphical model estimation). It is known that inference using pseudo-likelihood is consistent and enjoys the same $\sqrt{N}$ convergence rate as the MLE in general (Varin et al., 2011). This procedure can also be more robust to model misspecification (e.g., non-Gaussianity) in the sense that it assumes *only that the sub-models/conditional distributions (i.e., $\boldsymbol{\mathcal{X}}_i|\boldsymbol{\mathcal{X}}_{-i}$) are Gaussian.* Therefore, in practice, even if the data is not Gaussian, the Maximum Pseudolikelihood Estimation procedure is able to perform reasonably well. Wang et al. (2020c) also studied a different model misspecification scenario where the Kronecker product/sum and Sylvester structures are mismatched for SyGlasso.

From (2.1) we can define the sparse least-squares estimators for $\mathbf{\Psi}_k$'s as the solution of the following convex optimization problem:

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

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

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

with $\boldsymbol{\mathcal{W}}_{i_{[1:K]}} := \sum_{k=1}^{K} (\boldsymbol{\Psi}_k)_{i_k,i_k}$.

The optimization problem above can be put into the following matrix form:

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

where $\mathbf{S} \in \mathbb{R}^{d \times d}$ is the sample covariance matrix, i.e., $\mathbf{S} = \frac{1}{N} \sum_{i=1}^{N} \operatorname{vec}(\boldsymbol{\mathcal{X}}^i) \operatorname{vec}(\boldsymbol{\mathcal{X}}^i)^T$. Note that this is equivalent to the negative log-pseudolikelihood function that approximates the ℓ_1-penalized Gaussian negative log-likelihood in the log-determinant term by including only the Kronecker sum of the diagonal matrices instead of the Kronecker sum of the full matrices.

2.2 The Barzilai-Borwein Step Size

The BB method has been proven to be very successful in solving nonlinear optimization problems. In this section we outline the key ideas behind the BB method, which is motivated by quasi-Newton methods. Suppose we want to solve the unconstrained minimization problem

$$\min_x f(x),$$

where f is differentiable. A typical iteration of quasi-Newton methods for solving this problem is

$$x_{t+1} = x_t - B_t^{-1} \nabla f(x_t),$$

where B_t is an approximation of the Hessian matrix of f at the current iterate x_t. Here, B_t must satisfy the so-called secant equation: $B_t s_t = y_t$, where $s_t = x_t - x_{t-1}$ and $y_t = \nabla f(x_t) - \nabla f(x_{t-1})$ for $t \geq 1$. It is noted that in to get B_t^{-1} one needs to solve a linear system, which may be computationally expensive when B_t is large and

dense.

One way to alleviate this burden is to use the BB method, which replaces B_t by a scalar matrix $(1/\eta_t)\mathbf{I}$. However, it is hard to choose a scalar η_t such that the secant equation holds with $B_t = (1/\eta_t)\mathbf{I}$. Instead, one can find η_t such that the residual of the secant equation, i.e., $\|(1/\eta_t)s_t - y_t\|_2^2$, is minimized, which leads to the following choice of η_t:

$$\eta_t = \frac{\|s_t\|_2^2}{s_t^T y_t}.$$

Therefore, a typical iteration of the BB method for solving the original problem is

$$x_{t+1} = x_t - \eta_t \nabla f(x_t),$$

where η_t is computed via the previous formula.

For convergence analysis, generalizations and variants of the BB method, we refer the interested readers to Raydan (1993, 1997); Dai and Liao (2002); Fletcher (2005) and references therein. BB method has been successfully applied for solving problems arising from emerging applications, such as compressed sensing (Wright et al., 2009), sparse reconstruction (Wen et al., 2010) and image processing (Wang and Ma, 2007).

2.3 Proofs of Theorems

2.3.1 Proof of Theorem 3.4.1

We first state the regularity conditions needed for establishing convergence of the SG-PALM estimators $\{\hat{\mathbf{\Psi}}_k\}_{k=1}^K$ to their true value $\{\bar{\mathbf{\Psi}}_k\}_{k=1}^K$.

(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 $\mathbb{R}^d$, i.e., there exist a constant $c > 0$, such that for every $\mathbf{a} \in \mathbb{R}^d$, $\mathbb{E}e^{\mathbf{a}^T x} \leq e^{c\mathbf{a}^T \bar{\mathbf{\Sigma}} \mathbf{a}}$, and there exist $\rho_j > 0$ such that $\mathbb{E}e^{tx_j^2} \leq +\infty$ whenever $|t| < \rho_j$, for $1 \leq j \leq d$.

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

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

$$|\bar{\mathcal{L}}''_{ij,\mathcal{A}_k}(\bar{\mathbf{\Psi}})[\bar{\mathcal{L}}''_{\mathcal{A}_k,\mathcal{A}_k}(\bar{\mathbf{\Psi}})]^{-1}\text{sign}(\bar{\mathbf{\Psi}}_{\mathcal{A}_k,\mathcal{A}_k})| \leq \delta,$$

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

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

and

$$\mathcal{L}(\mathbf{\Psi}) = -\frac{N}{2}\log|(\bigoplus_{k=1}^{K}\text{diag}(\mathbf{\Psi}_k))^2| + \frac{N}{2}\text{tr}(\mathbf{S}\cdot(\bigoplus_{k=1}^{K}\mathbf{\Psi}_k)^2).$$

Given assumptions (A1-A3), the theorem follows from Theorem 3.3 in Wang et al. (2020c).

2.3.2 Proof of Theorem 3.4.2

We next turn to convergence of the iterates $\{\mathbf{\Psi}^t\}$ from SG-PALM to a global optimum of (3.3). The proof leverages recent results in the convergence of alternating minimization algorithms for non-strongly convex objective (Bolte et al., 2014; Karimi et al., 2016; Li and Pong, 2018; Zhang, 2020). We outline the proof strategy:

1. We establish Lipschitz continuity of the blockwise gradient $\nabla_k H(\mathbf{\Psi})$ for $k = 1, \ldots, K$.

2. We show that the objective function $\mathcal{L}_\lambda$ satisfies the Kurdyka - Łojasiewicz (KL) property. Further, it has a KL exponent of $\frac{1}{2}$ (defined later in the proofs).

3. The KL property (with exponent $\frac{1}{2}$) is equivalent to a generalized Error Bound (EB) condition, which enables us to establish linear iterative convergence of the objective function (3.3) to its global optimum.

Definition 2.3.1 (Subdifferentials)**.** Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous function. Its domain is defined by

$$\text{dom} f := \{x \in \mathbb{R}^d : f(x) < +\infty\}.$$

If we further assume that f is convex, then the subdifferential of f at $x \in \text{dom} f$ can be defined by

$$\partial f(x) := \{v \in \mathbb{R}^d : f(z) \geq f(x) + < v, z - x >, \forall z \in \mathbb{R}^d\}.$$

The elements of $\partial f(x)$ are called subgradients of f at x.

Denote the domain of ∂f by $\text{dom}\partial f := \{x \in \mathbb{R}^d : \partial f(x) \neq \emptyset\}$. Then, if f is proper, semicontinuous, convex, and $x \in \text{dom} f$, then $\partial f(x)$ is a nonempty closed convex set. In this case, we denote by $\partial^0 f(x)$ the unique least-norm element of $\partial f(x)$ for $x \in \text{dom}\partial f$, along with $\|\partial^0 f(x)\| = +\infty$ for $x \notin \text{dom}\partial f$. Points whose subdifferential contains 0 are critical points, denoted by $\mathbf{crit} f$. For convex f, $\mathbf{crit} f = \text{argmin} f$.

Definition 2.3.2 (KL property)**.** Let Γ_{c_2} stands for the class of functions $\phi : [0, c_2] \to \mathbb{R}_+$ for $c_2 > 0$ with the properties:

(i) ϕ is continuous on $[0, c_2]$;

(ii) ϕ is smooth concave on $(0, c_2)$;

(iii) $\phi(0) = 0, \phi'(s) > 0, \forall s \in (0, c_2)$.

Further, for $x \in \mathbb{R}^d$ and any nonempty $Q \subset \mathbb{R}^d$, define the distance function $d(x, Q) := \inf_{y \in Q} \|x - y\|$. Then, a function f is said to have the Kurdyka - Łojasiewicz (KL)

property at point x_0, if there exist $c_1 > 0$, a neighborhood B of x_0, and $\phi \in \Gamma_{c_2}$ such that for all

$$x \in B(x_0, c_1) \cap \{x : f(x_0) < f(x) < f(x_0) + c_2\},$$

the following inequality holds

$$\phi'\Big(f(x) - f(x_0)\Big)\mathrm{dist}(0, \partial f(x)) \geq 1.$$

If f satisfies the KL property at each point of $\mathrm{dom}\partial f$ then f is called a KL function.

We first present two lemmas that characterize key properties of the loss function.

Lemma 2.3.3 (Blockwise Lipschitzness)**.** *The function H is convex and continuously differentiable on an open set containing domG and its gradient, is block-wise Lipschitz continuous with block Lipschitz constant $L_k > 0$ for each k, namely for all $k = 1, \ldots, K$ and all $\mathbf{\Psi}_k, \mathbf{\Psi}'_k \in \mathbb{R}^{d_k \times d_k}$*

$$\begin{aligned}&\|\nabla_k H(\mathbf{\Psi}_{i<k}, \mathbf{\Psi}_k, \mathbf{\Psi}_{i>k}) - \nabla_k H(\mathbf{\Psi}_{i<k}, \mathbf{\Psi}'_k, \mathbf{\Psi}_{i>k})\| \\ &\leq L_k \|\mathbf{\Psi}_k - \mathbf{\Psi}'_k\|,\end{aligned}$$

where $\nabla_k H$ denotes the gradient of H with respect to $\mathbf{\Psi}_k$ with all remaining $\mathbf{\Psi}_i$, $i \neq k$ fixed. Further, the function G_k for each $k = 1, \ldots, K$ is a proper lower semicontinuous (lsc) convex function.

Proof. For simplicity of notation, in this and the following proofs we use $\mathbf{\Psi}$ (i.e., omitting the subscript) to denote the set $\{\mathbf{\Psi}_k\}_{k=1}^K$ or the K-tuple $(\mathbf{\Psi}_1, \ldots, \mathbf{\Psi}_K)$ whenever there is no confusion. Recall the blockwise gradient of the smooth part of the objective

function H with respect to $\mathbf{\Psi}_k$, for each $k = 1, \ldots, K$, is given by

$$\nabla_k H(\mathbf{\Psi}) = \mathrm{diag}\Big(\Big[\mathrm{tr}\{(\mathrm{diag}((\mathbf{\Psi}_k))_{ii} + \bigoplus_{j \neq k} \mathrm{diag}(\mathbf{\Psi}_j))^{-1}\} \quad i = 1 : d_k\Big]\Big)$$
$$+ \mathbf{S}_k \mathbf{\Psi}_k + \mathbf{\Psi}_k \mathbf{S}_k + 2 \sum_{j \neq k} \mathbf{S}_{j,k}.$$

Then for $\mathbf{\Psi}_k, \mathbf{\Psi}'_k$,

$$\begin{aligned}
&\|\mathbf{S}_k \mathbf{\Psi}_k + \mathbf{\Psi}_k \mathbf{S}_k + 2 \sum_{j \neq k} \mathbf{S}_{j,k} - (\mathbf{S}_k \mathbf{\Psi}'_k + \mathbf{\Psi}'_k \mathbf{S}_k + 2 \sum_{j \neq k} \mathbf{S}_{j,k})\| \\
&= \|\mathbf{S}_k \mathbf{\Psi}_k + \mathbf{\Psi}_k \mathbf{S}_k - \mathbf{S}_k \mathbf{\Psi}'_k - \mathbf{\Psi}'_k \mathbf{S}_k\| \\
&\leq 2\|\mathbf{S}_k\|\|\mathbf{\Psi}_k - \mathbf{\Psi}'_k\|.
\end{aligned}$$

To prove Lipschitzness of the remaining parts, we consider the case of $K = 2$ for simplicity of notations. The arguments easily carry over cases of $K > 2$. In this case, denote $\mathbf{A} = (a_{ij}) := \mathbf{\Psi}_1$ and $\mathbf{B} = (b_{kl}) := \mathbf{\Psi}_2$. Let $f(\mathbf{A}) := \frac{\partial}{\partial \mathbf{A}} \log |\mathrm{diag}(\mathbf{A} \oplus \mathbf{B})|$, then

$$f(\mathbf{A}) - f(\mathbf{A}') = \mathrm{diag}\Big(\Big[\sum_{i=1}^{m_2} (a_{jj} + b_{ii})^{-1} - \sum_{i=1}^{m_2} (a'_{jj} + b_{ii})^{-1} \quad j = 1, \ldots, m_1\Big]\Big)$$

and

$$\begin{aligned}
\|f(\mathbf{A}) - f(\mathbf{A}')\|_F &= \left(\sum_{j=1}^{m_1} \Big(\sum_{i=1}^{m_2} (a_{jj} + b_{ii})^{-1} - \sum_{i=1}^{m_2} (a'_{jj} + b_{ii})^{-1} \Big)^2 \right)^{1/2} \\
&\leq \left(\sum_{j=1}^{m_1} m_2 \sum_{i=1}^{m_2} \Big((a_{jj} + b_{ii})^{-1} - (a'_{jj} + b_{ii})^{-1} \Big)^2 \right)^{1/2} \\
&= \Big(m_2 \sum_{j=1}^{m_1} \sum_{i=1}^{m_2} (c_{ji}^{-1} - (c'_{ji})^{-1})^2 \Big)^{1/2} \\
&= \Big(m_2 \sum_{j=1}^{m_1} \sum_{i=1}^{m_2} (c'_{ji})^{-2} (c'_{ji} - c_{ji})^2 c_{ji}^{-2} \Big)^{1/2} \\
&= \Big(m_2 \sum_{j=1}^{m_1} (a_{jj} - a'_{jj})^2 \sum_{i=1}^{m_2} (c'_{ji} c_{ji})^{-2} \Big)^{1/2} \\
&\leq \Big(C m_2 \sum_{j=1}^{m_1} \sum_{i=1}^{m_2} (c'_{ji} c_{ji})^{-2} \Big)^{1/2} \|\mathbf{A} - \mathbf{A}'\|_F,
\end{aligned}$$

where the first inequality is due to Cauchy-Schwartz inequality; the third line is due to $c_{ji} := a_{jj} + b_{ii}$; and in the last inequality we upper-bound each $(a_{jj} - a'_{jj})^2$ by its maximum over all j, which is absorbed in a constant C. Note that the first term in the last line above is finite as long as the summations of the diagonal elements of the factors $\mathbf{A}$ and $\mathbf{B}$ are finite, which is implied if the precision matrix $\boldsymbol{\Omega}$ defined by the Sylvester generating equation as $(\mathbf{A} \oplus \mathbf{B})^2$ has finite diagonal elements. This follows from Theorem 3.1 of Oh et al. (2014), who proved that if a symmetric matrix $\boldsymbol{\Omega}$ satisfying $\boldsymbol{\Omega} \in \mathcal{C}_0$, where

$$\mathcal{C}_0 = \Big\{ \boldsymbol{\Omega} | \mathcal{L}_\lambda(\boldsymbol{\Omega}) \leq \mathcal{L}_\lambda(\boldsymbol{\Omega}^{(0)}) = M \Big\},$$

and $\boldsymbol{\Omega}^{(0)}$ is an arbitrary initial point with a finite function value $\mathcal{L}_\lambda(\boldsymbol{\Omega}^{(0)}) := M$, the diagonal elements of $\boldsymbol{\Omega}$ are bounded above and below by constants which depend only on M, the regularization parameter λ, and the sample covariance matrix $\mathbf{S}$.

Therefore, we have

$$\|f(\mathbf{A}) - f(\mathbf{A}')\|_F \leq \tilde{C}\|\mathbf{A} - \mathbf{A}'\|_F$$

for some constant $\tilde{C} \in (0, +\infty)$. Similarly, we can establish such an inequality for $\mathbf{B}$, proving that the first term in $\nabla_k H$ is Lipschitz continuous. □

As a consequence of Lemma 2.3.3, the gradient of H, $\nabla H = (\nabla_1 H, \ldots, \nabla_K H)$ is Lipschitz continuous on bounded subsets $\mathbb{B}_1 \times \cdots \times \mathbb{B}_K$ of $\mathbb{R}^{d_1 \times d_1} \times \cdots \times \mathbb{R}^{d_K \times d_K}$ with some constant $L > 0$, such that for all $(\mathbf{\Psi}_k, \mathbf{\Psi}'_k) \in \mathbb{B}_k \times \mathbb{B}_k$,

$$\begin{aligned}
&\|(\nabla_1 H(\mathbf{\Psi}_1, \mathbf{\Psi}_{i>1}) - \nabla_1 H(\mathbf{\Psi}'_1, \mathbf{\Psi}'_{i>1}), \ldots, \\
&\nabla_K H(\mathbf{\Psi}'_{i<K}, \mathbf{\Psi}'_K) - \nabla_K H(\mathbf{\Psi}'_{i<K}, \mathbf{\Psi}'_K))\| \\
&\leq L\|(\mathbf{\Psi}_1 - \mathbf{\Psi}'_1, \ldots, \mathbf{\Psi}_K - \mathbf{\Psi}'_K)\|,
\end{aligned}$$

and we have $L \leq \sum_{k=1}^{K} L_k$.

Lemma 2.3.4 (KL property of $\mathcal{L}_\lambda$)**.** *The objective function $\mathcal{L}_\lambda(\mathbf{\Psi})$ defined in* (3.3) *satisfies the KL property. Further, ϕ in this case can be chosen to have the form $\phi(s) = \alpha s^{1/2}$, where α is some positive real number. Functions satisfying the KL property with this particular choice of ϕ is said to have a KL exponent of $\frac{1}{2}$.*

Proof. This can be established in a few steps:

1. It can be shown that the function (of $\mathbf{X}$) $\operatorname{tr}(\mathbf{S}\mathbf{X}^2) + \|\mathbf{X}\|_{1,\text{off}}$ satisfies the KL property with exponent $\frac{1}{2}$ (Karimi et al., 2016). We then apply the calculus rules of the KL exponent (compositions and separable summations) studied in Li and Pong (2018) to prove that $\operatorname{tr}(\mathbf{S}(\bigoplus_j \mathbf{\Psi}_j)^2)$ and $\sum_j \|\mathbf{\Psi}_j\|_{1,\text{off}}$ are also KL functions with exponent $\frac{1}{2}$.

2. The $-\log\det\left(\bigoplus_j \operatorname{diag}(\mathbf{\Psi}_j)\right)$ term can be shown to be KL with exponent $\frac{1}{2}$ using a transfer principle studied in Lourenço and Takeda (2019).

3. Finally, using the calculus rules of KL exponent one more time, we combine the first two results and establish that $\mathcal{L}_\lambda$ has KL exponent of $\frac{1}{2}$.

Karimi et al. (2016) proved that the following function, parameterized by some symmetric matrix $\mathbf{X}$, satisfies the KL property with KL exponent $\frac{1}{2}$:

$$\operatorname{tr}(\mathbf{S}\mathbf{X}^2) + \|\mathbf{X}\|_{1,\text{off}} = \|\mathbf{A}\mathbf{X}\|_F^2 + \|\mathbf{X}\|_{1,\text{off}}$$

for $\mathbf{S} = \mathbf{A}\mathbf{A}^T$, even when $\mathbf{A}$ is not of full rank.

We apply the calculus rules of the KL exponent studied in Li and Pong (2018) to prove that $\operatorname{tr}(\mathbf{S}(\bigoplus_j \mathbf{\Psi}_j)^2)$ and $\sum_j \|\mathbf{\Psi}_j\|_{1,\text{off}}$ are KL functions with exponent $\frac{1}{2}$. Particularly, we observe that $\operatorname{tr}\Big(\mathbf{S}(\bigoplus_j \mathbf{\Psi}_j)^2\Big)$ is the composition of functions $\mathbf{X} \to \operatorname{tr}(\mathbf{S}\mathbf{X})$ and $(\mathbf{X}_1, \ldots, \mathbf{X}_K) \to \bigoplus_j \mathbf{X}_j$; and $\sum_j \|\mathbf{\Psi}_j\|_{1,\text{off}}$ is a separable block summation of functions $\mathbf{X}_j \to \|\mathbf{X}_j\|_{1,\text{off}}$.

Thus, by Theorem 3.2. (exponent for composition of KL functions) in Li and Pong (2018), since the Kronecker sum operation is linear and hence continuously differentiable, the trace function is KL with exponent $\frac{1}{2}$, and the mapping $(\mathbf{X}_1, \ldots, \mathbf{X}_K) \to \bigoplus_j \mathbf{X}_j$ is clearly one to one, the function $\operatorname{tr}(\mathbf{S}(\bigoplus_j \mathbf{\Psi}_j)^2)$ has the KL exponent of $\frac{1}{2}$. By Theorem 3.3. (exponent for block separable sums of KL functions) in Li and Pong (2018), since the function $\|\cdot\|_{1,\text{off}}$ is proper, closed, continuous on its domain, and is KL with exponent $\frac{1}{2}$, the function $\|\mathbf{X}_j\|_{1,\text{off}}$ is KL with an exponent of $\frac{1}{2}$.

It remains to prove that $-\log\det\Big(\bigoplus_j \operatorname{diag}(\mathbf{\Psi}_j)\Big)$ is also a KL function with an exponent of $\frac{1}{2}$. By Theorem 30 in Lourenço and Takeda (2019), if we have $f : \mathbb{R}^r \to \mathbb{R}$ a symmetric function and $F : \mathcal{E} \to \mathbb{R}$ the corresponding spectral function, the followings hold

(i) F satisfies the KL property at $\mathbf{X}$ iff f satisfies the KL property at $\lambda(\mathbf{X})$, i.e., the eigenvalues of $\mathbf{X}$.

(ii) F satisfies the KL property with exponent α iff f satisfies the KL property with

exponent α at $\lambda(\mathbf{X})$.

Here, take $f(\lambda(\mathbf{X})) := -\sum_{i=1}^{r} \log(\lambda_i(\mathbf{X}))$, and $F(\mathbf{X}) := -\log\det(\mathbf{X})$ the corresponding spectral function. Then, the function f is symmetric since its value is invariant to permutations of its arguments, and it is a strictly convex function in its domain, so it satisfies the KL property with an exponent of $\frac{1}{2}$. Therefore, F satisfies the KL property with the same KL exponent of $\frac{1}{2}$. Now, we apply the calculus rules for KL functions again. As both the Kronecker sum and the diag operators are linear, we conclude that $-\log\det\left(\bigoplus_j \operatorname{diag}(\mathbf{\Psi}_j)\right)$ is a KL function with an exponent of $\frac{1}{2}$.

Overall, we have that the negative log-pseudolikelihood function $\mathcal{L}(\mathbf{\Psi})$ satisfies the KL property with an exponent of $\frac{1}{2}$. □

Now we are ready to prove Theorem 3.4.2. We follow Zhang (2020) and divide the proof into three steps.

