Distributed learning, prediction and detection in probabilistic graphs

by

Zhaoshi Meng

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering: Systems)
in the University of Michigan
2014

Doctoral Committee:

Professor Alfred O. Hero III, Co-Chair
Assistant Professor Long Nguyen, Co-Chair
Assistant Professor Laura K. Balzano
Associate Professor Qiaozhu Mei

2014

To my parents and Kaiping

Acknowledgments

The completion of this dissertation marks the ending of a journey. There have been so many people who made this journey enjoyable and memorable, to whom I am giving my sincere acknowledgements.

My first and foremost acknowledgement goes to my advisor, Professor Alfred Hero. This dissertation would not be possible without his invaluable advice, guidance and vision. I am grateful for the freedom he has given to me to explore my research interests, to collaborate with other researchers, and to work off campus from time to time. Al's encouraging and supportive comments have given me enormous confidence and courage to challenge the unknown. His persistent enthusiasm in research will continue to motivate me in my own career ahead.

I am grateful to have Professor Long Nguyen as my co-advisor. Taking his STATS 601 has been one of the most inspiring and rewarding learning experiences. Those long-hour meetings with him have always been so intellectually enjoyable, thanks to his efforts in explaining insights in a transparent manner and his patience in guiding me through the research challenges. I have also benefited greatly from his style of formulating elegant solutions for practical problems. Long has also been a supportive, encouraging, and considerate mentor, for which I am truly thankful.

I am fortunate to have Professor Laura Balzano and Professor Qiaozhu Mei on my dissertation committee. Laura helped improve the dissertation through asking insightful questions and giving thoughtful suggestions on my writing. Qiaozhu has been extremely kind and encouraging, and I've learned lots of practical considerations in data mining through interactions with him. Several works in this dissertation are results of collaborations with some world-class researchers – Ami Wiesel, Dennis Wei and Brian Eriksson. I always benefited from Ami's to-the-point and insightful comments. I thank him for introducing distributed learning in graphical models to me and guiding me onto the track of doing research. Dennis has been the go-to person whenever I got stuck or having trouble clarifying my thoughts, and he always helped me greatly. Brian has been a phenomenal mentor during my internship at Technicolor Research Lab. His passion, humor, and devotion to work have all influenced me.

My years in Michigan as a graduate student have been very pleasant due to many people. I would like to thank Hamed Firouzi for his friendship, the time we spent playing tennis, preparing for conferences, and touring the Grand Canyon. I would also like to thank Mark Hsiao for always being energetic and inspiring, and for sharing so many good stories with me in Mandarin. I thank Kevin Xu for being so kind, knowledgable and helpful. I am fortunate to work alongside some brilliant researchers in the Hero Group – Tianpei Xie, Yu-Hui Chen, Kristjan Greenewald, Kevin Moon, Greg Newstadt, Tzu-Yu Liu, Sung Jin Hwang, Kumar Sricharan, Arnau Tibau Puig, Yilun Chen, Goran Marjanovic, Alex Kulesza, and many others. I am grateful to Becky Turanski, Michele Feldkamp, Elizabeth Zaenger (and other DCO staffs) for taking care of things and making my life much easier. I also thank Austin Lee, Hao Sun, Qingsi Wang, Hong Li, Yingze Bao, Jiangfeng Wu, Shengtao Wang, Yuanhao Zhai, and many other friends in Ann Arbor who have enriched my life. I would also like to acknowledge many friends outside Michigan. My friends in Stanford have made my occasional visits full of relaxing chats and laughters – Jack Xiaojiang Guo has been such a good friend in so many ways; Anita Varma has always been there when we needed help and providing considerate and meaningful advice; René Kizilcec is cheerfully smart; Jakki Bailey and Octavia Zahrt are caring and kind; Jason Lee, Yuekai Sun, and Wenzhao Lian have made my internship much more fun. I also thank Quan Geng, Yinfeng Qin, Zhaobo Lang, Yan Ji, Yang Gao, Jian Liu, Shi Qiu, Xiaoyou Wang for keeping in touch and sharing their life experiences. I apologize in advance to anyone whom I have neglected to mention.