Step 1. We obtain a sufficient decrease property for the loss function $\mathcal{L}$ in terms of the squared distance of two successive iterates:

$$\mathcal{L}(\mathbf{\Psi}^{(t)}) - \mathcal{L}(\mathbf{\Psi}^{(t+1)}) \geq \frac{L_{\min}}{2}\|\mathbf{\Psi}^{(t)} - \mathbf{\Psi}^{(t+1)}\|^2. \tag{2.2}$$

Here and below, $\mathbf{\Psi}^{(t+1)} := (\mathbf{\Psi}_1^{(t+1)}, \ldots, \mathbf{\Psi}_K^{(t+1)})$ and $L_{\min} := \min_k L_k$. First note that at iteration t, the line search condition is satisfied for step size $\frac{1}{\eta_k^{(t)}} \geq L_k$, where L_k is the Lipschitz constant for $\nabla_k H$. Further, it follows that for SG-PALM with backtracking one has for every $t \geq 0$ and each $k = 1, \ldots, K$,

$$\frac{1}{\eta_k^{(0)}} \leq \frac{1}{\eta_k^{(t)}} \leq cL_k,$$

where $c > 0$ is the backtracking constant.

Then by Lemma 3.1 in Shefi and Teboulle (2016), we get

$$\begin{aligned}\mathcal{L}(\mathbf{\Psi}^{(t)}) - \mathcal{L}(\mathbf{\Psi}^{(t+1)}) &\geq \frac{1}{2\eta_{\min}^{(t+1)}}\|\mathbf{\Psi}^{(t)} - \mathbf{\Psi}^{(t+1)}\|^2 \\ &\geq \frac{L_{\min}}{2}\|\mathbf{\Psi}^{(t)} - \mathbf{\Psi}^{(t+1)}\|^2\end{aligned}$$

for $\eta_{\min}^{(t)} := \min_k \eta_k^{(t)}$.

Step 2. By Lemma 2.3.4, $\mathcal{L}$ satisfies the KL property with an exponent of $\frac{1}{2}$. Then from Definition 2.3.2, this suggests that at $x = \mathbf{\Psi}^{t+1}$ and $f(x_0) = \min \mathcal{L}$

$$\|\partial^0 \mathcal{L}(\mathbf{\Psi}^{t+1})\| \geq \alpha\sqrt{\mathcal{L}(\mathbf{\Psi}^{t+1}) - \min \mathcal{L}}, \tag{2.3}$$

where $\alpha > 0$ is a fixed constant defined in Lemma 2.3.4. This property is equivalent to the error bound condition, $(\partial^0\mathcal{L}, \alpha, \Omega)$-(res-obj-EB), defined in Definition 5 in Zhang (2020), for $\Omega \subset \text{dom}\partial\mathcal{L}$. This is strictly weaker than strong convexity (see Section 4 in Zhang (2020)).

At iteration $t+1$, there exists $\xi_k^{(t+1)} \in \partial G_k(\mathbf{\Psi}_k^{(t+1)})$ satisfying the optimality condition:

$$\nabla_k H(\mathbf{\Psi}_{i<k}^{(t+1)}, \mathbf{\Psi}_{i\geq k}^{(t)}) + \frac{1}{\eta_k^{(t+1)}}(\mathbf{\Psi}_k^{(t+1)} - \mathbf{\Psi}_k^{(t)}) + \xi_k^{(t+1)} = 0.$$

Let $\xi^{(t+1)} := (\xi_1^{(t+1)}, \ldots, \xi_K^{(t+1)})$. Then,

$$\nabla H(\mathbf{\Psi}^{(t+1)}) + \xi^{(t+1)} \in \partial\mathcal{L}(\mathbf{\Psi}^{(t+1)})$$

and hence the error bound condition becomes

$$\mathcal{L}(\mathbf{\Psi}^{(t+1)}) - \min \mathcal{L} \leq \frac{\|\partial^0\mathcal{L}(\mathbf{\Psi}^{(t+1)})\|^2}{\alpha^2} \leq \frac{\|\nabla H(\mathbf{\Psi}^{(t+1)}) + \xi^{(t+1)}\|^2}{\alpha^2}.$$

It follows that

$$
\begin{aligned}
&\|\nabla H(\mathbf{\Psi}^{(t+1)}) + \xi^{(t+1)}\|^2 \\
&= \sum_{k=1}^{K} \|\nabla_k H(\mathbf{\Psi}^{(t+1)}) - \nabla_k H(\mathbf{\Psi}_{i<k}^{(t+1)}, \mathbf{\Psi}_{i\geq k}^{(t)}) - \frac{1}{\eta_k^{(t+1)}}(\mathbf{\Psi}_k^{(t+1)} - \mathbf{\Psi}_k^{(t)})\|^2 \\
&\leq \sum_{k=1}^{K} 2\|\nabla_k H(\mathbf{\Psi}^{(t+1)}) - \nabla_k H(\mathbf{\Psi}_{i<k}^{(t+1)}, \mathbf{\Psi}_{i\geq k}^{(t)})\|^2 + \sum_{k=1}^{K} \frac{2}{(\eta_k^{(t+1)})^2}\|\mathbf{\Psi}_k^{(t+1)} - \mathbf{\Psi}_k^{(t)}\|^2 \\
&\leq \sum_{k=1}^{K} 2\|\nabla H(\mathbf{\Psi}^{(t+1)}) - \nabla H(\mathbf{\Psi}_{i<k}^{(t+1)}, \mathbf{\Psi}_{i\geq k}^{(t)})\|^2 + \sum_{k=1}^{K} \frac{2}{(\eta_k^{(t+1)})^2}\|\mathbf{\Psi}_k^{(t+1)} - \mathbf{\Psi}_k^{(t)}\|^2 \\
&\leq \sum_{k=1}^{K} 2\Big(\sum_{j=1}^{K} \frac{1}{\eta_j^{(t+1)}}\Big)^2 \|\mathbf{\Psi}_{i\geq k}^{(t+1)} - \mathbf{\Psi}_{i\geq k}^{(t)}\|^2 + \sum_{k=1}^{K} \frac{2}{(\eta_k^{(t+1)})^2}\|\mathbf{\Psi}_k^{(t+1)} - \mathbf{\Psi}_k^{(t)}\|^2 \\
&\leq \left(2Kc^2\Big(\sum_{j=1}^{K} L_j\Big)^2 + 2c^2 L_{\max}\right)\|\mathbf{\Psi}^{(t+1)} - \mathbf{\Psi}^{(t)}\|^2.
\end{aligned}
$$

Therefore, we get

$$
\mathcal{L}(\mathbf{\Psi}^{(t+1)}) - \min \mathcal{L} \leq \frac{\left(2Kc^2\Big(\sum_{j=1}^{K} L_j\Big)^2 + 2c^2 L_{\max}\right)}{\alpha^2}\|\mathbf{\Psi}^{(t+1)} - \mathbf{\Psi}^{(t)}\|^2. \tag{2.4}
$$

Step 3. Combining (2.2) and (2.4), we have

$$
\begin{aligned}
\mathcal{L}(\mathbf{\Psi}^{(t)}) - \min \mathcal{L} &= \Big(\mathcal{L}(\mathbf{\Psi}^{(t)}) - \mathcal{L}(\mathbf{\Psi}^{(t+1)})\Big) + \Big(\mathcal{L}(\mathbf{\Psi}^{(t+1)}) - \min \mathcal{L}\Big) \\
&\geq \frac{L_{\min}}{2}\|\mathbf{\Psi}^{(t)} - \mathbf{\Psi}^{(t+1)}\|^2 + \Big(\mathcal{L}(\mathbf{\Psi}^{(t+1)}) - \min \mathcal{L}\Big) \\
&\geq \left(\frac{\alpha^2 L_{\min}}{4Kc^2(\sum_{j=1}^{K} L_j)^2 + 4c^2 L_{\max}} + 1\right)\Big(\mathcal{L}(\mathbf{\Psi}^{(t+1)}) - \min \mathcal{L}\Big).
\end{aligned}
$$

This completes the proof.

2.4 SG-PALM with Non-Convex Regularizers

The estimation algorithm for non-convex regularizer is largely the same as Algorithm III.1, except with an additional term added to the gradient term. Specifically, the updates are of the form

$$\boldsymbol{\Psi}_k^{(t+1)} = \text{prox}_{\eta_k^t \lambda_k}^{\|\cdot\|_{1,\text{off}}}\Big(\boldsymbol{\Psi}_k^t - \eta_k^t \nabla_k \bar{H}(\boldsymbol{\Psi}_{i<k}^{t+1}, \boldsymbol{\Psi}_{i\geq k}^t)\Big),$$

where

$$\bar{H}(\boldsymbol{\Psi}) = H(\boldsymbol{\Psi}) + \sum_{k=1}^{K}\sum_{i\neq j}\Big(g_{\lambda_k}([\boldsymbol{\Psi}_k]_{i,j}) - \lambda_k|[\boldsymbol{\Psi}_k]_{i,j}|\Big).$$

Here, the formulation covers a range of non-convex regularizations. Particularly, the SCAD penalty (Fan and Li, 2001) with parameter $a > 2$ is given by

$$g_\lambda(t) = \begin{cases} \lambda|t|, & \text{if} \quad |t| < \lambda \\ -\frac{t^2 - 2a\lambda|t| + \lambda^2}{2(a-1)}, & \text{if} \quad \lambda < |t| < a\lambda \\ \frac{(a+1)\lambda^2}{2}, & \text{if} \quad a\lambda < |t|, \end{cases}$$

which is linear for small $|t|$, constant for large $|t|$, and a transition between the two regimes for moderate $|t|$.

The MCP penalty (Zhang et al., 2010) with parameter $a > 0$ is given by

$$g_\lambda(t) = \text{sign}(t)\lambda \int_0^{|t|} \Big(1 - \frac{z}{a\lambda}\Big)_+ dz,$$

which gives a smoother transition between the approximately linear region and the constant region ($t > a\lambda$) as defined in SCAD.

The updates can also be written as

$$\mathbf{\Psi}_k^{(t+1)} = \operatorname{prox}_{\eta_k^t \lambda_k}^{\|\cdot\|_{1,\text{off}}}\left(\mathbf{\Psi}_k^t - \eta_k^t \nabla_k\Big(H(\mathbf{\Psi}_{i<k}^{t+1}, \mathbf{\Psi}_{i\geq k}^t) + Q'_{\lambda_k}(\mathbf{\Psi}_k)\Big)\right),$$

where $q'_\lambda(t) := \frac{d}{dt}(g_\lambda(t) - \lambda|t|)$ for $t \neq 0$ and $q'_\lambda(0) = 0$ and Q'_λ denotes q'_λ applied elementwise to a matrix argument. These updates can be inserted into the framework of Algorithm III.1. The details are summarized in Algorithm B.1.

Algorithm B.1: SG-PALM with non-convex regularizer

Input: Data tensor $\boldsymbol{\mathcal{X}}$, mode-k Gram matrix $\mathbf{S}_k$, regularizing parameter λ_k, backtracking constant $c \in (0,1)$, initial step size η_0, initial iterate $\mathbf{\Psi}_k$ for each $k = 1, \ldots, K$.

while not converged **do**

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

Line search: Let η_k^t be the largest element of $\{c^j \eta_{k,0}^t\}_{j=1,\ldots}$ such that condition (3.8) is satisfied for

$$\mathbf{\Psi}_k^{t+1} = \operatorname{prox}_{\eta_k^t \lambda_k}^{\|\cdot\|_{1,\text{off}}}\left(\mathbf{\Psi}_k^t - \eta_k^t \nabla_k\Big(H(\mathbf{\Psi}_{i<k}^{t+1}, \mathbf{\Psi}_{i\geq k}^t) + Q'_{\lambda_k}(\mathbf{\Psi}_k)\Big)\right).$$

Update:

$$\mathbf{\Psi}_k^{t+1} \longleftarrow \operatorname{prox}_{\eta_k^t \lambda_k}^{\|\cdot\|_{1,\text{off}}}\left(\mathbf{\Psi}_k^t - \eta_k^t \nabla_k\Big(H(\mathbf{\Psi}_{i<k}^{t+1}, \mathbf{\Psi}_{i\geq k}^t) + Q'_{\lambda_k}(\mathbf{\Psi}_k)\Big)\right).$$

end for

Next initial stepsize: Compute Barzilai-Borwein stepsize $\eta_0^{t+1} = \min_k \eta_{k,0}^{t+1}$, where $\eta_{k,0}^{t+1}$ is computed via (3.9).

end while

Output: Final iterates $\{\mathbf{\Psi}_k\}_{k=1}^K$.

2.4.1 Convergence Property

Consider a sequence of iterate $\{\mathbf{x}^t\}_{t\in\mathbb{N}}$ generated by a generic PALM algorithm for minimizing some objective function f. Specifically, assume

($\mathcal{H}_1$) $\inf f > -\infty$.

($\mathcal{H}_2$) The restriction of the function to its domain is a continuous function.

($\mathcal{H}_3$) The function satisfies the KL property.

Then, as in Theorem 2 of Attouch and Bolte (2009), if this objective function satisfying $(\mathcal{H}_1), (\mathcal{H}_2), (\mathcal{H}_3)$ in addition satisfies the KL property with

$$\phi(s) = \alpha s^{1-\theta},$$

where $\alpha > 0$ and $\theta \in (0, 1]$. Then, for $\mathbf{x}^*$ some critical point of f, the following estimations hold

(i) If $\theta = 0$ then the sequence of iterates converges to $\mathbf{x}^*$ in a finite number of steps.

(ii) If $\theta \in (0, \frac{1}{2}]$ then there exist $\omega > 0$ and $\tau \in [0, 1)$ such that $\|\mathbf{x}^t - \mathbf{x}^*\| \leq \omega\tau^t$.

(iii) If $\theta \in (\frac{1}{2}, 1)$ then there exist $\omega > 0$ such that $\|\mathbf{x}^t - \mathbf{x}^*\| \leq \omega t^{-\frac{1-\theta}{1\theta-1}}$.

In the case of SG-PALM with non-convex regularizations, so long as the non-convex $\mathcal{L}$ satisfies the KL property with an exponent in $(0, \frac{1}{2}]$, the algorithm remains linearly convergent (to a critical point). We argue that this is true for SG-PALM with MCP or SCAD penalty. Li and Pong (2018) showed that penalized least square problems with such penalty functions satisfy the KL property with an exponent of $\frac{1}{2}$. The proof strategy for the convex case can be easily adopted, incorporating the KL results for MCP and SCAD in Li and Pong (2018), to show that the new $\mathcal{L}$ still has KL exponent of $\frac{1}{2}$. Therefore, SG-PALM with MCP or SCAD penalty converges linearly in the sense outlined above.

2.5 Additional Details of the Solar Flare Experiments

2.5.1 HMI and AIA Data

The Solar Dynamics Observatory (SDO)/Helioseismic & Magnetic Imager (HMI) data characterize solar variability including the Sun's interior and the various components of magnetic activity; the SDO/Atmospheric Imaging Assembly (AIA) data contain a set of measurements of the solar atmosphere spectrum at various wavelengths.

In general, HMI produces data that is particularly useful in determining the mechanisms of solar variability and how the physical processes inside the Sun that are related to surface magnetic field and activity. AIA contains structural information about solar flares, and the the high AIA pixel values are correlated with the flaring intensities. We are interested in examining if combination of multiple instruments enhances our understanding of the solar flares, comparing to the case of single instrument. Both HMI and AIA produce multi-band (or multi-channel) images, for this experiment we use all three channels of the HMI images and $9.4, 13.1, 17.1, 19.3$ nm wavelength channels of the AIA images. For a detailed descriptions of the instruments and all channels of the images, see `https://en.wikipedia.org/wiki/Solar_Dynamics_Observatory` and the references therein. Furthermore, for training and testing involved in this study, we used the data described in (Galvez et al., 2019), which are further pre-processed HMI and AIA imaging data for machine learning methods.

2.5.2 Classification of Solar Flares/Active Regions (AR)

The classification system for solar flares uses the letters A, B, C, M or X, according to the peak flux in watts per square metre (W/m^2) of X-rays with wavelengths 100 to 800 picometres (1 to 8 angstroms), as measured at the Earth by the GOES spacecraft (`https://en.wikipedia.org/wiki/Solar_flare#Classification`). Here, A usually refers to a "quite" region, which means that the peak flux of that region is not high enough to be classified as a real flare; B usually refers a "weak" region, where the flare is not strong enough to have impact on spacecrafts, earth, etc; and M or X refers to a "strong" region that is the most detrimental. Differentiating between a weak and a strong flare/region ahead of time is a fundamental task in space physics and has recently attracted attentions from the machine learning community (Chen et al., 2019; Jiao et al., 2020a; Sun et al., 2019). In our study, we also focus on B and M/X flares and attempt to predict the videos that lead to either one of these two

types of flares.

2.5.3 Run Time Comparison

We compare run times of the SG-PALM algorithm for estimating the precision matrix from the solar flare data with SyGlasso. Table B.1 illustrates that the SG-PALM algorithm converges faster in wallclock time. Note that in this real dataset, which is potentially non-Gaussian, the convergence behavior of the algorithms is different compare to synthetic examples. Nonetheless, SG-PALM enjoys an order of magnitude speed-up over SyGlasso.

Table B.1: Run time (in seconds) comparisons between SyGlasso and SG-PALM on solar flare data for different regularization parameters. Note that the SG-PALM is an order of magnitude faster that SyGlasso.

λ	**SyGlasso**		**SG-PALM**	
	iter	**sec**	**iter**	**sec**
0.28	47	5772.1	89	583.7
0.41	43	5589.0	86	583.4
0.54	45	5673.7	85	568.8
0.67	42	5433.0	77	522.6
0.79	39	4983.2	82	511.4
0.92	40	5031.9	72	498.0
1.05	39	4303.7	76	452.2
1.18	41	4234.7	64	437.6
1.30	40	4039.5	58	406.9
1.43	35	3830.7	64	364.9

2.5.4 Examples of Predicted Magnetogram Images

Figure B.1 depicts examples of the predicted HMI channels by SG-PALM. We observe that the proposed method was able to reasonably capture various components of the magnetic field and activity. Note that the spatial behaviors of the HMI components are quite different from those of AIA channels, that is, the structures tend to be less smooth and continuous (e.g., separated holes and bright spots) in HMI.

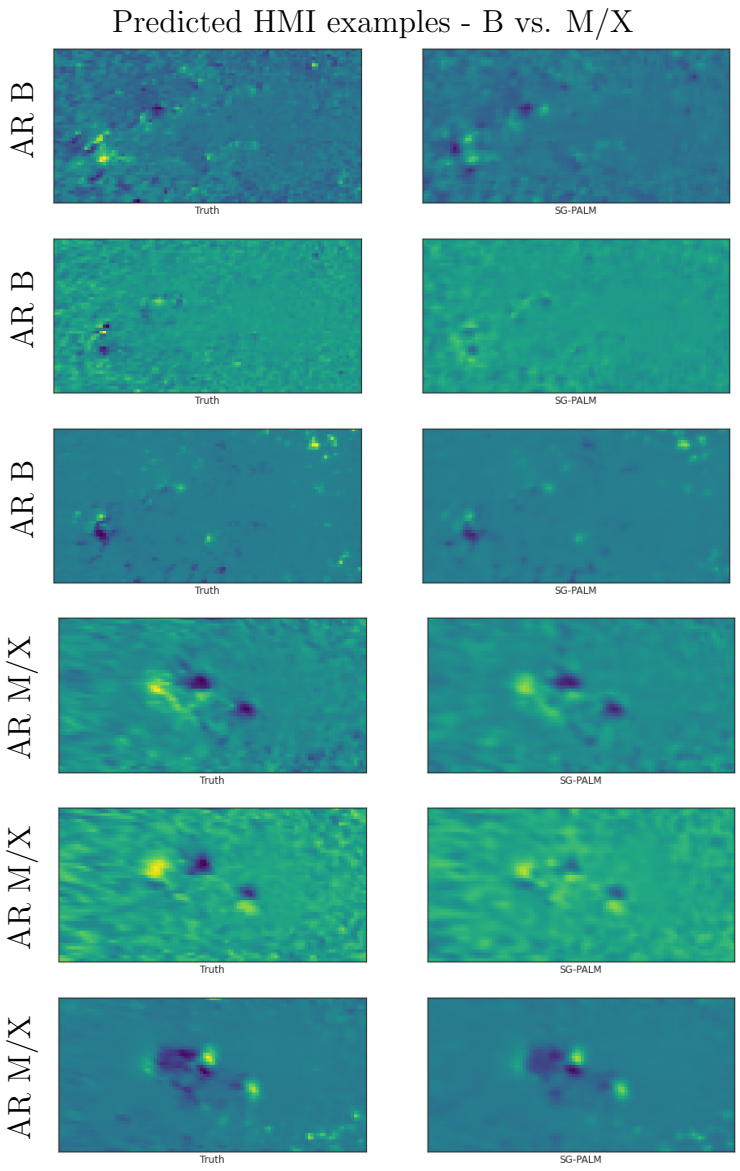


Figure B.1: Examples of one-hour ahead prediction of the first three channels (HMI components) of ending frames of 13-frame videos, leading to B- (first three rows) and MX-class (last three rows) flares, produced by the SG-PALM, comparing to the real image (left column). Similarly to AIA predictions, linear forward predictors tend to underestimate the contrast ratio of the images. Nonetheless, the SG-PALM algorithm was able to both capture the spatial structures of the underlying magnetic fields. HMI images tend to be harder to predict, as indicated by the increased number and decreased degree of smoothness of features, signifying the underlying magnetic activity on the solar surface.

2.5.5 Multi-instrument vs. Single Instrument Prediction

To illustrate the advantages of multi-instrument analysis, we compare the NRMSEs between an AIA-only (i.e., last four channels of the dataset) and an HMI&AIA (i.e., all seven channels of the dataset) study in predicting the last frames of 13-frame AIA videos, for each flare class, respectively, using the proposed SG-PALM. The results are depicted in Figure B.2, where the average, standard deviation, and range of the NRMSEs across pixels are also shown for each error image. By leveraging the cross-instrument correlation structure, there is a $0.5\% - 1\%$ drop in the averaged error rates and a $2\% - 4\%$ drop in the range of the errors.

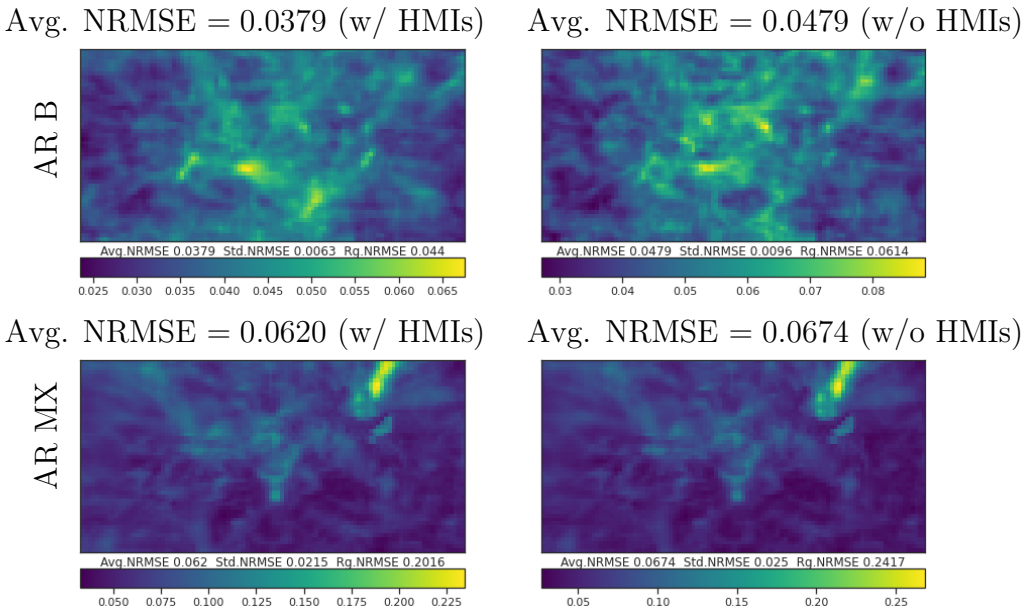


Figure B.2: Comparison of the SG-PALM performance measured by NRMSE in predicting the AIA channels (i.e., last four channels) of the ending frame of 13-frame videos leading to B- and MX-class solar flares, by using all HMI&AIA channels (left column) and AIA-only channels (right column). The NRMSEs are computed by averaging across both testing samples and channels for each pixel. Note that there are improvements in both the averaged errors rates and the uncertainty in those errors (i.e., range of the errors) by including multi-instrument image channels.

2.5.6 Illustration of the Difficulty of Predictions for Two Flares Classes

We demonstrate the difficulty of forward predictions of video frames. Figure B.3 depicts two different channels of multiple frames from two videos leading to MX-class solar flares. Note that the current frame is the 13th frame in the sequence that we are trying to predict. We observe that the prediction task is particularly difficult if there is a sudden transition of either the brightness or spatial structure of the frames near the end of the video. These sudden transitions are more frequent for MX flares than for B flares. In addition, as MX flares are generally considered as rare events (i.e., less frequent than B flares), it is harder for SG-PALM or related methods to learn a common correlation structures from training data.

On the other hand, typical image sequences leading to B flares exhibit much smoother transitions from frame to frame. As shown in Figure B.4, the SG-PALM was able to produce remarkably good predictions of the current frames.

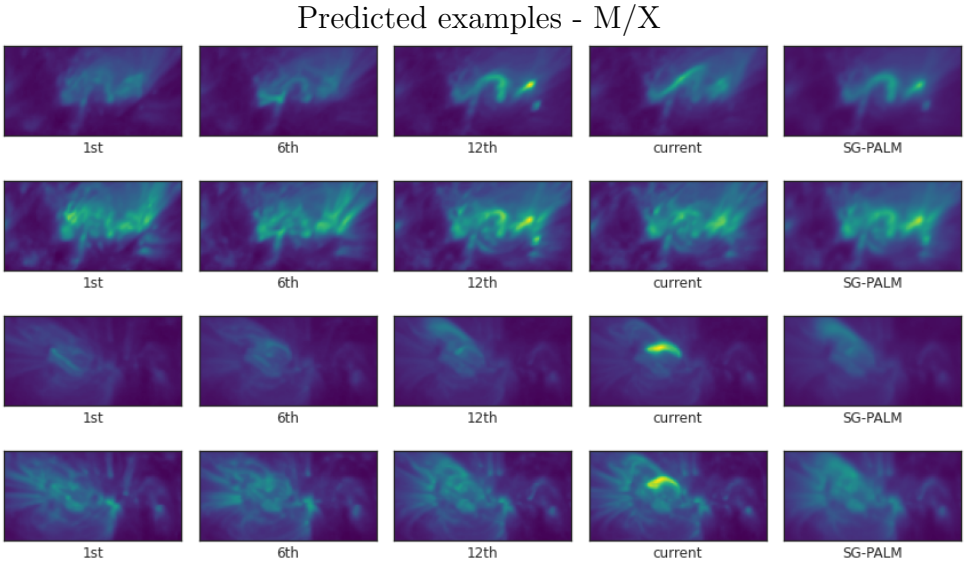


Figure B.3: Examples of frames at various timestamps of videos preceding the predictions of the last frames (last column) that lead to MX flares. Here, the first two rows correspond to the same video as the last two rows in Figure III.3. Note that the prediction tasks are difficult in these two extreme cases, where there are dramatic changes from the 12th to the current (13th) frames.

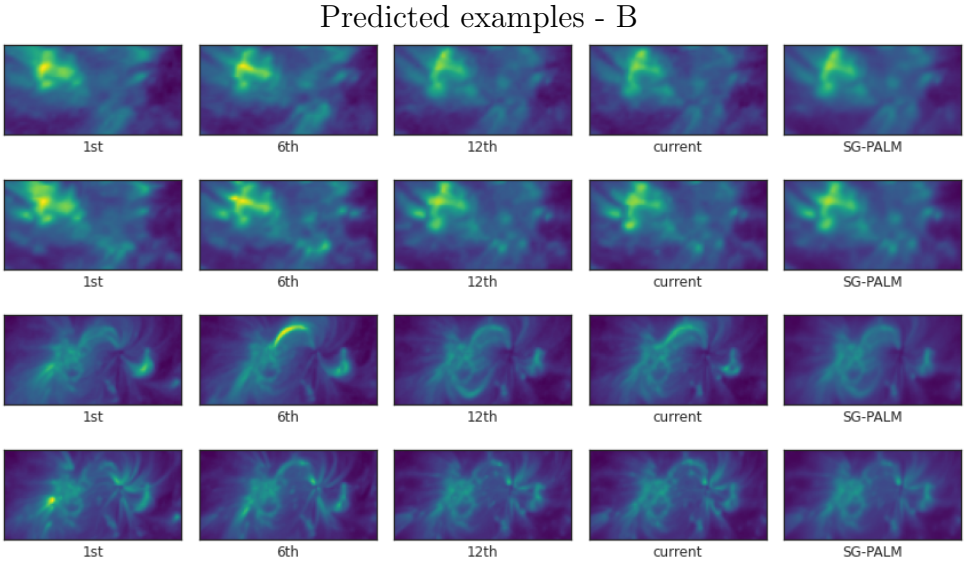


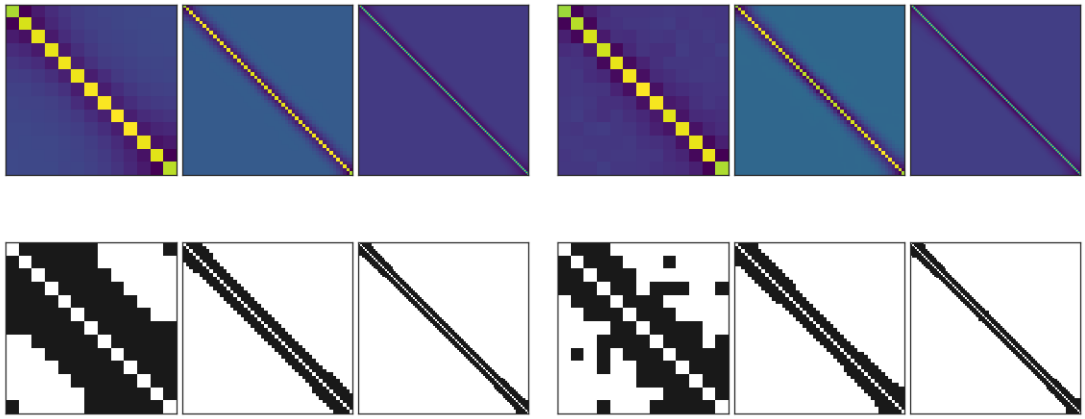
Figure B.4: Examples of frames at various timestamps of videos preceding the predictions of the last frames (last column) that lead to B flares. Here, the first two rows correspond to the same video as the first two rows in Figure III.3. Note that the prediction tasks are easier than those illustrated in Figure B.3, since the transitions near the end of the videos are much smoother.

2.5.7 Illustration of the Estimated Sylvester Generating Factors

Figure B.5 illustrates the patterns of the estimated Sylvester generating factors ($\mathbf{\Psi}_k$'s) for each flare class. Here, the videos from both classes appear to form Markov Random Fields, that is, each pixel only depends on its close neighbors in space and time given all other pixels. This is demonstrated by observing that the temporal or each of the spatial generating factor, which can be interpreted as conditional dependence graph for the corresponding mode, has its energies concentrate around the diagonal and decay as the nodes move far apart (in space or time).

The spatial patterns are similar for different flares. Although the exact spatial patterns are different from one frame to another, they always have their energies being concentrated at certain region (i.e., the brightest spot) that is usually close to the center of the images. This is due to the way how these images were curated and pre-processed before analysis. On the other hand, the temporal structures are quite

different. Specifically, B flares tend to have longer range dependencies, as the frames leading to these types flares are smooth, which is consistent with results from the previous section.



(a) Estimated precision matrices - B flares (b) Estimated precision matrices - M/X flares

Figure B.5: Estimated spatial and two (longitude and latitude) temporal Sylvester generating factors for B and MX solar flares, along with their off-diagonal sparsity patterns (second row in each subplot). Both classes exhibit autoregressive dependence structures (across time or space). Note the significant difference in the temporal components, where the B flares exhibit longer range dependency. This is consistent with the smooth transition property of the corresponding videos as illustrated previously.

APPENDIX C

Appendix of Chapter IV

In this Appendix, we discuss the blocked versions of the Possion-AR(1) and convection-diffusion processes.

Poisson-AR(1) Process. The first extension, which we call the Poisson-AR(1) process, imposes an autoregressive temporal model of order 1 on the source function f in the Poisson equation (4.3). Specifically, we say a sequence of discretized spatial observations $\{\mathbf{U}^k \in \mathbb{R}^{d_1 \times d_2}\}_k$ indexed by time step $k = 1, \cdots, T$ is from a Poisson-AR(1) process if

$$(\mathbf{A}_{d_1} \oplus \mathbf{A}_{d_2}) \operatorname{vec}(\mathbf{U}^k) = \operatorname{vec}(\mathbf{Z}^k), \tag{3.1}$$

$$\operatorname{vec}(\mathbf{Z}^k) = a \operatorname{vec}(\mathbf{Z}^{k-1}) + \operatorname{vec}(\mathbf{W}^k), \quad |a| < 1, \tag{3.2}$$

where $\{\mathbf{W}^k \in \mathbb{R}^{d_1 \times d_2}\}_k$ is spatiotemporal white noise, i.e., $W^k_{i,j} \sim \mathcal{N}(0, \sigma_w^2)$, i.i.d. Assuming $\mathbf{Z}^0 = \mathbf{0}$ and defining the T-by-T matrix

$$\mathbf{B} = \begin{bmatrix} 1 & -a & & \\ & 1 & \ddots & \\ & & \ddots & -a \\ & & & 1 \end{bmatrix},$$

the above linear system of equations can be written as $(\mathbf{A}_{d_1} \oplus \mathbf{A}_{d_2})\mathbf{U}\mathbf{B} = \mathbf{W}$, or equivalently,

$$\left(\mathbf{B}^T \otimes (\mathbf{A}_{d_1} \oplus \mathbf{A}_{d_2})\right) \mathrm{vec}(\mathbf{U}) = \mathrm{vec}(\mathbf{W}), \tag{3.3}$$

where $\mathbf{U} = [\mathrm{vec}(\mathbf{U}^1)\, \mathrm{vec}(\mathbf{U}^2) \ldots \mathrm{vec}(\mathbf{U}^T)] \in \mathbb{R}^{d_1 d_2 \times T}$ and $\mathbf{W}$ is defined likewise. The inverse covariance of $\mathbf{U}$, despite having a large size of $d_1 d_2 T \times d_1 d_2 T$, is sparse and has a mixed Kronecker sum and product structure.

Convection-diffusion Process. The second time-varying extension of the Poisson PDE model (4.3) is based on the convection-diffusion process Chandrasekhar (1943)

$$\frac{\partial u}{\partial t} = \theta \sum_{i=1}^{2} \frac{\partial^2 u}{\partial x_i^2} - \epsilon \sum_{i=1}^{2} \frac{\partial u}{\partial x_i}. \tag{3.4}$$

Here, $\theta > 0$ is the diffusivity; and $\epsilon \in \mathbb{R}$ is the convection velocity of the quantity along each coordinate. Note that for simplicity of discussion here, we assume these coefficients do not change with space and time (see, Stocker (2011), for example, for a detailed discussion). These equations are closely related to the Navier-Stokes equation commonly used in stochastic modeling for weather and climate prediction (Chandrasekhar, 1943; Stocker, 2011). Coupled with Maxwell's equations, these equations can be used to model magneto-hydrodynamics (Roberts, 2006), which characterize solar activities including flares.

A solution of Equation (4.9) can be approximated similarly as in the Poisson equation case, through a finite difference approach. Denote the discrete spatial samples of $u(\mathbf{x}, t)$ at time t_k as a matrix $\mathbf{U}^k \in \mathbb{R}^{d_1 \times d_2}$. We obtain a discretized update

propagating $u(\mathbf{x}, t)$ in space and time, which locally satisfies

$$\begin{aligned}\frac{U_{i,j}^k - U_{i,j}^{k-1}}{\Delta t} = \theta \left(\frac{U_{i+1,j}^k + U_{i-1,j}^k + U_{i,j+1}^k + U_{i,j-1}^k - 4U_{i,j}^k}{h^2} \right) \\ - \epsilon \left(\frac{U_{i+1,j}^k - U_{i-1,j}^k + U_{i,j+1}^k - U_{i,j-1}^k}{2h} \right),\end{aligned} \tag{3.5}$$

where $\Delta t = t_{k+1} - t_k$ is the time step and h is the mesh step (spatial grid spacing). Similarly to the Poisson-AR(1) process, in the following, we consider a "blocked" version of the convection-diffusion process.

We define the first-order and second-order discretized differential operators, denote by $\mathbf{D}$ and $\mathbf{A}$, respectively:

$$\mathbf{D} = \begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix}.$$

Then, Equation (4.10) can be written as

$$\begin{aligned}\frac{1}{\Delta t}(\mathbf{D} \otimes \mathbf{I} \otimes \mathbf{I}) \operatorname{vec} \mathbf{U} = &\frac{\theta}{h^2}(\mathbf{I} \otimes \mathbf{A} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{I} \otimes \mathbf{A}) \operatorname{vec} \mathbf{U} \\ &- \frac{\epsilon}{2h}(\mathbf{I} \otimes \mathbf{D} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{I} \otimes \mathbf{D}) \operatorname{vec} \mathbf{U},\end{aligned} \tag{3.6}$$

where $\mathbf{U} = [\operatorname{vec}(\mathbf{U}^1) \operatorname{vec}(\mathbf{U}^2) \dots \operatorname{vec}(\mathbf{U}^T)] \in \mathbb{R}^{d_1 d_2 \times T}$. Assuming the process is driven by some white noise $\mathbf{W}$, similarly defined as in the Poisson-AR equation, the inverse covariance of $\mathbf{U}$ is again sparse and has a mixed Kronecker sum and product structure.

We consider a spatio-temporal process (2D space + time) on a 8×8 spatial grid, and generated instances of state trajectories, which we call true states, according to the Poisson-AR(1) and the convection-diffusion dynamics for $T = 50$ time steps. Several realizations of the true state variables are shown in Figure C.1 to illustrate

how the states evolve over time under each model.

We generated $N = 50$ independent realizations of random tensors of dimension 64×50 and estimated the state covariance / inverse covariance (with $K = 2$) using several sparse (multiway) inverse covariance estimation methods described in Section 4.2 of Chapter IV, including Glasso, KPCA, Tlasso, TeraLasso, SG-PALM. Note that none of the above-mentioned models operate under the true generative processes (i.e., there is model mismatch with the data). Here, the sparsity-regularized methods are all implemented with an ℓ_1 penalty function, and the penalty parameters were selected similarly and guided by the theoretical results in Table IV.1. For example, for SG-PALM, we use a penalty parameter of $\lambda_k = C\sqrt{\frac{d_k \log d}{N}}$ where C is chosen by optimizing a normalized Frobenius norm error between the estimate and the truth, over a range of λ values parameterized by C. For the KPCA alorithm, both the nuclear norm penalty parameter and the separation rank are selected by optimizing a normalized Frobenius norm error via grid search.

Summary of the estimation accuracy in terms of the recovery of the matrix entries measured normalized Frobenius norm error as well as the recovery of the sparsity patterns measured by Mathews Correlation Coefficient (Matthews, 1975) are reported in Table C.1. In Figure C.2 and C.3 we show the true and the estimated inverse covariance matrices obtained for all the methods except KPCA, under both the Poisson-AR (panel (a)) and the Convection-Diffusion processes (panel (b)). The inverse covariances under both generating processes admit structures with a mix of Kronecker sums and Kronecker products of sparse matrices. In both of the cases, the SG-PALM produces the estimates with the closest and richest structures, which we believe is due to the nature of the Sylvester graphical model that imposes a squared KS structure on the precision matrix. Tlasso has comparable performances and achieves the best matrix approximation error under the convection-diffusion generating process. This might be due to the fact that the KP model corresponds to an underlying spatio-

temporal autoregressive process. TeraLasso seems to produce the biggest model mismatch as indicated by the MCC scores. Although Glasso works reasonable well given that it ignores any multiway structure, this also leads to an increased computational cost for the vector-variate estimating algorithm. In Figure C.4, we also show compare the true covariance matrix and the estimate obtained by KPCA. Here, although the KPCA model does not match the underlying generating process, the estimates were able to capture certain blocking patterns that similarly exist in the true covariance.

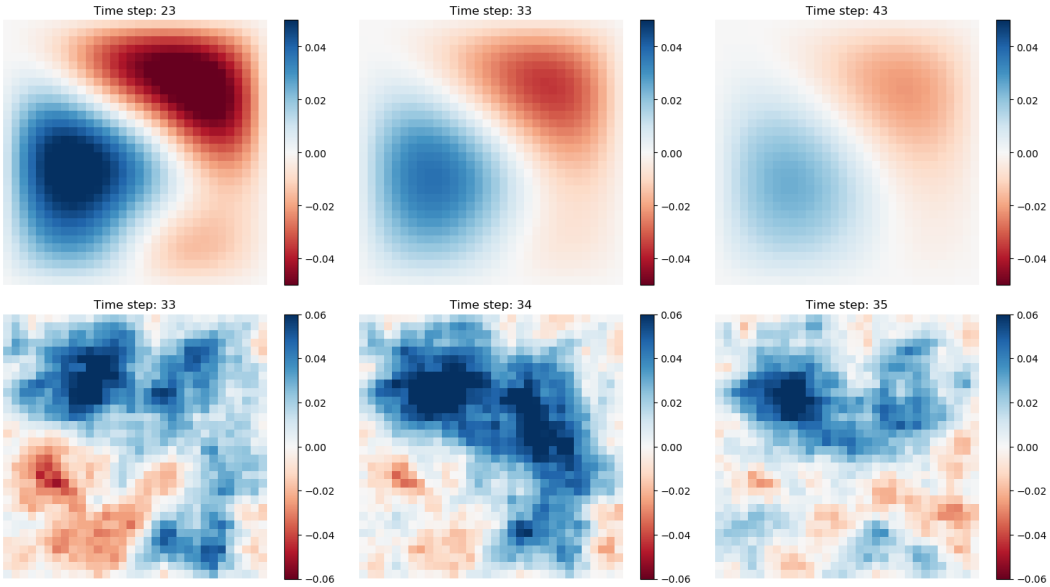
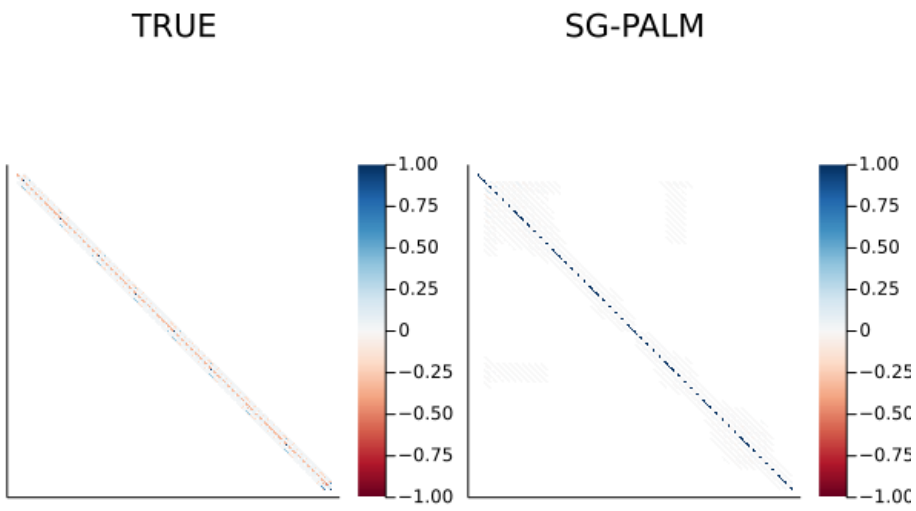
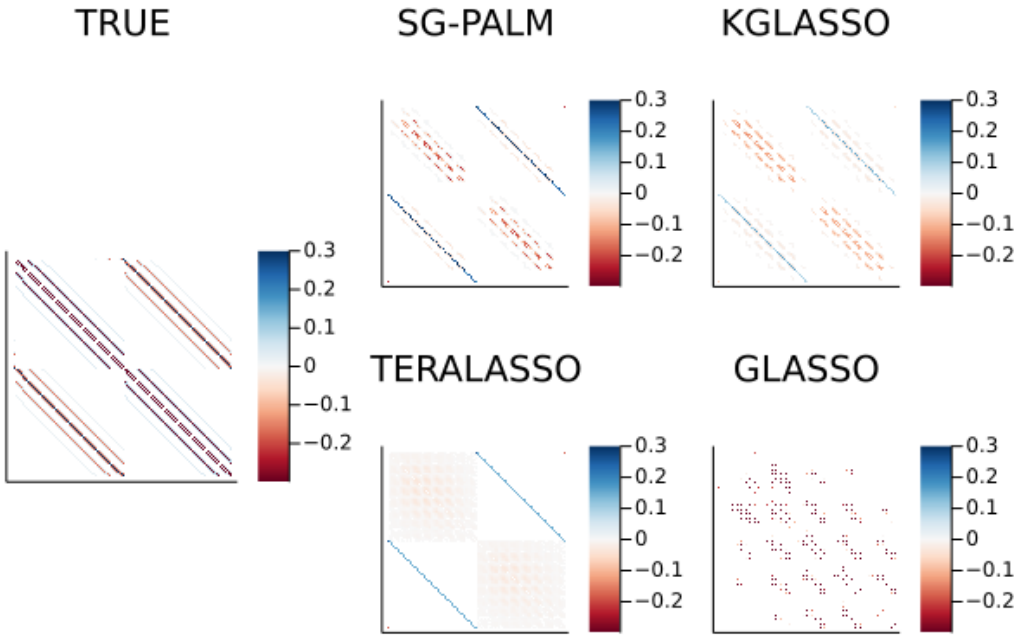


Figure C.1: 2D Convection-diffusion (top) and Poisson-AR(1) state variables at three different time steps.