Finally, this dissertation is dedicated to my family. To my parents, Mr. Guangwei Meng and Ms. Fei Gao, and my mother-in-law, Ms. Yuhua Wang, for their unconditional love, support, and belief in me. To my wife, Kaiping Zhang, for accompanying me through all the ups and downs in this journey, for being brave in difficult times, for understanding and making the sacrifice, and for always reminding me the good things in life with her smile.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

ABSTRACT

Distributed learning, prediction and detection in probabilistic graphs

by

Zhaoshi Meng

Co-Chairs: Alfred O. Hero III and XuanLong Nguyen

Critical to high-dimensional statistical estimation is to exploit the structure in the data distribution. *Probabilistic graphical models* provide an efficient framework for representing complex joint distributions of random variables through their conditional dependency graph, and can be adapted to many high-dimensional machine learning applications.

This dissertation develops the probabilistic graphical modeling technique for three statistical estimation problems arising in real-world applications: distributed and parallel learning in networks, missing-value prediction in recommender systems, and emerging topic detection in text corpora. The common theme behind all proposed methods is a combination of parsimonious representation of uncertainties in the data, optimization surrogate that leads to computationally efficient algorithms, and fundamental limits of estimation performance in high dimension.

More specifically, the dissertation makes the following theoretical contributions:

(1) We propose a distributed and parallel framework for learning the parameters in Gaussian graphical models that is free of iterative global message passing. The proposed distributed estimator is shown to be asymptotically consistent, improve with increasing local neighborhood sizes, and have a high-dimensional error rate comparable to that of the centralized maximum likelihood estimator.

(2) We present a family of latent variable Gaussian graphical models whose marginal precision matrix has a "low-rank plus sparse" structure. Under mild conditions, we analyze the high-dimensional parameter error bounds for learning this family of models using regularized maximum likelihood estimation.

(3) We consider a hypothesis testing framework for detecting emerging topics in topic models, and propose a novel surrogate test statistic for the standard likelihood ratio. By leveraging the theory of empirical processes, we prove asymptotic consistency for the proposed test and provide guarantees of the detection performance.

CHAPTER 1

Introduction

Recent advances in information technology have transformed the data generation and acquisition landscape, resulting in diverse and high volumes of data. Examples include measurements of various sensing devices (such as ubiquitous sensors, smartphones, and agents in a social network), behavioral data from users of large-scale Internet services such as recommendation systems, and digital publication media.

Analyzing and deriving insights from such data sets has emerged as a topic of great interest to researchers in both academia and industry. While great effort has been devoted to improving the data processing capability, the statistical aspects of the problems are often less studied in practice. One of the fundamental problems is that of statistical estimation, *i.e.,* to recover or extract distributional characteristics from observations, which is often the first step towards statistical inference tasks such as prediction and detection. Unfortunately, the increasing data dimensions pose tremendous challenges for statistical estimation. For instance, modern high-resolution imaging systems typically generate samples from millions of highly correlated pixel variables. Similarly, an online recommender system (such as the movie streaming platform *Netflix*, or the E-commerce website *Amazon.com*) often has hundreds of millions of users and even more products for recommendation. The estimation problem at such scale requires a judicious choice of model, scalable computational techniques, and deep understanding of its statistical efficiency with respect to the number of