Computational efficiencies of the various covariance/precision estimation algorithms are also vitally important in practice to facilitate real-time tracking of physical systems. Table C.2 shows the runtime of different covariance and inverse covariance estimation algorithms for the synthetic experiments. It shows that by recognizing and exploiting multiway structures in the data, sparse multiway inverse covariance estimation methods, TeraLasso, Tlasso, and SG-PALM significantly reduce the runtime complexity of Glasso that ignores such special multiway structures. Remark that KPCA runs considerably slower than other methods as it involves expensive singular

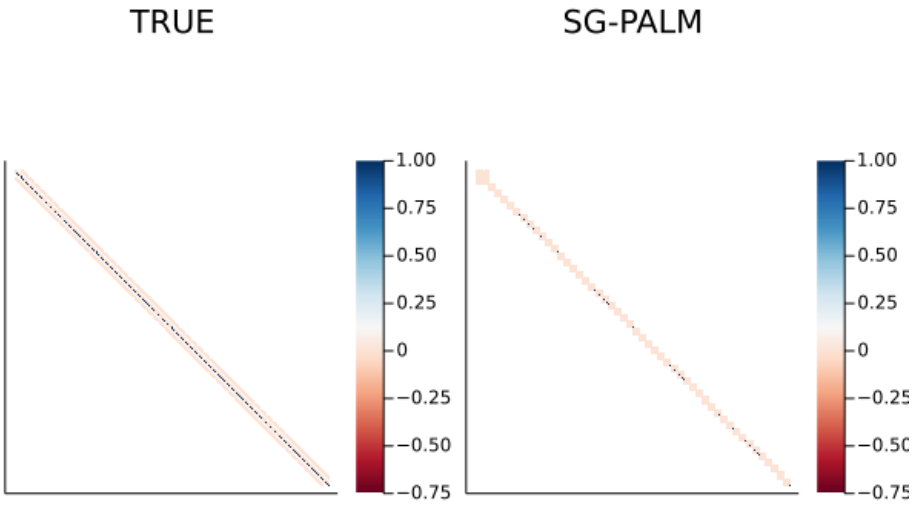


(a) Poisson-AR inverse covariance (left) and the SG-PALM estimate (right).

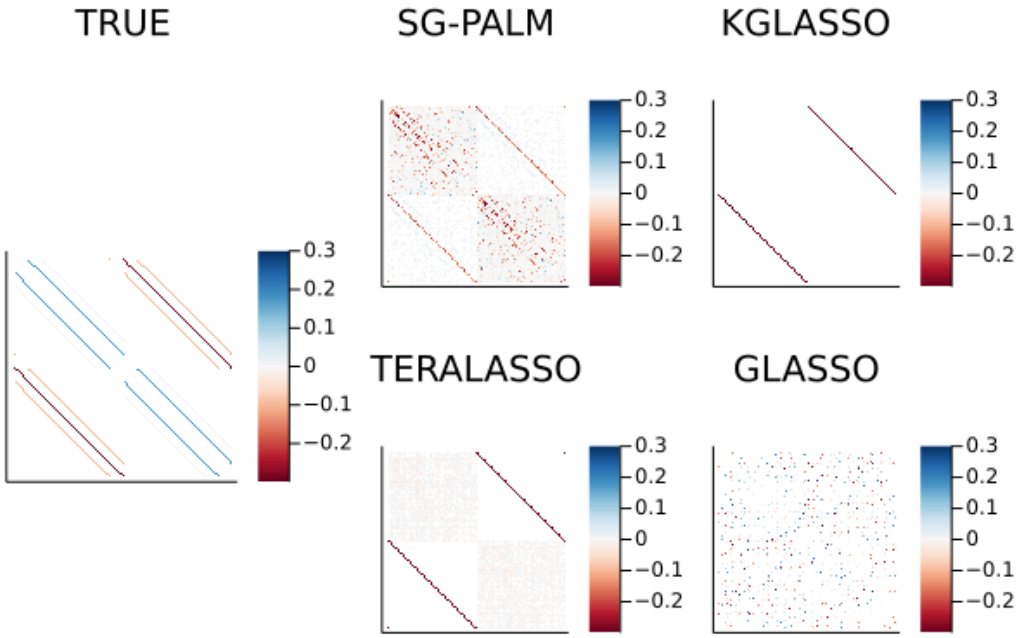


(b) Zoomed-in (middle 128 rows and columns) Poisson-AR inverse covariance structure (left) and the estimate obtained by SG-PALM, KGlasso, Glasso, TeraLasso (right, clockwise).

Figure C.2: Inverse covariance structures for Poisson-AR(1) and its estimates. Here, white entries indicate zeros in the inverse covariance matrices. The zoomed-in plots show two temporal blocks (each of size 64×64) of spatial inverse correlation structures with the diagonal elements removed for clearer visualization. SG-PALM and the associated Sylvester graphical model produce the richest structures.

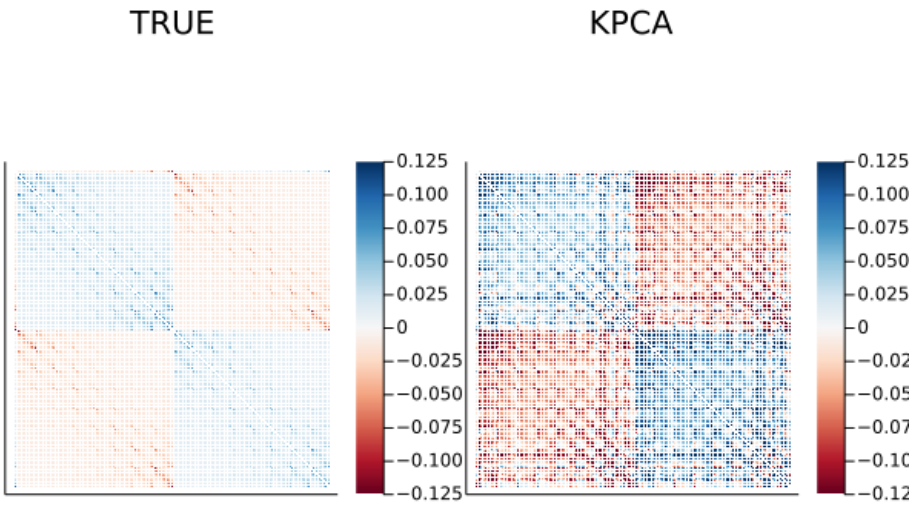


(a) Convection-diffusion inverse covariance (left) and the SG-PALM estimate (right).

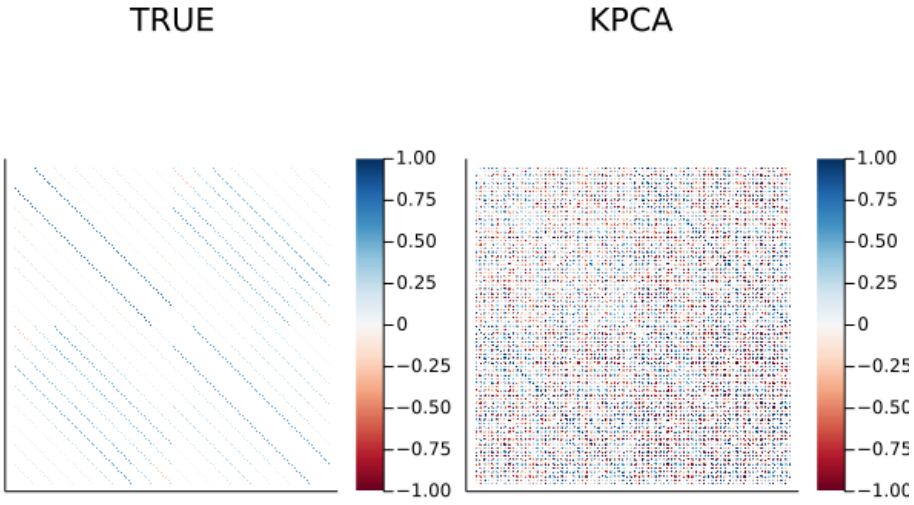


(b) Zoomed-in (middle 128 rows and columns) convection-diffusion inverse covariance (left) and the estimates by SG-PALM, KGlasso, Glasso, TeraLasso (right, clockwise).

Figure C.3: Inverse covariance structures for the Convection-Diffusion and its estimates. Here, white entries indicate zeros in the inverse covariance matrices. The zoomed-in plots show two temporal blocks (64×64) of spatial inverse correlation structures with the diagonal elements removed for clearer visualization. SG-PALM and the associated Sylvester graphical model produce the richest structures.



(a) Poisson-AR covariance structure (left) and the estimate obtained by KPCA (right).



(b) Convection-Diffusion covariance (left) and the estimate obtained by KPCA (right).

Figure C.4: Visualizations of the middle 128 rows and columns of the covariance structures for Poisson-AR(1) and Convection-Diffusion dynamics and their estimates, which show two temporal blocks of spatial correlation structures, each of size 64×64, with the diagonal elements removed for clearer visualization of the pattern. Here, white entries indicate zeros in the covariance matrices. Since the covariances are not sparse in general, all matrices are thresholded for clearer inspections of patterns.

Table C.1: Comparisons of performances measured by $\log\Big(\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F \backslash \|\mathbf{\Sigma}\|_F\Big)$ for KPCA as well as $\log\Big(\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}\|_F \backslash \|\mathbf{\Omega}\|_F\Big)$ and the Mathews Correlation Coefficient (MCC) for SG-PALM, Tlasso, TeraLasso, Glasso. The MCC is a measure of the quality of sparsity recovery considered as a binary classification problem, where ± 1 indicates perfect agreement or disagreement between the truth and the estimation. Here the Frobenius norm errors are included in the first row under each generating type while the MCCs are in the second row. Note that the best performers under each type/criteria are highlighted.

Type	Metric	**SG-PALM**	**KGlasso**	**TeraLasso**	**Glasso**	**KPCA**
P-AR	Fnorm	$\mathbf{-0.2622}$	1.1777	0.6312	0.9775	0.3289
	MCC	**0.4300**	0.3395	0.2061	0.0560	N/A
C-D	Fnorm	$\mathbf{-0.0420}$	1.4919	-0.0208	2.2041	0.0642
	MCC	**0.2122**	0.1884	0.2018	0.0349	N/A

value decomposition of a large-dimensional re-arranged sample covariance matrix of the data.

Table C.2: Runtime (in seconds) of estimating spatio-temporal (inverse) covariance matrices of size $d \times 50$, where d is varying, using various algorithms. Comparisons under various problem sizes (i.e., different d and N) are shown. Note the sparse multiway precision models (SG-PALM, KGlasso, TeraLasso) are comparably fast and are all faster than Glasso (for large problems) and KPCA.

d	N	**Glasso**	**SG-PALM**	**TeraLasso**	**KGlasso**	**KronPCA**
		sec	**sec**	**sec**	**sec**	**sec**
8^2	25	0.40(0.20)	0.46(0.15)	0.15(0.35)	0.65(0.11)	37.22(0.20)
	50	0.48(0.21)	0.47(0.08)	0.22(0.50)	0.70(0.10)	38.22(0.55)
	100	0.76(0.05)	0.44(0.13)	0.26(0.28)	0.69(0.30)	39.09(1.05)
16^2	25	6.43(1.45)	3.37(1.09)	5.38(0.58)	5.14(1.99)	495.47(2.69)
	50	9.12(0.98)	3.27(1.81)	4.62(1.98)	3.39(2.00)	516.64(2.19)
	100	11.84(2.01)	4.85(1.10)	6.71(0.72)	5.67(0.57)	498.04(4.01)

APPENDIX D

Appendix of Chapter V

4.1 Additional details of the Twitter latent Dirichlet allocation (T-LDA) algorithm

The generation processes for a T-LDA and an LDA are illustrated side-by-side in Figure D.1. Here, the key differences exhibited in T-LDA are aggregation (pooling tweets from users) and regularization (restricting a tweet to be generated from only one topic). In our study, the aggregation is done by pooling tweets generated from the same day.

All numerical results presented in the article were produced with the following implementation details of the T-LDA algorithm: the collapsed Gibbs sampler has been run for 2000 iterations with the first 1000 samples discarded as burn-in. The latent variable β is assumed to be symmetric Dirichlet with hyperparameter $\eta = 0.01$ for all topics; and θ is assumed to be symmetric Dirichlet with hyperparameter $\alpha = 0.5$ for all time stamps. Additionally, for weakly-supervised T-LDA implemented on the TalkLife data, additional weights are added to η such that $\eta_n = 0.01 + w_n$ for each seed word n and $\eta_n = 0.01$ otherwise, where the details of the weights can be found in Section 4.12.

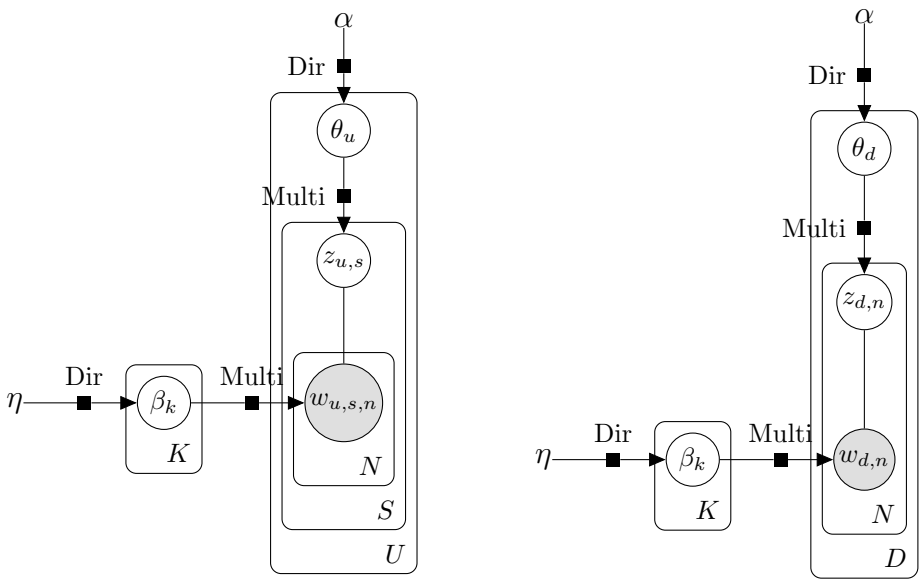


Figure D.1: Plate notation comparison for the Twitter Latent Dirichlet Allocation (T-LDA) (left) and the standard Latent Dirichlet Allocation (LDA) (right) models. Here nodes are random variables; edges indicate dependence through probability distributions (e.g., Dirichlet or multinomial). Shaded nodes are observed; unshaded nodes are latent. Plates indicate replicated variables. Note that the T-LDA model aggregates tweets from each user into a document and constrains each tweet to be drawn from only one topic.

4.2 Summary statistics characterizing shortest paths

To characterize the smoothness and continuity of the learned shortest paths of topics, we present summaries of the 'skips' (days where there are no topics connected to either a topic immediately before or after the current timestamp) they made. Table D.1 depicts the number of skips and the length of the skips for four topic paths (see Appendix 4.9 for details on the path names). We note that the length of a whole path (number of topics connected) could be different because 1) the different numbers of skips, and 2) the different time span as some topics appeared only for a certain time range (e.g., the wash hands topic). The lengths of those paths shown in the table are: COVID NEWS (presidential election), 70; COVID (health care), 58; STAY HOME (executive order), 59; SANITIZING (wash hands) 19. Clearly, longer paths could make longer skips. However, the paths remain fairly continuous (small numbers of short skips) during their time span. This is partly due to the corpora smoothing being applied–the topics learned at time t should usually be very similar to those learned at nearby timestamps.

Table D.1: Summary of the number of skips along with the length of those skips for four different topic paths. The paths are discovered by the shortest path algorithm using 10-nearest neighbor weighted graph. Note that all paths exhibit small numbers of short-length skips.

Path Name	Days Skipped			
	1	2	3	4
COVID (health care)	10	0	1	1
COVID NEWS (presidential election)	3	1	2	2
SANITIZING (wash hands)	1	1	0	0
STAY HOME (executive order)	4	0	0	0

4.3 Shortest path on MDS, ISOMAP, and PHATE

We desire a low-dimensional embedding that preserves the trajectory structures of shortest paths, so that we can visualize and interpret any results computed using methods described in Section 5.2.2. Here we compare PHATE with MDS and ISOMAP. MDS does not take any local structural information into account when building the embedding; ISOMAP applies MDS using shortest path distances computed on neighborhood graphs; finally, PHATE applies MDS on potential distances computed on neighborhood graphs while striking a balance between local and global trajectory structures. Figure D.2 shows that MDS failed to identify any path between two points. ISOMAP identifies a cleaner structure but there are interrupting background points on the path. PHATE identifies a clean path that is also well separated from background points. The comparison also highlights the importance of working with neighborhood graphs, instead of the fully connected graph, when trying to identify local structures in data.

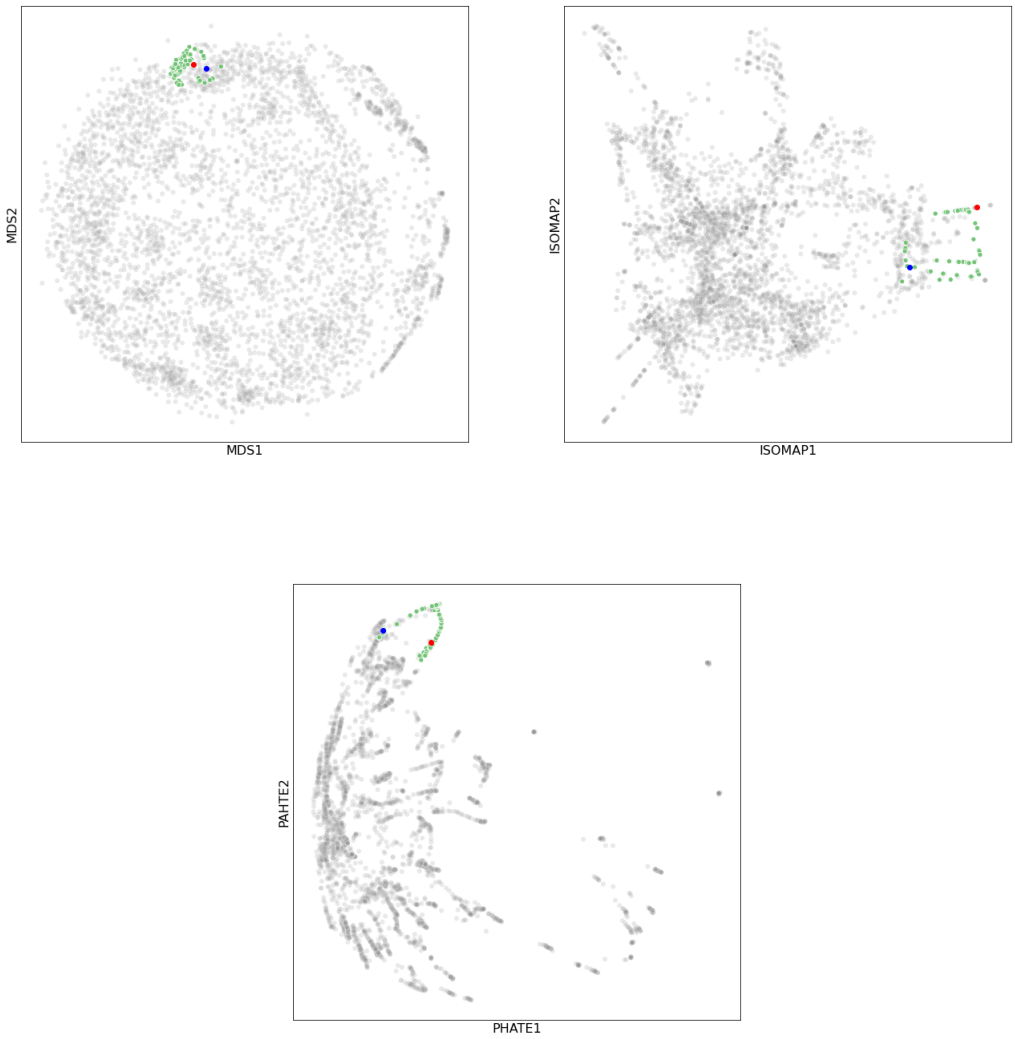


Figure D.2: Multidimensional scaling (MDS), isometric feature mapping (ISOMAP), and potential of heat-diffusion for affinity-based transition embedding (PHATE) for the same set of word distributions. A shortest path computed on 10 nearest neighbors graph is highlighted on each embedding with red and blue points indicating the starting and ending points of the path. Note that PHATE identifies the cleanest path connecting the red and blue points, with minimal background noises (grey points) included in between.

4.4 Detailed descriptions of PHATE

Algorithm D.1 outlines the steps for obtaining a low-dimensional embedding using PHATE with the Hellinger distance metric.

Algorithm D.1: PHATE with Hellinger distance

Input: N observations of some objects

1: Compute pairwise Hellinger distance matrix (denoted as D) from all pairs of multinomial topic distributions (stored as columns in a matrix X).
2: Compute k-nearest neighbor distance (denoted as $\epsilon_k(x)$) from each column of X.
3: Compute local affinity matrix $K_{k,\alpha}$ from D and ϵ_k.
4: Form a diffusion operator P, which is a Markov transition matrix computed by normalizing $K_{k,\alpha}$.
5: Compute time scale via Von Neumann Entropy. The time scale is then used to diffuse P to obtain P^t.
6: Compute potential representation of the diffusion matrix as $U_t = -\log(P^t)$ and compute potential distance matrix $D_{U,t}$ from U_t.
7: Apply MDS on $D_{U,t}$ to embed the data in lower dimension.

Output: An $N \times L$ matrix that contains L-dimensional coordinates for each observation.

In Algorithm D.1 we use the Hellinger distance to compute D and ϵ_k. This ensures that PHATE is being used to perform dimension reduction on a statistical manifold (Amari, 2012).

The PHATE construction is based on computing local similarities between data points, and then diffusing through the data using a Markovian random-walk diffusion process to infer more global relations. The local similarities between points are computed by first computing pairwise distances and then transforming the distances into similarities, via a kernel named the α-decaying kernel with locally adaptive bandwidth. It is defined as

$$K_{k,\alpha}(x,y) = \frac{1}{2}\exp\Big(-\Big(\frac{\|x-y\|}{\epsilon_k(x)}\Big)^{\alpha}\Big) + \frac{1}{2}\exp\Big(-\Big(\frac{\|x-y\|}{\epsilon_k(y)}\Big)^{\alpha}\Big). \tag{4.1}$$

Here the k-nearest neighbor distance ϵ_k is used to ensure that the bandwidth is locally

adaptive and varies based on the local density of the data. The exponent α controls the rate of decay of the tails in the kernel $K_{k,\alpha}$. Setting $\alpha = 2$ is equivalent to the use of a Gaussian kernel and choosing $\alpha > 2$ results in lighter tails in the kernel. The kernel is then normalized by row-sums that results in a row-stochastic matrix $P = P_{k,\alpha}$ (the diffusion operator), which is used for following steps.

In Step 5, the diffusion operator is powered by a time scale t. In particular, for a data point x and diffusion operator P, and let δ_x be the Dirac delta that is defined to be a row vector of length N (length of the data) with a one at entry corresponding to x and zero elsewhere. The t-step distribution of x is the row in P^t corresponding to x:

$$p_x^t := \delta_x P^t = [P^t]_{(x,\cdot)}. \tag{4.2}$$

This distribution captures multiscale (where t serves as the scale) local neighborhoods of data points, where the local neighborhoods are explored by randomly walking or diffusing over the intrinsic manifold geometry of the data. The scale parameter t affects the embedding. It can be selected based on any prior knowledge of the data or, as proposed in Moon et al. (2019), by quantifying the information in the powered diffusion operator with different values of t, via computing the Von Neumann Entropy (von Neumann, 2013; Anand et al., 2011) of the diffusion affinity, and choosing the one that explains the maximum amount of variability in the data.

Finally, a new type of distance, called the potential distance in Moon et al. (2019), is recovered in the end from the powered diffusion operator, which is obtained by taking the negative log of the transition probabilities. This transforms these transition probabilities into the heat-potential context.

4.5 Additional simulation studies for PHATE

To illustrate the idea of probability vectors on a sphere, in Figure D.3 we present a simple example of a sphere in 3D and probability vectors (simulated as in Section 5.2.3) lying on the sphere. The trajectories in this simulated example exhibit different progressive structures. In particular, the trajectory in dark blue evolves smoothly and remains roughly on the same path; the trajectory in brown exhibits a sharp turn in the direction at a certain position; finally, the trajectory in light blue behaves more chaotically and exhibits clustering structures. The PHATE embedding presented in Figure V.3 of Section 5.2.3 was able to uncover all these types of structures in low dimension.

To further demonstrate the advantage of PHATE over traditional methods for uncovering progressive structures, we present a similar example to that in Moon et al. (2019), which uses artificial tree-structured data and compare principle component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE) and PHATE in constructing low-dimensional embedding. In particular, we generate tree-structured data with 10 branches and 200 dimensions, and each branch has length 300. Thus, we have 3000 observations of 200-dimensional data, and the goal is to find a 2-dimensional embedding for visualization. Figure D.4 shows the results of embedding for three different methods. PCA is good for finding an optimal linear transformation that gives the major axes of variation in the data. However, the underlying data structure in this case is nonlinear in which case PCA is not ideal. t-SNE is able to embed nonlinear data; however, it is optimized for cluster structure and as a result will destroy any continuous progression structure in the data. PHATE for this example separates the clusters and is able to clearly represent the trajectory structure of the data. Additionally, PHATE neatly captures the branching/splitting points of different trajectories. This feature is vital for our study of tweeting behaviors as we are interested in learning how different conversations converge to a similar one or

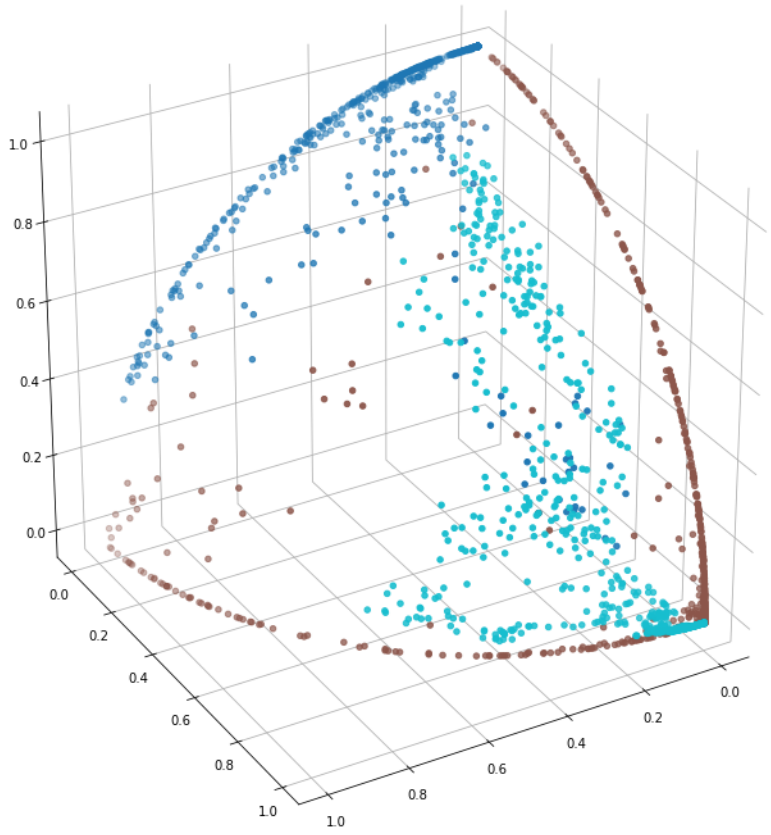


Figure D.3: Three simulated trajectories of probability vectors on a sphere. Each color signifies a trajectory simulated using a specific σ in the random-walk structure described in Section 5.2.3. Here, three trajectories started at the same point exhibit different progressive structures: stable (dark blue), chaotic and clustering (light blue), and sharp transition (brown).

diverge to different topics.

Additionally, we also demonstrate that PHATE does not 'create' spurious trajectories, although it does not preclude the existence of such structures. Here, 3000 independent data points were simulated from a 3-component (with weights $0.6, 0.3, 0.1$) 10-dimensional Gaussian mixture model and transformed through softmax (i.e., $z_j \rightarrow \frac{\exp(z_j)}{\sum_{i=1}^{10}\exp(z_i)}, j = 1, \ldots, 10$). Figure D.5 depicts 2-dimensional embedding computed by PCA, t-SNE, uniform manifold approximation and projection (UMAP), and PHATE using Hellinger distance. Clearly, PHATE did not artificially

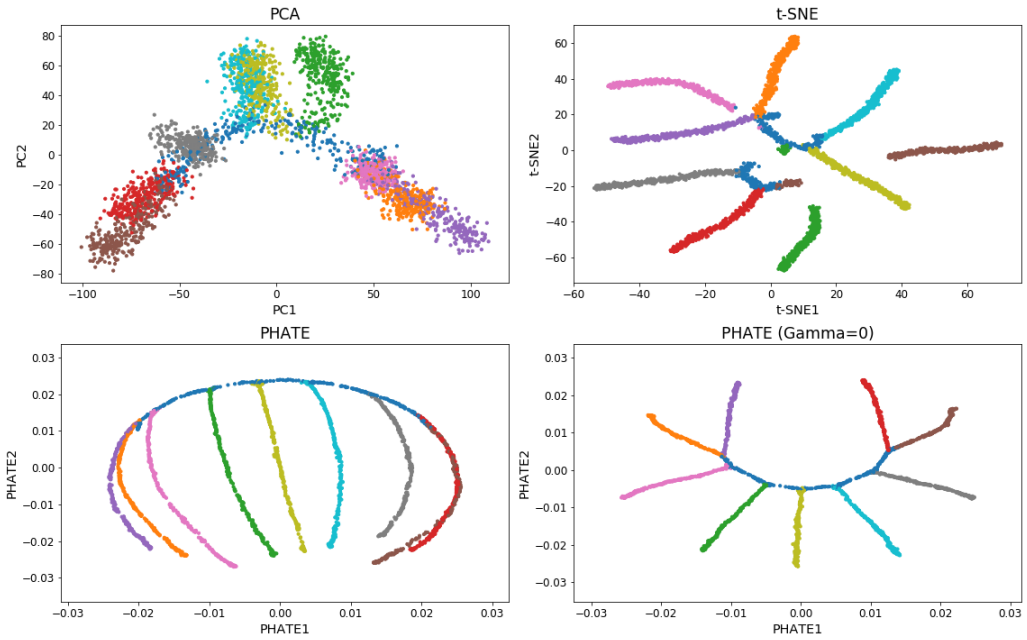


Figure D.4: Comparison of principle component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and potential of heat-diffusion for affinity-based transition embedding (PHATE). Two versions of PHATE with different tuning parameters are illustrated. The data are 3000 tree-structured observations with 10 branches. Various branches are colored differently. Note that for this truly trajectory-based data, PHATE gives the clearest low-dimensional representation of the data.

'trajectorize' the data; t-SNE seems to perform the best in terms of clustering as it often tries to separate data as much as possible; UMAP separated the clusters well but generated artificial segments and trajectories in the embedding.

Lastly, we compare PHATE (and other) embeddings using different distance metrics. In particular, we compute 2-dimensional embeddings for the data generated in Section 5.2.3 using Euclidean and cosine distances/similarities. Figure D.5 depicts the results comparing PCA, t-SNE, UMAP, and PHATE. It shows that the Hellinger metric (for t-SNE, UMAP, and PHATE) outperforms the other two in terms of generating the clearest low-dimensional embedding that preserves the true data geometry.

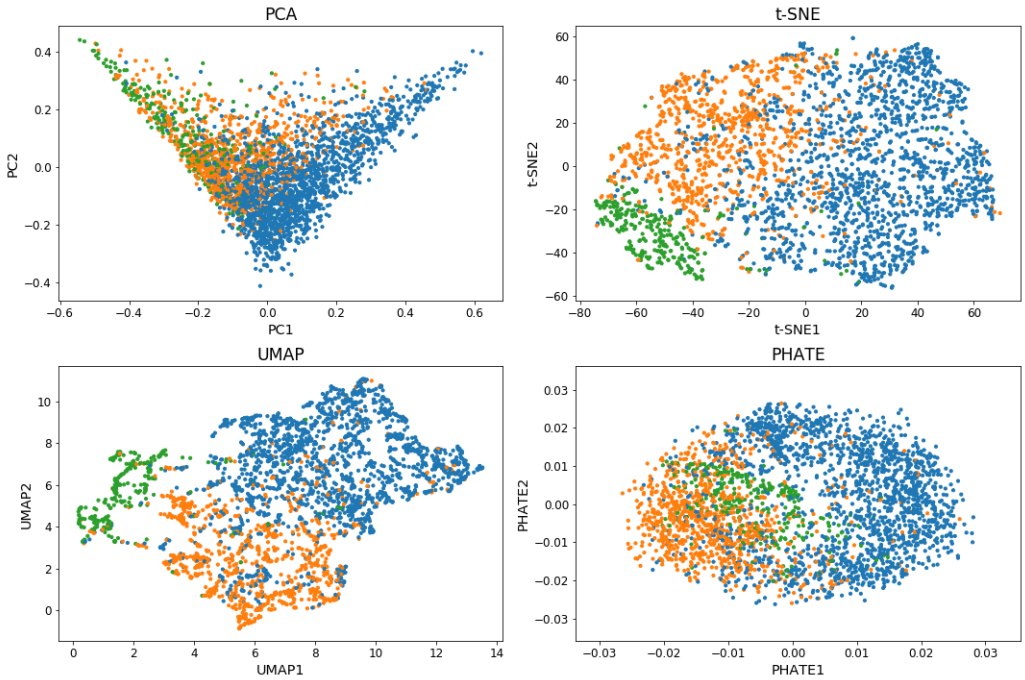
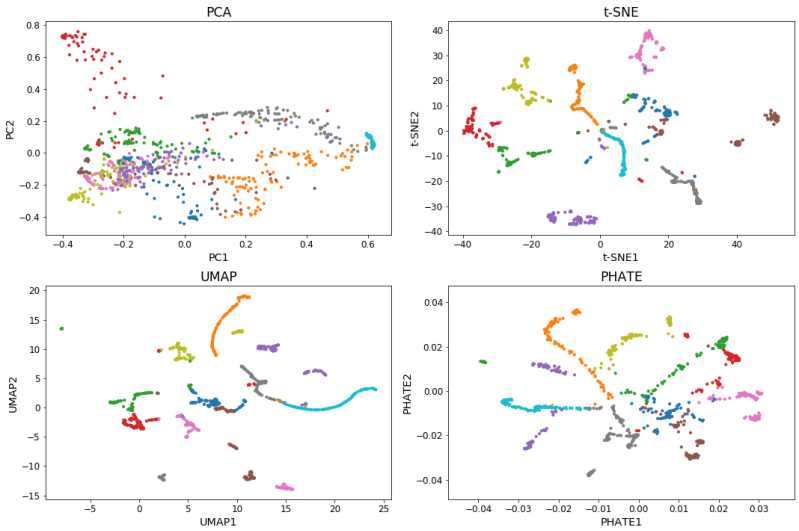
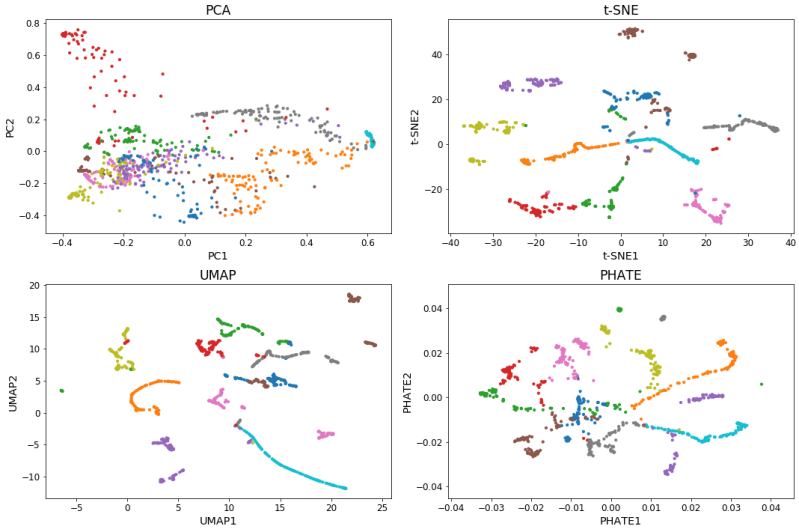


Figure D.5: Comparison of principle component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), uniform manifold approximation and projection (UMAP), and potential of heat-diffusion for affinity-based transition embedding (PHATE). Here $3,000$ independent data points were generated from a 3-component (with weights $0.6, 0.3, 0.1$) 10-dimensional Gaussian mixture model. Here, data were transformed via softmax to resemble a probability vector. Note that for this random nonstructured data, PHATE did not 'create' spurious trajectories in the low-dimensional embedding.



(a) Embedding using Euclidean metric.



(b) Embedding using Cosine metric.

Figure D.6: Principle component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), uniform manifold approximation and projection (UMAP), and potential of heat-diffusion for affinity-based transition embedding (PHATE) using Euclidean and cosine metrics. Here 10 trajectories of 100-dimensional probability vectors are generated, where the trajectories are colored differently. PHATE gives the clearest 2D representation of the inputs that preserves their high-dimensional progressive structures, regardless of the distance metric used. Comparing with Figure V.3, the Hellinger metric outperforms the other two metrics in recovering the data geometry.

4.6 Comparison with TopicFlow for topic trend mining

TopicFlow (Malik et al., 2013) is an analysis framework for Twitter data over adjacent time slices, binned topic models, and alignment, which is an application of LDA to timestamped documents at independent time intervals and alignment of the resulting topics. The key differences between TopicFlow and the proposed framework are: 1) a different similarity measure between topics, that is, cosine similarity metric for TopicFlow; 2) a different mechanism for topic alignment and connection–TopicFlow connects every pair of adjacent topics that has similarity above a certain threshold. The advantages of Hellinger metric over other metrics for comparing/embedding word distributions have been made clear in the previous section. Here, we demonstrate the advantages of the proposed shortest path mechanism over TopicFlow for obtaining natural temporal evolution of topics.

We analyze a particular topic cluster–the presidential election cluster discussed in Section 5.3–and compare the connections computed by the proposed shortest path algorithm and the TopicFlow algorithm. Here, for a fair and direct comparison, we fix the bins and the topic detection algorithms to be the same for both frameworks–using the smoothed temporal corpus and the T-LDA; the shortest path is performed on a 10-nearest neighbor weighted graph and the TopicFlow is performed with a connection threshold of 0.2. For the latter, we obtain a path by localizing the connection that has the largest cosine similarity at each pair of adjacent timestamp. For illustration, in Table D.2, we highlight a time segment that exhibits differences between two paths. In particular, the shortest path skipped 3 days, March 23 to March 26, while the TopicFlow remain continuously connected. The top row of Figure D.7 depicts the top word clouds of topics at timestamps March 23, 24, 27, and May 15 on the TopicFlow path. It shows a sharp transition from a voting/election topic to general political topics and finally to a relatively nonpolitical topic. On the other hand, the shortest path automatically skipped the timestamps where these new topics emerged

and maintained the major theme of the path, which is voting/election and later on general politics. This offers a more natural and much smoother transition.

This comparison demonstrates particularly that the mechanism for topic trend discovery used by TopicFlow is restrictive as it potentially results in nonsmooth and nonintuitive transitions. Although one could tune the connection threshold, it increases the computational burden and there is no obvious objective (e.g., prediction score, loss, etc.) that could help with the tuning process.

Table D.2: A portion of connected presidential election topics via the shortest path mechanism (left column) and the TopicFlow mechanism (right column). Here topics are indicated by their indices, e.g., $0-49$, at each timestamp (row index). *NA* indicates that no connection has been made by the algorithm.

	SP topic index	**TF topic index**
Feb 15	37	37
⋮	⋮	⋮
Mar 23	1	1
Mar 24	*NA*	44
Mar 25	*NA*	27
Mar 26	*NA*	26
Mar 27	22	26
⋮	⋮	⋮
May 15	2	15

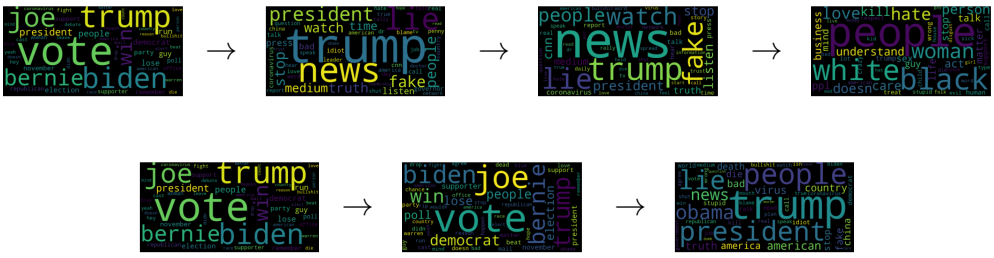


Figure D.7: Top word clouds showing evolution of topics on the presidential election topic paths computed via the shortest path algorithm (bottom) and the TopicFlow (top) algorithm. The sample timestamps at which the topics are learned are March 23, 24, 27, and May 15 (top); March 23, 27, and May 15 (bottom). Note that the shortest path algorithm produces much smoother and more intuitive transitions among topics within a general theme.

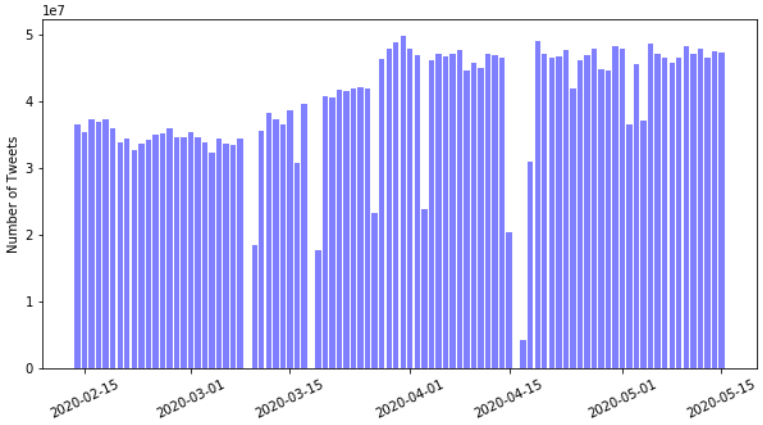
To further investigate the two different topic aligning methods, we fix the distance metric to be Hellinger, and compare the shortest path mechanism and the TopicFlow mechanism for the same set of topics. Table D.3 depicts a similar pattern for the time range March 23 to 27, where the restrictive TopicFlow mechanism for topic connection exhibits a sharp transition as shown in Figure D.7. Similar to Table D.2, from February to March 23, the two paths are mostly the same. However, we observe that the two paths also exhibit similar topics near the end of the time period. This again demonstrates the superiority of Hellinger distance for measuring topic similarity.

Table D.3: A portion of connected presidential election topics via the shortest path mechanism (left column) and the TopicFlow mechanism (right column) using the same distance metric (Hellinger). Here topics are indicated by their indices, e.g., $0-49$, at each timestamp timestamps (row index). *NA* indicates that no connection has been made by the algorithm. Note that the restriction imposed by TopicFlow impacts the topic path similar (from March 23 to 27) to that in Table D.2

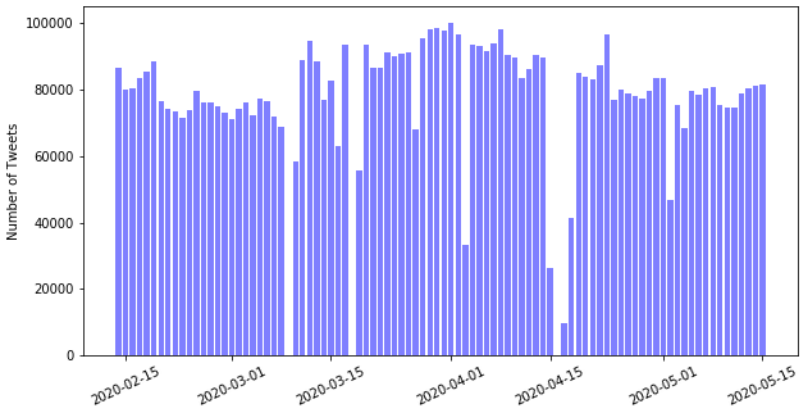
	SP topic index	**TF topic index**
Feb 15	37	37
⋮	⋮	⋮
Mar 23	1	1
Mar 24	NA	44
Mar 25	NA	27
Mar 26	NA	26
Mar 27	22	26
⋮	⋮	⋮
May 12	8	8
May 13	42	42
May 14	0	0
May 15	2	2

4.7 Volume plots of raw Twitter Decahose data

Figure D.8b shows the Decahose Twitter volume plots before (top) and after (bottom) processing. Although Twitter officially claims the percentage of geotagged tweets to be around 1-2% of the total tweets (`https://developer.twitter.com/en/docs/tutorials/Tweet-geo-metadata`), we found the percentage to much smaller. Note that there are several time points where the data is either incomplete (i.e., low volumes) or missing (i.e., 0 volumes).



(a) Raw Decahose tweets volume (on a scale of 10^7 tweets) from Feb 15 to May 15.



(b) Geotagged U.S., non-retweet, English Decahose tweets volume from Feb 15 to May 15.

Figure D.8: Volume of all and geotagged Decahose tweets for each day during the study period. The Decahose stream generates around $30-50$ million raw tweets and $50-100$ thousand geotagged English language tweets per day, except for several missing/incomplete cases with 0 or abnormally small volumes.

4.8 Sensitivity analysis for hyperparameters

In this section, we perform sensitivity analyses for the hyperparameters k and γ in Algorithm V.2, namely the number of neighbors in the nearest neighbor graph and the smoothing parameter for constructing new corpora. Further, we perform model selection for varying choices of K, the number of topics in T-LDA.

In Figure D.9, the two shortest paths computed using neighborhood graphs of $k = 8$ and 12 are illustrated. For comparison, the same starting and ending topics as well as the two intermediate topics at the same time points as those in Figure V.2 are used. It is clear from the word clouds that the shortest paths are not sensitive to the choice of k in the neighborhood of 10.

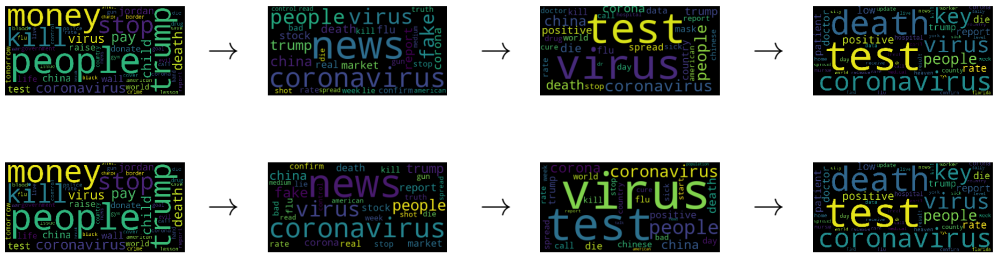


Figure D.9: Evolution along the shortest paths of a COVID-19 topic on the first day to a COVID-19 health care focused topic on the last day illustrated as top word clouds. The paths are computed on a 8- (top) and a 12- (bottom) nearest neighbor graph. The middle two word clouds are illustrations of two of the topics on the paths at the same time points as those in Figure V.2. Note that the intermediate topics in both cases represent natural transformations from the beginning to the end topics, confirming that the shortest path is not sensitive to small perturbations of k around 10.

Additionally, we quantify the similarities between any two shortest paths computed on different neighborhood graphs by computing the average Hellinger distance between topics (at the same time point) on the paths. Particularly, in Table D.4 we show the average Hellinger distances. For this particular cluster of topics, the average Hellinger distances are negligible and are stable across all pairs of different paths, which suggests that the shortest path is not sensitive to different k in the

neighborhood of $k = 10$.

Table D.4: Average Hellinger distances between any two topics paths generated using various neighborhood parameters k as the column/row indices. Examples are shown for the COVID (health care) topics. Note that the average Hellinger distances are identically 0 across all pairs of paths, indicating that the shortest paths are stable under different choices of k.

	8	**10**	**12**
8	0	0	0
10	0	0	0
12	0	0	0

Figure D.10 shows the contributions (in terms of the number of tweets) from each document to the temporally smoothed corpus constructed for March 31, using smoothing parameters of $0.65, 0.75, 0.85$. With 0.75, the contents span the whole study period (Feb 15 to May 15) but concentrate on tweets within a month, centered at March 31.

Moreover, Figure D.11 shows the PHATE embedding of all topics learned by T-LDA, using corpus constructed with smoothing parameters 0.65 and 0.85. Here we highlight two clusters (COVID and COVID NEWS) and one shortest path (presidential election) similar to Figure V.4. Comparing the three PHATE plots, the overall structures are similar and the highlighted trajectories remain relatively stable (i.e., presidential election paths exhibit similar 'U' shapes in all cases). Note that the 'split-and-merge' behaviors within the COVID NEWS cluster are being captured in all cases as well. The only notable difference in the PHATE produced with different temporal smoothing is the length of the trajectories, with those in the embedding produced using smoothing parameter 0.85 being the longest. This is reasonable because a larger smoothing parameter assumes a longer range temporal dependence structure of the data.

Additionally, we quantify the similarities between any two shortest paths from different smoothed corpora by computing the average Hellinger distance of the topics

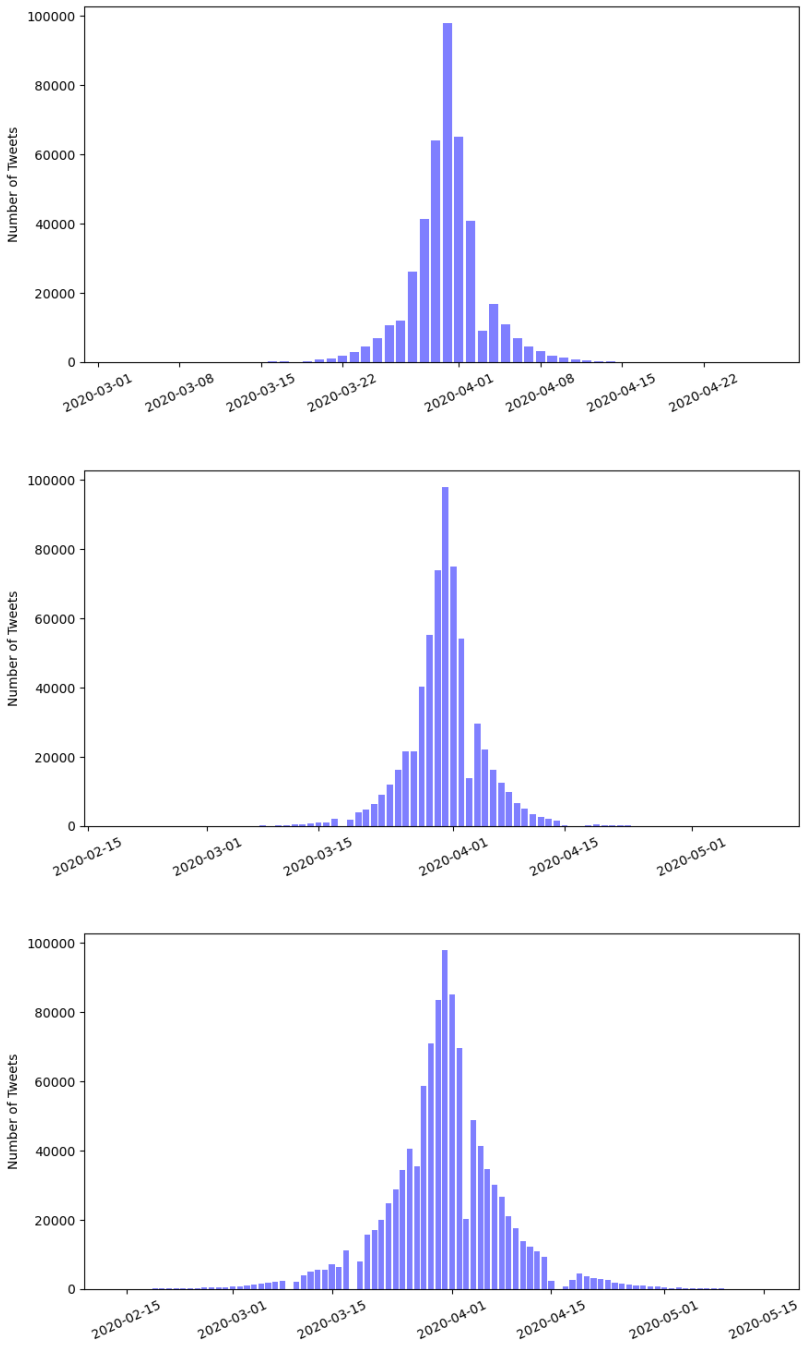


Figure D.10: Contributions of tweet volume from various time points for temporally smoothed corpora. The examples are constructed for March 31, using smoothing parameters $0.65, 0.75, 0.85$ (from top to bottom). Although the plots exhibit different resolutions and spans of the histograms, the shapes of the contribution distributions are similar in all cases. This illustrates robustness of the proposed method to the choice of smoothing parameters.

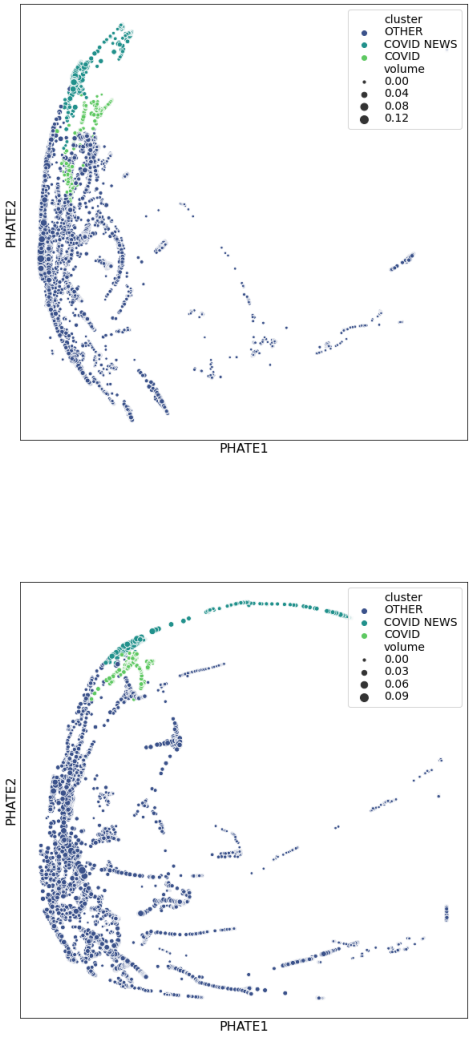


Figure D.11: Potential of heat-diffusion for affinity-based transition embedding (PHATE) for all word distributions. The topics here are learned by T-LDA on tweet collections constructed with smoothing parameters 0.65 (top) and 0.85 (bottom). Here two clusters and one shortest path are highlighted for comparison with Figure V.4. Note that the overall structures as well as the trajectories for highlighted points are similar in all three cases, while the lengths of the trajectories are different, which are the result of different assumptions on the range of the temporal dependence (i.e., a smoothing using 0.85 assumes longer range dependence by including more old tweets).

on the paths. Particularly, in Table D.5 we show the average Hellinger distances between any two paths computed under different smoothing conditions, for the COVID

NEWS (presidential election) and COVID (health care) topics. In these two cases, the average Hellinger distances are around 0.35 and are stable across all pairs, which suggests that the shortest paths of key topics of interest are not sensitive to different smoothing parameters.

Table D.5: Average Hellinger distances between any two topics paths generated from corpora with various smoothing parameters as the column/row indices. Examples are shown for the COVID NEWS (presidential election) and the COVID (health care) topics in the top and bottom tables, respectively. Note that the average Hellinger distances are both relatively small and stable in the sense that all pairwise distances are similar in magnitude, indicating that the shortest paths are stable under different choices of smoothing parameters.

	0.65	**0.75**	**0.85**
0.65	0	0.3520	0.3578
0.75	0.3520	0	0.3056
0.85	0.3578	0.3056	0

	0.65	**0.75**	**0.85**
0.65	0	0.3697	0.4112
0.75	0.3697	0	0.3652
0.85	0.4112	0.3652	0

Lastly, for the choice of the number of topics for T-LDA, we propose to compute a Bayesian Information Criteria (BIC) score at each timestamp defined as

$$-\text{log-likelihood} + \frac{C \log(D)}{2}$$

where the model complexity is computed by $C := Kp + (K-1)D$ with p denoting the length of the vocabulary. The log-likelihood of the T-LDA model is defined as

$$\prod_{k=1}^{K} \text{Dirichlet}(\beta_k; \eta) \times \prod_{d=1}^{D} \text{Dirichlet}(\theta_d; \alpha)$$
$$\times \prod_{s=1}^{S_d} \text{Categorical}(z_{s,d}; \theta_d) \times \prod_{n=1}^{N_s} \text{Categorical}(w_{n,s,d}; \beta_{z_{s,d}}).$$

Here, the categorical distribution is a special case of the multinomial distribution, in that it gives the probabilities of potential outcomes of a single drawing rather than multiple drawings; S_d denotes the number of tweets in document d ; and N_s denotes the number of words in a tweet s. This criteria is similar to the topic model selection criteria proposed in Taddy (2012).

In Figure D.12, we show the computed scores across all timestamps for various choices of the numbers of topics. The model with $K = 50$ consistently produces the lowest scores for the first half of the time range and is comparable to the model with $K = 100$ for the second half.

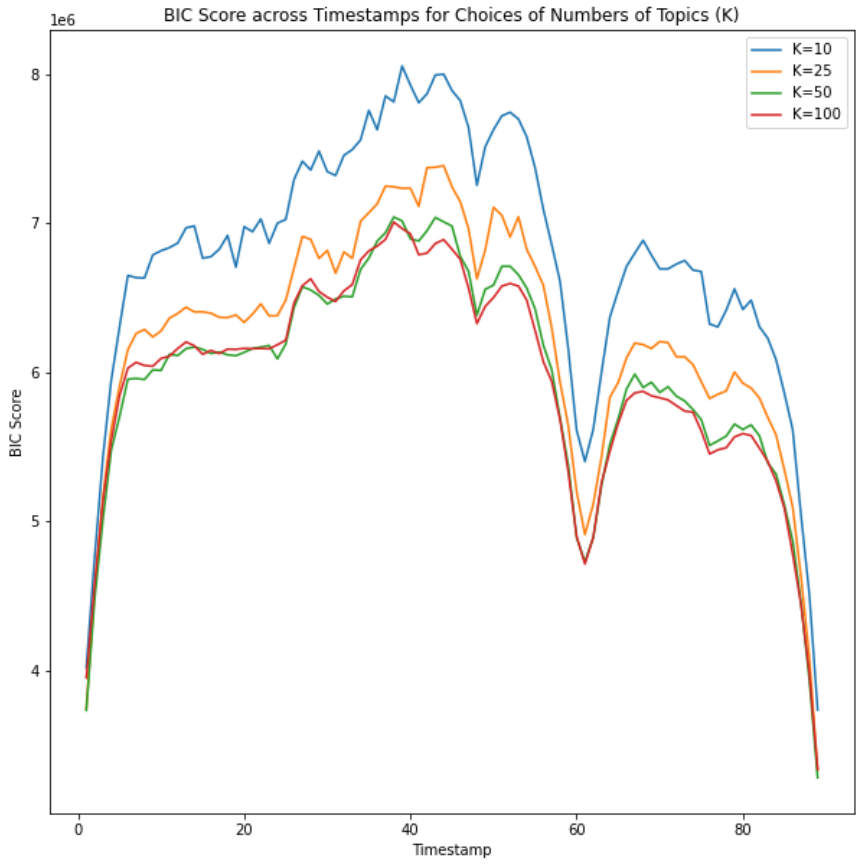


Figure D.12: Bayesian information criteria (BIC) scores across timestamps for different choices of the numbers of topics.

4.9 PHATE dictionary of clusters and trajectories

We explain the labeling of the PHATE plots for visualization and interpretation:

- Colors signify clusters of topics. Clusters are computed by a hierarchical clustering algorithm using Hellinger distance between topics. Only selected COVID-19 topics are colored differently, and all others are grouped into a single color. Selected COVID-19 topics are:
  - COVID: topics where the top words are mostly general COVID-19 terms such as coronavirus, virus, covid, etc.
  - COVID NEWS: topics where the top words are related to government officials or politicians discussing COVID-19 related issues. Typical top words include: Trump, government, news, covid, etc.
  - SANITIZING: topics where the top words are mostly wash hands, sanitizing, virus, etc.
  - STAY HOME: topics where the top words are mostly stay home, safe, covid, etc.
- Sizes represent normalized number of tweets that is generated from each topic.
- Shapes highlight selected COVID-19 related shortest paths computed on the neighborhood graph. Different shapes represent
  - COVID (health care): a subset of topics in the COVID topic cluster that are all on a shortest path starting from a general COVID topic at the first time point and finishing at a health care focused COVID topic (e.g., testing, death).
  - COVID (politics): a shortest path that starts from a topic that is second-closest in distance to the starting topic of the COVID (health care) and finishing at a politics focused COVID topic (e.g., president, news, etc.).

- COVID NEWS (presidential election): a subset of topics in the COVID NEWS cluster that are all closely related to presidential election and are on a shortest path starting from a election-related topic at the first time point.
- SANITIZING (wash hands): a subset of topics in the SANITIZING cluster that are on a shortest path starting from a topic related to washing hands due to COVID-19.
- STAY HOME (executive order): a subset of topics in the STAY HOME cluster that are on a shortest path starting from a topic related to stay home executive order due to COVID-19.
- General: topics that are not on selected shortest paths of interest.

4.10 Additional PHATE trajectories

Figure D.13 shows two linear trajectories, the SANITIZING (wash hands) and the STAY HOME (executive order), on the PHATE embedding. In contrast to nonlinear trajectories presented in Figure V.5 and Figure V.7, topics on linear trajectories exhibit no obvious deviation in terms of the top words.

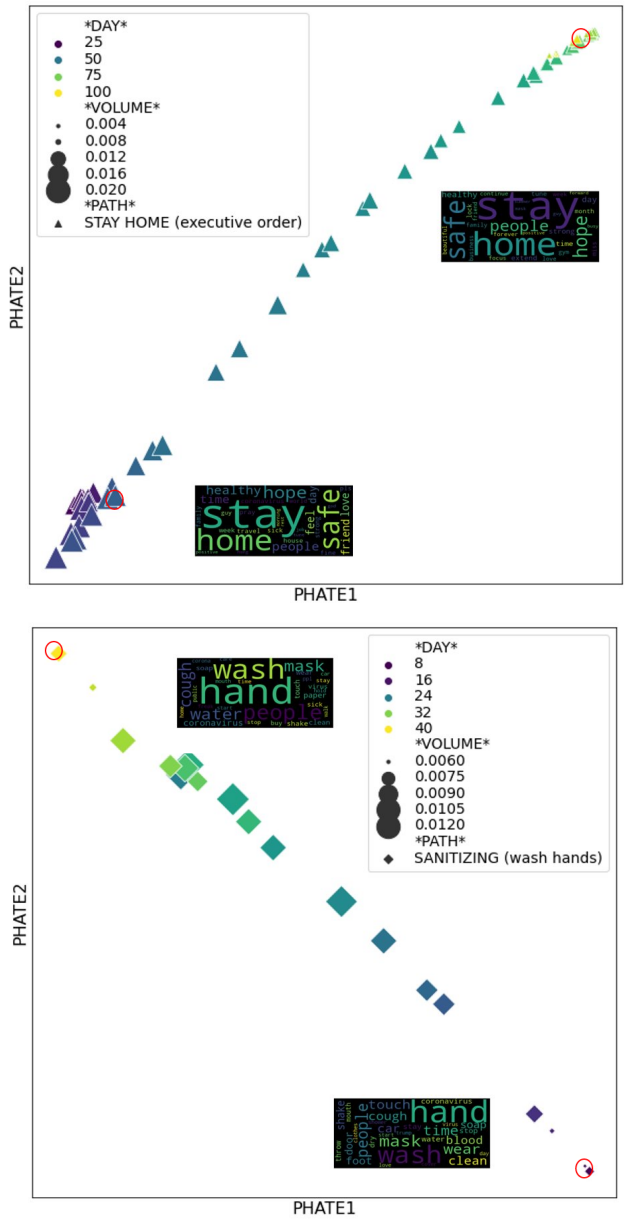


Figure D.13: Potential of heat-diffusion for affinity-based transition embedding (PHATE) for subsets of topics lie on the executive order path (top) and the wash hands path (bottom). Colors and sizes of points highlight time and tweet volume, respectively. Here two word clouds containing top 30 words in corresponding topics are shown for the time points highlighted by red circles in each path. Note that in both cases, the topic near the beginning of the study period is similar to that near the end of the study period. This shows the stability of topics on linear trajectories.

4.11 State-level trend in Tweet proportions

Here, we illustrate state-level variations in estimated tweet volumes generated by topics on the presidential election path, normalized by total tweet volume at each time point. From Figure D.14, we see that although tweet proportions vary state by state, the overall trend is clear with peaks roughly correspond to the time points of key event highlighted.

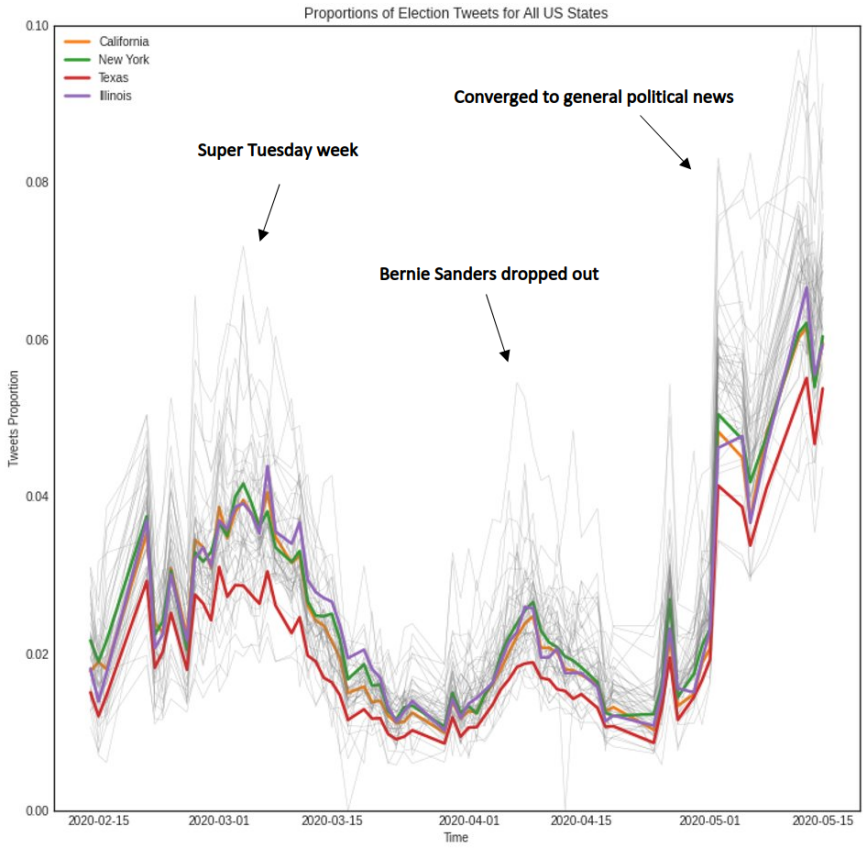


Figure D.14: State-level spatial distribution of tweet proportions generated from all topics on the COVID NEWS (presidential election) path. California, New York, Texas, and Illinois are highlighted for illustration, while all other states are ll plotted in grey. Note that similar three events (annotated using texts) as in Figure V.5 correspond roughly to the three peaks in the time-course plot, indicating validations of the quality of the shortest path using real-world events.

For the COVID (health care) topic path, at the state level, tweet proportions follow global trends at the beginning of the study period in February and March but

become chaotic starting in April. One possible explanation is that the COVID-19 pandemic in the United States started in several hot spots but quickly spread into other states, which then started to implement state-specific control measures. In addition, the overall new cases and death toll in the country reached a few record highs in April, starting with New York, which became an epicenter of the pandemic, with a record 12274 new cases reported on April 4 (`https://en.wikipedia.org/wiki/COVID-19_pandemic_in_New_York_(state)`). This explains the difference in tweet proportions trend in New York, compared with the other three highlighted states.

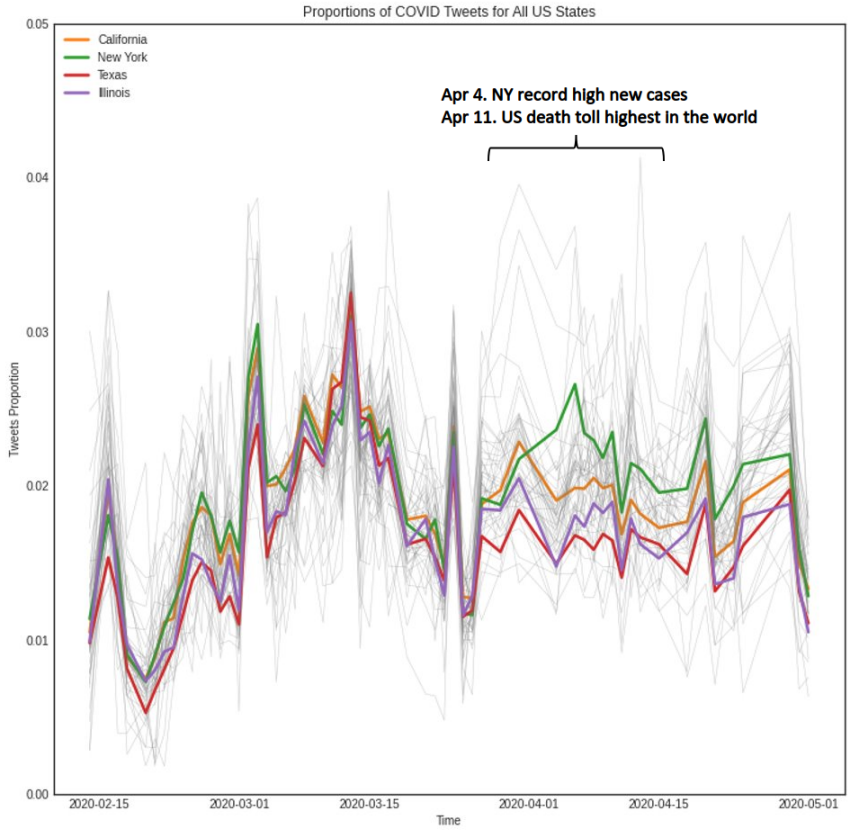


Figure D.15: State-level spatial distribution of Tweet proportions generated from all topics on the COVID (health care) path. California, New York, Texas, and Illinois are highlighted for illustration, while all other states are plotted in grey. Note that a time period in April is annotated with relevant events explaining the surge in tweet proportions in many states. This validates the quality of this shortest path using real-world events.

4.12 Additional details of TalkLife data

Table D.6: Label names and corresponding percentage volume in all posts generated in 2019. Note the label "Other" indicates a post is not labeled by any other labels.

Label	**Percentage volume**
Other	0.5840
DepressedMoodSuspected	0.0799
AgitationOrIrritationSuspected	0.07421
LonelinessSuspected	0.0682
FamilyIssuesSuspected	0.0674
BehavorialSymptomsSuspected	0.0618
DistortedThinkingSuspected	0.0607
AnxietyPanicFearSuspected	0.0566
SelfHarmSuspectedTakeTwo	0.0506
BodyImageEatingDisordersSuspected	0.0495
SuicidalIdeationAndBehaviorSuspected	0.0433
NumbnessEmptinessSuspected	0.0371
NssiIdeationAndBehaviorSuspected	0.0327
SelfHarmRelapseSuspected	0.0325
TiredFatiguedLowEnergySuspected	0.0257
MentalHealthTreatmentSuspected	0.0241
CryingSuspected	0.0229
DeathOfOtherSuspected	0.0201
AlcoholAndSubstanceAbuseSuspected	0.0191
HelplessnessHopelessnessSuspected	0.0163
SelfHarmRemissionSuspected	0.0126
EmotionalExhaustionSuspected	0.0110
FinalTiredFatiguedLowEnergySuspected	0.0105
FailureSuspected	0.0094
SongLyricsSuspected	0.0078
SuicidalPlanningSuspected	0.0078
InpatientOutPatientMedicationSuspected	0.0059
EmptinessSuspected	0.0054
NumbnessSuspected	0.0050
SuicideAttemptSuspected	0.0017
NssiUrgeSuspected	0.0017
NauseaWithEatingDisorderSuspected	0.0015
NauseaSuspected	0.0012
SelfHarmRemissionOrRelapseSuspected	0.0011

Table D.7: Seed words and the associated weights that are used in the weakly-supervised LDA algorithm. Weights are computed as natural log of the volume (number of occurrences) of the corresponding word in the entire year of 2019, multiplied by a tune-able constant (equals 10 here).

Word	**Weight**	**Word**	**Weight**
afraid	101.72	listen	105.46
anxiety	106.67	live	113.41
anymore	112.89	lonely	108.59
attempt	86.54	long	111.18
band	80.62	lose	112.61
body	104.47	lyric	79.44
clean	93.92	medication	86.72
cry	112.81	mental	101.90
cut	104.78	mom	109.04
dad	103.89	month	108.48
dead	102.16	numb	93.79
death	98.71	pain	110.16
depression	105.84	parent	105.64
die	114.56	plan	97.73
drink	99.31	play	104.71
drug	93.76	pretty	103.28
drunk	92.33	relapse	81.82
eat	107.72	sad	111.77
emptiness	77.17	scar	106.90
empty	95.84	school	109.36
end	113.06	sick	103.66
energy	93.67	smoke	94.36
exhaust	91.71	song	102.18
eye	104.10	stop	114.58
fail	96.32	suicide	99.23
failure	90.22	tear	97.94
family	110.45	throw	97.05
favorite	96.82	tire	108.81
feeling	112.08	tired	101.14
food	99.55	ugly	100.59
harm	96.07	urge	86.28
health	97.70	vomit	71.62
heart	112.15	weak	92.34
hospital	91.96	week	107.86
hurt	115.45	worry	101.11
kill	108.86	write	101.68
leave	115.83		

4.13 Additional details of clustering of TalkLife labels

Table D.8: Clustered labels.

Cluster	Labels
1	BehavorialSymptomsSuspected, CryingSuspected, DepressedMoodSuspected
2	InpatientOutPatientMedicationSuspected, MentalHealthTreatmentSuspected
3	EmptinessSuspected, NumbnessSuspected
4	AgitationOrIrritationSuspected, AlcoholAndSubstanceAbuseSuspected, AnxietyPanicFearSuspected, BodyImageEatingDisordersSuspected, DeathOfOtherSuspected, FamilyIssuesSuspected, NssiIdeationAndBehaviorSuspected, SelfHarmRelapseSuspected, SelfHarmRemissionSuspected, SelfHarmSuspectedTakeTwo, SuicidalIdeationAndBehaviorSuspected
5	FinalTiredFatiguedLowEnergySuspected, TiredFatiguedLowEnergySuspected
6	NauseaSuspected, NauseaWithEatingDisorderSuspected
7	SuicidalPlanningSuspected, SuicideAttemptSuspected
8	DistortedThinkingSuspected, EmotionalExhaustionSuspected, FailureSuspected, HelplessnessHopelessnessSuspected
9	NssiUrgeSuspected, SelfHarmRemissionOrRelapseSuspected, SongLyricsSuspected
10	LonelinessSuspected, NumbnessEmptinessSuspected

Figure D.16: Top words visualization of the sparse word distributions of each label before clustering.

BIBLIOGRAPHY

BIBLIOGRAPHY

Ahmed, N. K., J. Neville, and R. Kompella (2013), Network sampling: From static to streaming graphs, *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *8*(2), 1–56.

Amari, S.-I. (2012), *Differential-geometrical methods in statistics*, vol. 28, Springer Science & Business Media.

Amir, S., M. Dredze, and J. W. Ayers (2019), Mental health surveillance over social media with digital cohorts, in *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pp. 114–120.

Anand, K., G. Bianconi, and S. Severini (2011), Shannon and von Neumann entropy of random networks with heterogeneous expected degree, *Physical Review E*, *83*(3), Article 036,109.

Anandkumar, A., R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky (2014), Tensor decompositions for learning latent variable models, *Journal of Machine Learning Research*, *15*, 2773–2832.

Anderson, J. (2009), Spatially and temporally varying adaptive covariance inflation for ensemble filters, *Tellus A: Dynamic meteorology and oceanography*, *61*(1), 72–83.

Anderson, J. L. (2007), An adaptive covariance inflation error correction algorithm for ensemble filters, *Tellus A: Dynamic meteorology and oceanography*, *59*(2), 210–224.

Arora, S., R. Ge, and A. Moitra (2012), Learning topic models – Going beyond SVD, in *2012 IEEE 53th Annual Symposium on Foundations of Computer Science*, pp. 1–10, IEEE Computer Society, Los Alamitos, CA, USA.

Attouch, H., and J. Bolte (2009), On the convergence of the proximal algorithm for nonsmooth functions involving analytic features, *Mathematical Programming*, *116*(1-2), 5–16.

Bai, L., Y. Bi, B. Yang, J.-C. Hong, Z. Xu, Z.-H. Shang, H. Liu, H.-S. Ji, and K.-F. Ji (2021), Predicting the evolution of photospheric magnetic field in solar active regions using deep learning, *Research in Astronomy and Astrophysics*, *21*(5), 113.

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

Banerjee, A., I. S. Dhillon, J. Ghosh, S. Sra, and G. Ridgeway (2005), Clustering on the unit hypersphere using von mises-fisher distributions., *Journal of Machine Learning Research*, *6*(9).

Barzilai, J., and J. M. Borwein (1988), Two-point step size gradient methods, *IMA journal of numerical analysis*, *8*(1), 141–148.

Batmanghelich, K., A. Saeedi, K. Narasimhan, and S. Gershman (2016), Nonparametric spherical topic modeling with word embeddings, in *Proceedings of the conference. Association for computational linguistics. Meeting*, vol. 2016, p. 537, NIH Public Access.

Belkin, M., and P. Niyogi (2004), Semi-supervised learning on Riemannian manifolds, *Machine Learning*, *56*(1-3), 209–239.

Bernstein, M., V. D. Silva, J. C. Langford, and J. B. Tenenbaum (2000), Graph approximations to geodesics on embedded manifolds.

Besag, J. (1977), Efficiency of pseudolikelihood estimation for simple gaussian fields, *Biometrika*, pp. 616–618.

Bhadury, A., J. Chen, J. Zhu, and S. Liu (2016), Scaling up dynamic topic models, in *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pp. 381–390, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE.

Bickel, P. J., and A. Chen (2009), A nonparametric view of network models and newman–girvan and other modularities, *Proceedings of the National Academy of Sciences*, *106*(50), 21,068–21,073.

Bishop, C., and D. Hodyss (2009a), Ensemble covariances adaptively localized with eco-rap. part 1: Tests on simple error models, *Tellus A: Dynamic Meteorology and Oceanography*, *61*(1), 84–96.

Bishop, C., and D. Hodyss (2009b), Ensemble covariances adaptively localized with eco-rap. part 2: A strategy for the atmosphere, *Tellus A: Dynamic Meteorology and Oceanography*, *61*(1), 97–111.

Bishop, C. H., B. J. Etherton, and S. J. Majumdar (2001), Adaptive sampling with the ensemble transform kalman filter. part i: Theoretical aspects, *Monthly weather review*, *129*(3), 420–436.

Blei, D. M. (2012), Probabilistic topic models, *Communications of the ACM*, *55*(4), 77–84.

Blei, D. M., and J. D. Lafferty (2006), Dynamic topic models, in *Proceedings of the 23rd International Conference on Machine Learning*, edited by W. Cohen and A. Moore, pp. 113–120, Omni Press, New York, NY, USA.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003), Latent Dirichlet allocation, *Journal of Machine Learning Research*, *3*, 993–1022.

Bolte, J., S. Sabach, and M. Teboulle (2014), Proximal alternating linearized minimization for nonconvex and nonsmooth problems, *Mathematical Programming*, *146*(1-2), 459–494.

Boon-Itt, S., and Y. Skunkan (2020), Public perception of the COVID-19 pandemic on Twitter: sentiment analysis and topic modeling study, *JMIR Public Health and Surveillance*, *6*(4), Article e21,978.

Brown, R. G. (1959), *Statistical forecasting for inventory control*, McGraw/Hill.

Burgers, G., P. Jan van Leeuwen, and G. Evensen (1998), Analysis scheme in the ensemble kalman filter, *Monthly weather review*, *126*(6), 1719–1724.

Campbell, W. F., C. H. Bishop, and D. Hodyss (2010), Vertical covariance localization for satellite radiances in ensemble kalman filters, *Monthly Weather Review*, *138*(1), 282–290.

Carter, K. M., R. Raich, W. G. Finn, and A. O. Hero III (2009), Fine: Fisher information nonparametric embedding, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(11), 2093–2098.

Chandrasekhar, S. (1943), Stochastic problems in physics and astronomy, *Reviews of modern physics*, *15*(1), 1.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (2002), SMOTE: synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, Y., et al. (2019), Identifying solar flare precursors using time series of sdo/hmi images and sharp parameters, *Space Weather*, *17*(10), 1404–1426.

Chu, M. T., and M. M. Lin (2021), Nonlinear power-like and svd-like iterative schemes with applications to entangled bipartite rank-1 approximation, *SIAM Journal on Scientific Computing*, *43*(5), S448–S474, doi:10.1137/20M1336059.

Chuang, J., S. Gupta, C. Manning, and J. Heer (2013), Topic model diagnostics: Assessing domain relevance via topical alignment, in *International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, vol. 28, edited by S. Dasgupta and D. McAllester, pp. 612–620, PMLR, Atlanta, Georgia, USA.

Chuang, J., M. E. Roberts, B. M. Stewart, R. Weiss, D. Tingley, J. Grimmer, and J. Heer (2015), TopicCheck: Interactive alignment for assessing topic model stabilityy, in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 175–184, Association for Computational Linguistics, Denver, Colorado.

Conway, M., and D. O'Connor (2016), Social media, big data, and mental health: current advances and ethical implications, *Current opinion in psychology*, *9*, 77–82.

Costa, J. A., and A. O. Hero (2006), *Determining Intrinsic Dimension and Entropy of High-Dimensional Shape Spaces*, pp. 231–252, Birkhäuser Boston, Boston, MA.

Cressie, N. (2015), *Statistics for spatial data*, John Wiley & Sons.

Cui, W., S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong (2011), TextFlow: Towards better understanding of evolving topics in text, *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2412–2421.

Dai, B., Y. Wang, J. Aston, G. Hua, and D. Wipf (2018), Connections with robust pca and the role of emergent sparsity in variational autoencoder models, *The Journal of Machine Learning Research*, *19*(1), 1573–1614.

Dai, Y.-H., and L.-Z. Liao (2002), R-linear convergence of the Barzilai and Borwein gradient method, *IMA Journal of Numerical Analysis*, *22*(1), 1–10.

Dalal, O., and B. Rajaratnam (2017), Sparse gaussian graphical model estimation via alternating minimization, *Biometrika*, *104*(2), 379–395.

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

De Choudhury, M. (2013), Role of social media in tackling challenges in mental health, in *Proceedings of the 2nd international workshop on Socially-aware multimedia*, pp. 49–52.

De Choudhury, M., E. Kiciman, M. Dredze, G. Coppersmith, and M. Kumar (2016), Discovering shifts to suicidal ideation from mental health content in social media, in *Proceedings of the 2016 CHI conference on human factors in computing systems*, pp. 2098–2110.

Dempsey, W. (2020), The hypothesis of testing: paradoxes arising out of reported coronavirus case-counts, *arXiv preprint arXiv:2005.10425*.

DeRosa, M., et al. (2015), The influence of spatial resolution on nonlinear force-free modeling, *The Astrophysical Journal*, *811*(2), 107.

Diao, Q., J. Jiang, F. Zhu, and E.-P. Lim (2012), Finding bursty topics from microblogs, in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, edited by H. Li, C.-Y. Lin, M. Osborne, G. G. Lee, and J. C. Park, pp. 536–544, Association for Computational Linguistics, Jeju Island, Korea.

Donoho, D. L., and C. Grimes (2003), Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data, *Proceedings of the National Academy of Sciences*, *100*(10), 5591–5596.

Doogan, C., W. Buntine, H. Linger, and S. Brunt (2020), Public perceptions and attitudes toward COVID-19 nonpharmaceutical interventions across six countries: A topic modeling analysis of Twitter data, *Journal of Medical Internet Research*, *22*(9), Article e21,419.

Doshi-Velez, F., and B. Kim (2017), Towards a rigorous science of interpretable machine learning, *arXiv preprint arXiv:1702.08608*.

Doxas, I., S. Dennis, and W. L. Oliver (2010), The dimensionality of discourse, *Proceedings of the National Academy of Sciences*, *107*(11), 4866–4871.

Du, M., N. Liu, and X. Hu (2019), Techniques for interpretable machine learning, *Communications of the ACM*, *63*(1), 68–77.

Ellen, J. (2011), All about microtext - a working definition and a survey of current microtext research within artificial intelligence and natural language processing, in *Proceedings of the 3rd International Conference on Agents and Artificial Intelligence*, edited by J. Filipe and A. Fred, pp. 329–336, SciTePress.

Evensen, G. (1994), Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics, *Journal of Geophysical Research: Oceans*, *99*(C5), 10,143–10,162.

Evensen, G. (2003), The ensemble kalman filter: Theoretical formulation and practical implementation, *Ocean dynamics*, *53*(4), 343–367.

Evensen, G. (2004), Sampling strategies and square root analysis schemes for the enkf, *Ocean dynamics*, *54*(6), 539–560.

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

Fletcher, R. (2005), On the Barzilai-Borwein method, in *Optimization and control with applications*, pp. 235–256, Springer.

Friedman, J., T. Hastie, and R. Tibshirani (2008), Sparse inverse covariance estimation with the graphical lasso, *Biostatistics*, *9*(3), 432–441.

Friedman, J., T. Hastie, and R. Tibshirani (2010), Applications of the lasso and grouped lasso to the estimation of sparse graphical models, *Tech. rep.*, Technical report, Stanford University.

Galvez, R., et al. (2019), A machine-learning data set prepared from the nasa solar dynamics observatory mission, *The Astrophysical Journal Supplement Series*, *242*(1), 7.

Gao, X., and J. Liang (2011), The dynamical neighborhood selection based on the sampling density and manifold curvature for isometric data embedding, *Pattern Recognition Letters*, *32*(2), 202–209.

Gkotsis, G., A. Oellrich, T. Hubbard, R. Dobson, M. Liakata, S. Velupillai, and R. Dutta (2016), The language of mental health problems in social media, in *Proceedings of the third workshop on computational linguistics and clinical psychology*, pp. 63–73.

Godinez, H. C., and J. D. Moulton (2012), An efficient matrix-free algorithm for the ensemble kalman filter, *Computational Geosciences*, *16*(3), 565–575.

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

Gopal, S., and Y. Yang (2014), Von mises-fisher clustering models, in *International Conference on Machine Learning*, pp. 154–162, PMLR.

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

Greenewald, K., and A. O. Hero (2015), Robust kronecker product pca for spatio-temporal covariance estimation, *IEEE Transactions on Signal Processing*, *63*(23), 6368–6378.

Greenewald, K., S. Zhou, and A. Hero (2019), Tensor graphical lasso (teralasso), *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *81*(5), 901–931.

Greenewald, K. H., and A. O. Hero (2014), Kronecker pca based spatio-temporal modeling of video for dismount classification, in *Algorithms for Synthetic Aperture Radar Imagery XXI*, vol. 9093, p. 90930V, International Society for Optics and Photonics.

Greybush, S. J., E. Kalnay, T. Miyoshi, K. Ide, and B. R. Hunt (2011), Balance and ensemble Kalman filter localization techniques, *Monthly Weather Review*, *139*(2), 511–522.

Griffiths, T. L., and M. Steyvers (2002), A probabilistic approach to semantic representation, *Proceedings of the Annual Meeting of the Cognitive Science Society*, *24*.

Griffiths, T. L., and M. Steyvers (2004), Finding scientific topics, *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5228–5235.

Guggenberger, P., F. Kleibergen, and S. Mavroeidis (2022), A test for kronecker product structure covariance matrix, *Journal of Econometrics*.

Guillot, D., B. Rajaratnam, B. T. Rolfs, A. Maleki, and I. Wong (2012), Iterative thresholding algorithm for sparse inverse covariance estimation, *arXiv preprint arXiv:1211.2532*.

Halko, N., P.-G. Martinsson, and J. A. Tropp (2011), Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions, *SIAM review*, *53*(2), 217–288.

Hamill, T. M., J. S. Whitaker, and C. Snyder (2001), Distance-dependent filtering of background error covariance estimates in an ensemble kalman filter, *Monthly Weather Review*, *129*(11), 2776–2790.

Hayden, E. P., R. E. Wiegand, E. T. Meyer, L. O. Bauer, S. J. O'Connor, J. I. Nurnberger Jr, D. B. Chorlian, B. Porjesz, and H. Begleiter (2006), Patterns of regional brain activity in alcohol-dependent subjects, *Alcoholism: Clinical and Experimental Research*, *30*(12), 1986–1991.

Hickmann, K. S., H. C. Godinez, C. J. Henney, and C. N. Arge (2015), Data assimilation in the adapt photospheric flux transport model, *Solar Physics*, *290*(4), 1105–1118.

Hoffman, M. D., D. M. Blei, C. Wang, and J. Paisley (2013), Stochastic variational inference, *The Journal of Machine Learning Research*, *14*(1), 1303–1347.

Holland, P. W., K. B. Laskey, and S. Leinhardt (1983), Stochastic blockmodels: First steps, *Social networks*, *5*(2), 109–137.

Holt, C. C. (2004), Forecasting seasonals and trends by exponentially weighted moving averages, *International Journal of Forecasting*, *20*(1), 5–10.

Hong, L., D. Yin, J. Guo, and B. D. Davison (2011), Tracking trends: incorporating term volume into temporal topic models, in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 484–492.

Hou, E., E. Lawrence, and A. O. Hero (2021), Penalized ensemble kalman filters for high dimensional non-linear systems, *PloS one*, *16*(3), e0248,046.

Houtekamer, P. L., and H. L. Mitchell (2001), A sequential ensemble kalman filter for atmospheric data assimilation, *Monthly Weather Review*, *129*(1), 123–137.

Hsieh, C.-J., M. A. Sustik, I. S. Dhillon, P. K. Ravikumar, and R. Poldrack (2013), Big & quic: Sparse inverse covariance estimation for a million variables, *Advances in neural information processing systems*, *26*.

Hunt, B. R., E. J. Kostelich, and I. Szunyogh (2007), Efficient data assimilation for spatiotemporal chaos: A local ensemble transform kalman filter, *Physica D: Nonlinear Phenomena*, *230*(1-2), 112–126.

Hwang, S. J., S. B. Damelin, and A. O. Hero III (2016), Shortest path through random points, *The Annals of Applied Probability*, *26*(5), 2791–2823.

Hyman, J. M., and B. Nicolaenko (1986), The kuramoto-sivashinsky equation: a bridge between pde's and dynamical systems, *Physica D: Nonlinear Phenomena*, *18*(1-3), 113–126.

Jager, L., and J. A. Wellner (2007), Goodness-of-fit tests via phi-divergences, *The Annals of Statistics*, *35*(5), 2018–2053.

Jang, H., E. Rempel, G. Carenini, and N. Janjua (2020), Exploratory analysis of COVID-19 related tweets in North America to inform public health institutes, in *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, edited by K. Verspoor, K. B. Cohen, M. Conway, B. de Bruijn, M. Dredze, R. Mihalcea, and B. Wallace, Association for Computational Linguistics, Online.

Jiao, Z., H. Sun, X. Wang, W. Manchester, T. Gombosi, A. Hero, and Y. Chen (2020a), Solar flare intensity prediction with machine learning models, *Space Weather*, *18*(7), e2020SW002,440.

Jiao, Z., H. Sun, X. Wang, W. Manchester, T. Gombosi, A. Hero, and Y. Chen (2020b), Solar flare intensity prediction with machine learning models, *Space Weather*, *18*(7), e2020SW002,440.

Kalaitzis, A., J. Lafferty, N. D. Lawrence, and S. Zhou (2013), The bigraphical lasso, in *International Conference on Machine Learning*, pp. 1229–1237, PMLR.

Karimi, H., J. Nutini, and M. Schmidt (2016), Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition, in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 795–811, Springer.

Katzfuss, M., J. R. Stroud, and C. K. Wikle (2016), Understanding the ensemble kalman filter, *The American Statistician*, *70*(4), 350–357.

Ke, Y., S. Minsker, Z. Ren, Q. Sun, and W.-X. Zhou (2019), User-friendly covariance estimation for heavy-tailed distributions, *Statistical Science*, *34*(3), 454–471.

Khare, K., and B. Rajaratnam (2014), Convergence of cyclic coordinatewise l1 minimization, *arXiv preprint arXiv:1404.5100*.

Khare, K., S.-Y. Oh, and B. Rajaratnam (2015), A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees, *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pp. 803–825.

Kim, J., D. Lee, E. Park, et al. (2021), Machine learning for mental health in social media: bibliometric study, *Journal of Medical Internet Research*, *23*(3), e24,870.

Kindermann, R. (1980), Markov random fields and their applications, *American mathematical society*.

Kleibergen, F., and R. Paap (2006), Generalized reduced rank tests using the singular value decomposition, *Journal of econometrics*, *133*(1), 97–126.

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

Koanantakool, P., A. Ali, A. Azad, A. Buluc, D. Morozov, L. Oliker, K. Yelick, and S.-Y. Oh (2018), Communication-avoiding optimization methods for distributed massive-scale sparse inverse covariance estimation, in *International Conference on Artificial Intelligence and Statistics*, pp. 1376–1386, PMLR.

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

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

Krige, D. G. (1951), A statistical approach to some mine valuation and allied problems on the Witwatersrand, Ph.D. thesis, University of the Witwatersrand.

Lauritzen, S. L. (1996), *Graphical models*, vol. 17, Clarendon Press.

Lawson, W. G., and J. A. Hansen (2004), Implications of stochastic and deterministic filters as ensemble-based data assimilation methods in varying regimes of error growth, *Monthly weather review*, *132*(8), 1966–1981.

Leka, K., and G. Barnes (2003), Photospheric magnetic field properties of flaring versus flare-quiet active regions. ii. discriminant analysis, *The Astrophysical Journal*, *595*(2), 1296.

Lemen, J. R., et al. (2011), The atmospheric imaging assembly (aia) on the solar dynamics observatory (sdo), in *The solar dynamics observatory*, pp. 17–40, Springer.

Li, G., and T. K. Pong (2018), Calculus of the exponent of kurdyka–łojasiewicz inequality and its applications to linear convergence of first-order methods, *Foundations of computational mathematics*, *18*(5), 1199–1232.

Li, H., E. Kalnay, and T. Miyoshi (2009), Simultaneous estimation of covariance inflation and observation errors within an ensemble kalman filter, *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, *135*(639), 523–533.

Liese, F. (2012), phi-divergences, sufficiency, Bayes sufficiency, and deficiency, *Kybernetika*, *48*(4), 690–713.

Lindgren, F., H. Rue, and J. Lindström (2011), An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(4), 423–498.

Liu, Q., et al. (2020), Health communication through news media during the early stage of the COVID-19 outbreak in China: Digital topic modeling approach, *Journal of Medical Internet Research*, *22*(4), Article e19,118.

Long, Z., Y. Lu, X. Ma, and B. Dong (2018), Pde-net: Learning pdes from data, in *International Conference on Machine Learning*, pp. 3208–3216, PMLR.

Lourenço, B. F., and A. Takeda (2019), Generalized subdifferentials of spectral functions over euclidean jordan algebras, *arXiv preprint arXiv:1902.05270*.

Lu, B., M. Ott, C. Cardie, and B. K. Tsou (2011), Multi-aspect sentiment analysis with topic models, in *2011 IEEE 11th international conference on data mining workshops*, pp. 81–88, IEEE.

Lyu, X., W. W. Sun, Z. Wang, H. Liu, J. Yang, and G. Cheng (2019), Tensor graphical model: Non-convex optimization and statistical inference, *IEEE transactions on pattern analysis and machine intelligence*, *42*(8), 2024–2037.

Malik, S., A. Smith, T. Hawes, P. Papadatos, J. Li, C. Dunne, and B. Shneiderman (2013), TopicFlow: visualizing topic alignment of Twitter data over time, in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, edited by J. Rokne and C. Faloutsos, pp. 720–726, IEEE.

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

Mcauliffe, J., and D. Blei (2007), Supervised topic models, *Advances in neural information processing systems*, *20*.

McInnes, L., J. Healy, and J. Melville (2018), UMAP: Uniform manifold approximation and projection for dimension reduction, *arXiv preprint arXiv:1802.03426*.

Meinshausen, N., and P. Bühlmann (2006), High-dimensional graphs and variable selection with the lasso, *The annals of statistics*, *34*(3), 1436–1462.

Meng, Y., J. Shen, C. Zhang, and J. Han (2019), Weakly-supervised hierarchical text classification, in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, pp. 6826–6833.

Mimno, D., M. D. Hoffman, and D. M. Blei (2012), Sparse stochastic inference for latent Dirichlet allocation, in *Proceedings of the 29th International Conference on Machine Learning*, edited by J. Langford and J. Pineau, p. 1515–1522, Omni press, Madison, WI, USA.

Minka, T., and J. Lafferty (2002), Expectation-propagation for the generative aspect model, in *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, edited by A. Darwiche and N. Friedman, p. 352–359, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Miyoshi, T. (2011), The gaussian approach to adaptive covariance inflation and its implementation with the local ensemble transform kalman filter, *Monthly Weather Review*, *139*(5), 1519–1535.

Moon, K. R., et al. (2019), Visualizing structure and transitions in high-dimensional biological data, *Nature Biotechnology*, *37*(12), 1482–1492.

Murdoch, W. J., C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu (2019), Definitions, methods, and applications in interpretable machine learning, *Proceedings of the National Academy of Sciences*, *116*(44), 22,071–22,080.

Nerger, L., T. Janjić, J. Schröter, and W. Hiller (2012), A unification of ensemble square root kalman filters, *Monthly Weather Review*, *140*(7), 2335–2345.

Newman, D., P. Smyth, M. Welling, and A. U. Asuncion (2008), Distributed inference for latent Dirichlet allocation, in *Advances in Neural Information Processing Systems*, pp. 1081–1088.

Newman, M. E., and M. Girvan (2004), Finding and evaluating community structure in networks, *Physical review E*, *69*(2), 026,113.

Nguyen, X., M. J. Wainwright, and M. I. Jordan (2009), On surrogate loss functions and f-divergences, *The Annals of Statistics*, *37*(2), 876–904.

Niu, L., X. Liu, and J. Zhao (2020), Robust estimator of the correlation matrix with sparse kronecker structure for a high-dimensional matrix-variate, *Journal of Multivariate Analysis*, *177*, 104,598.

Oh, S., O. Dalal, K. Khare, and B. Rajaratnam (2014), Optimization methods for sparse pseudo-likelihood graphical model selection, *Advances in Neural Information Processing Systems*, *27*, 667–675.

Ott, E., B. R. Hunt, I. Szunyogh, A. V. Zimin, E. J. Kostelich, M. Corazza, E. Kalnay, D. Patil, and J. A. Yorke (2004), A local ensemble kalman filter for atmospheric data assimilation, *Tellus A: Dynamic Meteorology and Oceanography*, *56*(5), 415–428.

Papernot, N., and P. McDaniel (2018), Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning, *arXiv preprint arXiv:1803.04765*.

Parikh, N., and S. Boyd (2014), Proximal algorithms, *Foundations and Trends in optimization*, *1*(3), 127–239.

Park, S., W. Lee, and I.-C. Moon (2015), Supervised dynamic topic models for associative topic extraction with a numerical time series, in *Proceedings Of the 2015 Workshop On Topic Models: Post-Processing And Applications*, pp. 49–54.

Park, S., K. Shedden, and S. Zhou (2017), Non-separable covariance models for spatio-temporal data, with applications to neural encoding analysis, *arXiv preprint arXiv:1705.05265*.

Peng, J., P. Wang, N. Zhou, and J. Zhu (2009), Partial correlation estimation by joint sparse regression models, *Journal of the American Statistical Association*, *104*(486), 735–746.

Petterson, J., W. Buntine, S. Narayanamurthy, T. Caetano, and A. Smola (2010), Word features for latent dirichlet allocation, *Advances in Neural Information Processing Systems*, *23*.

Qiang, J., Z. Qian, Y. Li, Y. Yuan, and X. Wu (2020), Short text topic modeling techniques, applications, and performance: a survey, *IEEE Transactions on Knowledge and Data Engineering*.

Qiao, X., S. Guo, and G. M. James (2019), Functional graphical models, *Journal of the American Statistical Association*, *114*(525), 211–222.

Qiu, M., F. Zhu, and J. Jiang (2013), It is not just what we say, but how we say them: LDA-based behavior-topic model, in *Proceedings of the 2013 SIAM International Conference on Data Mining*, edited by J. Ghosh, Z. Obradovic, J. Dy, Z.-H. Zhou, C. Kamath, and S. Parthasarathy, pp. 794–802, SIAM.

Ramage, D., D. Hall, R. Nallapati, and C. D. Manning (2009), Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora, in *Proceedings of the 2009 conference on empirical methods in natural language processing*, pp. 248–256.

Rao, C. R. (1995), *The use of Hellinger distance in graphical displays of contingency table data*, pp. 143–161, VSP, Utrecht.

Raydan, M. (1993), On the Barzilai and Borwein choice of steplength for the gradient method, *IMA Journal of Numerical Analysis*, *13*(3), 321–326.

Raydan, M. (1997), The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem, *SIAM Journal on Optimization*, *7*(1), 26–33.

Reichstein, M., G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, et al. (2019), Deep learning and process understanding for data-driven earth system science, *Nature*, *566*(7743), 195–204.

Roberts, B. (2006), Slow magnetohydrodynamic waves in the solar atmosphere, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *364*(1839), 447–460.

Rohe, K., S. Chatterjee, B. Yu, et al. (2011), Spectral clustering and the high-dimensional stochastic blockmodel, *The Annals of Statistics*, *39*(4), 1878–1915.

Rozansky, M. (2020), In a holiday season unlike any other, avoid unfounded claims about suicide, `https://www.annenbergpublicpolicycenter.org/in-a-holiday-season-unlike-any-other-avoid-unfounded-claims-about-suicide/`, accessed: 2022-06-24.

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

Rudin, C. (2019), Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nature Machine Intelligence*, *1*(5), 206–215.

Rudin, C., C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong (2022), Interpretable machine learning: Fundamental principles and 10 grand challenges, *Statistics Surveys*, *16*, 1–85.

Rue, H., and L. Held (2005), *Gaussian Markov random fields: theory and applications*, Chapman and Hall/CRC.

Samko, O., A. D. Marshall, and P. L. Rosin (2006), Selection of the optimal parameter value for the isomap algorithm, *Pattern Recognition Letters*, *27*(9), 968–979.

Santomauro, D. F., et al. (2021), Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the covid-19 pandemic, *The Lancet*, *398*(10312), 1700–1712.

Sareen, J., B. J. Cox, T. O. Afifi, R. de Graaf, G. J. Asmundson, M. Ten Have, and M. B. Stein (2005), Anxiety disorders and risk for suicidal ideation and suicide attempts: a population-based longitudinal study of adults, *Archives of general psychiatry*, *62*(11), 1249–1257.

Schrijver, C., et al. (2008), Nonlinear force-free field modeling of a solar active region around the time of a major flare and coronal mass ejection, *The Astrophysical Journal*, *675*(2), 1637.

Servidea, J. D., and X.-L. Meng (2006), Statistical physics and statistical computing: A critical link, in *Frontiers In Statistics*, edited by J. Fan and H. L. Koul, pp. 327–344, World Scientific.

Sha, H., M. A. Hasan, G. Mohler, and P. J. Brantingham (2020), Dynamic topic modeling of the covid-19 twitter narrative among us governors and cabinet executives, *arXiv preprint arXiv:2004.11692*.

Sharma, A., M. Choudhury, T. Althoff, and A. Sharma (2020a), Engagement patterns of peer-to-peer interactions on mental health platforms, in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, pp. 614–625.

Sharma, A., A. S. Miner, D. C. Atkins, and T. Althoff (2020b), A computational approach to understanding empathy expressed in text-based mental health support, *arXiv preprint arXiv:2009.08441*.

Sharma, A., I. W. Lin, A. S. Miner, D. C. Atkins, and T. Althoff (2021), Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach, in *Proceedings of the Web Conference 2021*, pp. 194–205.

Shefi, R., and M. Teboulle (2016), On the rate of convergence of the proximal alternating linearized minimization algorithm for convex problems, *EURO Journal on Computational Optimization*, *4*(1), 27–46.

Shemyakin, A. (2014), Hellinger distance and non-informative priors, *Bayesian Analysis*, *9*(4), 923–938.

Song, X., C.-Y. Lin, B. L. Tseng, and M.-T. Sun (2005), Modeling and predicting personal information dissemination behavior, in *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, edited by R. Grossman, R. Bayardo, and K. Bennett, pp. 479–488, Association for Computing Machinery.

Srivastava, A., and E. P. Klassen (2016), *Functional and shape data analysis*, vol. 1, Springer.

Srivastava, A., and C. Sutton (2017), Autoencoding variational inference for topic models, *arXiv preprint arXiv:1703.01488*.

Stocker, T. (2011), *Introduction to climate modelling*, Springer Science & Business Media.

Stokes, D. C., A. Andy, S. C. Guntuku, L. H. Ungar, and R. M. Merchant (2020), Public priorities and concerns regarding COVID-19 in an online discussion forum: Longitudinal topic modeling, *Journal of General Internal Medicine*, *35*(7), 2244–2247.

Su, Y., A. Venkat, Y. Yadav, L. B. Puglisi, and S. J. Fodeh (2021), Twitter-based analysis reveals differential covid-19 concerns across areas with socioeconomic disparities, *Computers in Biology and Medicine*, *132*, Article 104,336.

Sun, H., W. Manchester, Z. Jiao, X. Wang, and Y. Chen (2019), Interpreting lstm prediction on solar flare eruption with time-series clustering, *arXiv preprint arXiv:1912.12360*.

Sun, W. W., Z. Wang, X. Lyu, H. Liu, and G. Cheng (2016), *Tlasso: Non-Convex Optimization and Statistical Inference for Sparse Tensor Graphical Models*, r package version 1.0.1.

Sun, Z., M. Bobra, X. Wang, Y. Wang, H. Sun, T. Gombosi, Y. Chen, and A. Hero (2021), Predicting solar flares using cnn and lstm on two solar cycles of active region data, *Earth and Space Science Open Archive*, p. 32, doi:10.1002/essoar.10508256.1.

Taddy, M. (2012), On estimation and selection for topic models, in *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, *Proceedings of Machine Learning Research*, vol. 22, edited by N. D. Lawrence and M. Girolami, pp. 1184–1193, PMLR, La Palma, Canary Islands.

Takeishi, N., Y. Kawahara, and T. Yairi (2017), Learning koopman invariant subspaces for dynamic mode decomposition, *arXiv preprint arXiv:1710.04340*.

Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei (2006), Hierarchical Dirichlet processes, *Journal of the American Statistical Association*, *101*(476), 1566–1581.

Teh, Y. W., K. Kurihara, and M. Welling (2008), Collapsed variational inference for HDP, in *Advances in Neural Information Processing Systems*, vol. 20, edited by J. Platt, D. Koller, Y. Singer, and S. Roweis, pp. 1481–1488, Curran Associates, Inc.

Tenenbaum, J. B., V. De Silva, and J. C. Langford (2000), A global geometric framework for nonlinear dimensionality reduction, *Science*, *290*(5500), 2319–2323.

Thomas, J. W. (2013), *Numerical partial differential equations: finite difference methods*, vol. 22, Springer Science & Business Media.

Tibshirani, R. (1996), Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Tippett, M. K., J. L. Anderson, C. H. Bishop, T. M. Hamill, and J. S. Whitaker (2003), Ensemble square root filters, *Monthly weather review*, *131*(7), 1485–1490.

Tödter, J., and B. Ahrens (2015), A second-order exact ensemble square root filter for nonlinear data assimilation, *Monthly Weather Review*, *143*(4), 1347–1367.

Tsiligkaridis, T., and A. O. Hero (2013), Covariance estimation in high dimensions via kronecker product expansions, *IEEE Transactions on Signal Processing*, *61*(21), 5347–5360.

Tsiligkaridis, T., A. O. Hero, and S. Zhou (2013), On convergence of kronecker graphical lasso algorithms, *IEEE transactions on signal processing*, *61*(7), 1743–1755.

van der Maaten, L., and G. Hinton (2008), Visualizing data using t-SNE, *Journal of Machine Learning Research*, *9*, 2579–2605.

van Driel-Gesztelyi, L., and L. M. Green (2015), Evolution of active regions, *Living Reviews in Solar Physics*, *12*(1), 1–98.

Van Loan, C. F., and N. Pitsianis (1993), Approximation with kronecker products, in *Linear algebra for large scale and real-time applications*, pp. 293–314, Springer.

Varin, C., N. Reid, and D. Firth (2011), An overview of composite likelihood methods, *Statistica Sinica*, pp. 5–42.

Vlachas, P. R., W. Byeon, Z. Y. Wan, T. P. Sapsis, and P. Koumoutsakos (2018), Data-driven forecasting of high-dimensional chaotic systems with long short-term memory networks, *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *474*(2213), 20170,844.

von Neumann, J. (2013), *Mathematische grundlagen der quantenmechanik*, vol. 38, Springer-Verlag.

Wang, C., D. Blei, and D. Heckerman (2008), Continuous time dynamic topic models, in *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, edited by D. McAllester and P. Myllymaki, pp. 579–586, AUAI Press, Arlington, Virginia, USA.

Wang, H., Y. Lu, and C. Zhai (2010), Latent aspect rating analysis on review text data: a rating regression approach, in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 783–792.

Wang, R., K. Kashinath, M. Mustafa, A. Albert, and R. Yu (2020a), Towards physics-informed deep learning for turbulent flow prediction, in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1457–1466.

Wang, X., and A. McCallum (2006), Topics over time: a non-Markov continuous-time model of topical trends, in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, edited by T. Eliassi-Rad, L. Ungar, M. Craven, and D. Gunopulos, pp. 424–433, Association for Computing Machinery.

Wang, X., N. Mohanty, and A. McCallum (2005), Group and topic discovery from relations and text, in *Proceedings of the 3rd International Workshop on Link Discovery*, edited by J. Adibi, M. Grobelnik, D. Mladenic, and P. Pantel, pp. 28–35, Association for Computing Machinery.

Wang, X., T. M. Hamill, J. S. Whitaker, and C. H. Bishop (2007), A comparison of hybrid ensemble transform kalman filter–optimum interpolation and ensemble square root filter analysis schemes, *Monthly weather review*, *135*(3), 1055–1076.

Wang, X., et al. (2020b), Predicting solar flares with machine learning: investigating solar cycle dependence, *The Astrophysical Journal*, *895*(1), 3.

Wang, Y., and A. Hero (2021a), Multiway ensemble kalman filter, *arXiv preprint arXiv:2112.04322*.

Wang, Y., and A. Hero (2021b), Sg-palm: a fast physically interpretable tensor graphical model, *International Conference on Machine Learning (ICML), arXiv preprint arXiv:2105.12271*.

Wang, Y., and S. Ma (2007), Projected Barzilai-Borwein method for large-scale nonnegative image restoration, *Inverse Problems in Science and Engineering*, *15*(6), 559–583.

Wang, Y., B. Jang, and A. Hero (2020c), The sylvester graphical lasso (syglasso), in *International Conference on Artificial Intelligence and Statistics*, pp. 1943–1953, PMLR.

Wang, Y., C. Hougen, B. Oselio, W. Dempsey, and A. Hero (2021), A Geometry-Driven Longitudinal Topic Model, *Harvard Data Science Review*, *3*(2), https://hdsr.mitpress.mit.edu/pub/0v7qw6jf.

Wang, Y., Z. Sun, and A. Hero (2022), Tensorgraphicalmodels: A julia toolbox for multiway covariance models and ensemble kalman filter, *Software Impacts*, *13*, 100,308, doi:https://doi.org/10.1016/j.simpa.2022.100308.

Wei, X., and S. Minsker (2017), Estimation of the covariance structure of heavy-tailed distributions, in *Advances in Neural Information Processing Systems*, vol. 30, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Curran Associates, Inc.

Weinan, E., J. Han, and L. Zhang (2020), Integrating machine learning with physics-based modeling, *arXiv*.

Wen, Z., W. Yin, D. Goldfarb, and Y. Zhang (2010), A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization, and continuation, *SIAM Journal on Scientific Computing*, *32*(4), 1832–1857.

Wheatland, M., and S. Gilchrist (2013), The state of nonlinear force-free magnetic field extrapolation, in *Journal of Physics: Conference Series*, vol. 440, p. 012037, IOP Publishing.

Whitaker, J. S., and T. M. Hamill (2002), Ensemble data assimilation without perturbed observations, *Monthly weather review*, *130*(7), 1913–1924.

Winters, P. R. (1960), Forecasting sales by exponentially weighted moving averages, *Management science*, *6*(3), 324–342.

Witten, D. M., and R. Tibshirani (2010), A framework for feature selection in clustering, *Journal of the American Statistical Association*, *105*(490), 713–726.

Wood, S. N. (2017), *Generalized additive models: an introduction with R*, CRC Press.

Wright, S. J., R. D. Nowak, and M. A. Figueiredo (2009), Sparse reconstruction by separable approximation, *IEEE Transactions on signal processing*, *57*(7), 2479–2493.

Xu, Y., and W. Yin (2013), A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion, *SIAM Journal on imaging sciences*, *6*(3), 1758–1789.

Xue, J., J. Chen, C. Chen, C. Zheng, S. Li, and T. Zhu (2020), Public discourse and sentiment during the COVID-19 pandemic: Using latent Dirichlet allocation for topic modeling on Twitter, *PloS One*, *15*(9), Article e0239,441.

Yang, Y., C. Chen, and F. S. Bao (2016), Aspect-based helpfulness prediction for online product reviews, in *2016 IEEE 28th International Conference on Tools with Artificial Intelligence*, edited by N. Bourbakis, A. Esposito, A. Mali, and M. Alamaniotis, pp. 836–843, IEEE.

Yin, J., and H. Li (2012), Model selection and estimation in the matrix normal graphical model, *Journal of multivariate analysis*, *107*, 119–140.

Yuan, M., B. Van Durme, and J. L. Ying (2018), Anchoring: Interactive topic modeling and alignment across languages, in *Advances in Neural Information Processing Systems*, vol. 31, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, pp. 8653–8663, Curran Associates, Inc.

Zhang, C.-H., et al. (2010), Nearly unbiased variable selection under minimax concave penalty, *The Annals of statistics*, *38*(2), 894–942.

Zhang, H. (2020), New analysis of linear convergence of gradient-type methods via unifying error bound conditions, *Mathematical Programming*, *180*(1), 371–416.

Zhang, L., J. Han, H. Wang, W. Saidi, R. Car, and W. E (2018), End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems, in *Advances in Neural Information Processing Systems*, vol. 31, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Curran Associates, Inc.

Zhang, X. L., H. Begleiter, B. Porjesz, W. Wang, and A. Litke (1995), Event related potentials during object recognition tasks, *Brain Research Bulletin*, *38*(6), 531–538.

Zhao, P., and B. Yu (2006), On model selection consistency of lasso, *Journal of Machine learning research*, *7*(Nov), 2541–2563.

Zhao, W. X., J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li (2011), Comparing Twitter and traditional media using topic models, in *Advances in Information Retrieval*, edited by P. Clough, C. Foley, C. Gurrin, G. J. F. Jones, W. Kraaij, H. Lee, and V. Mudoch, pp. 338–349, Springer Berlin Heidelberg, Berlin, Heidelberg.

Zhou, S. (2014), Gemini: Graph estimation with matrix variate normal instances, *The Annals of Statistics*, *42*(2), 532–562.

Zhou, S., P. Rütimann, M. Xu, and P. Bühlmann (2011), High-dimensional covariance estimation based on gaussian graphical models, *Journal of Machine Learning Research*, *12*(Oct), 2975–3026.

Zhu, H., N. Strawn, and D. B. Dunson (2016), Bayesian graphical models for multivariate functional data, *The Journal of Machine Learning Research*, *17*(1), 7157–7183.

Zhu, J., H. Wang, B. K. Tsou, and M. Zhu (2009), Multi-aspect opinion polling from textual reviews, in *Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 1799–1802.

Zhu, J., A. Ahmed, and E. P. Xing (2012), Medlda: maximum margin supervised topic models, *the Journal of machine Learning research*, *13*(1), 2237–2278.

Zou, H. (2006), The adaptive lasso and its oracle properties, *Journal of the American statistical association*, *101*(476), 1418–1429.