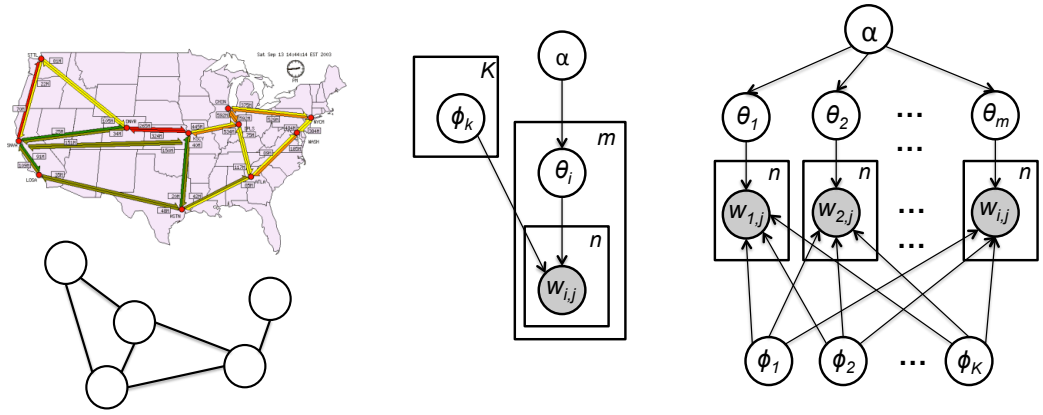
observations.

This dissertation addresses these three aspects of modeling, computation, and performance in the context of statistical estimation under the framework of probabilistic graphical modeling. Three broad types of applications are discussed, namely distributed and parallel estimation in networks, missing-value prediction in recommender systems, and emerging topic detection in text corpora. The theory and algorithms developed in this dissertation also apply to other related problems. For each of the three applications, we propose probabilistic graphical modeling that is adapted to the structure of the problem, and provide both algorithmic, experimental and theoretical analysis. The common theme behind all proposed methods is a combination of parsimonious representation of uncertainties in the data, optimization surrogate that leads to computationally efficient algorithms, and fundamental limits of estimation performance in high dimension.

1.1 Overview of probabilistic graphical models

One of the most recent and fruitful directions in high-dimensional statistical estimation is the use of regularization that promotes simple and low-dimensional structures. Examples abound in machine learning, statistics, and signal processing, including sparse regression models, low-rank matrix factorization models, and their variants and combinations. These methods essentially impose sparsity (*i.e.,* a small number of non-zero parameters) in a certain transformed domain.

Probabilistic graphical modeling (PGM) is a principled framework for compactly modeling high-dimensional distributions through the *conditional dependency* graph among the variables (Lauritzen, 1996; Wainwright & Jordan, 2008; Koller & Friedman, 2009). In PGM, each node often represents a random variable (could be either observed or latent), and an edge denotes the corresponding conditional dependence,



(a) The Abilene network and an undirected graphical model

(b) The LDA model (the "plate" and a partially unrolled representation)

Figure 1.1: Examples of two probabilistic graphical models: (a) a Markov random field for the traffic flows in the computer network *Abilene*, and (b) a hierarchical Bayesian network, the latent Dirichlet allocation model, for text documents. In (a), each node of the graphical model represents the traffic flows in a loop of the physical connectivity graph of Abilene. In (b), the "plate" (left) and a partially unrolled (right) representation are shown, where the shaded nodes denote the observed words in the documents and the unshaded nodes denote latent and shared variables and parameters (see text for details).

i.e., the statistical dependence between two random variables conditioning on all remaining variables. This conditional dependency graph is often assumed to be sparse, which leads to a low-dimensional yet still richly structured representation. Note that two variables could be conditionally independent but *marginally* dependent, due to their mutual dependence with other variables. This phenomenon implies that the set of conditional dependencies is a more parsimonious representation of the distribution than the set of (marginal) dependencies. Such modeling efficiency and flexibility makes PGM especially suitable to high-dimensional distributions.

Two most commonly studied PGM methods are *Markov random fields* (MRF) and *Bayesian networks* (see Figure. 1.1 for illustrations). Markov random fields are often used for modeling random variables that are spatially correlated, for example, pixel values in an images or a video, or sensory signals in a sensor network. The common fundamental assumption is that the statistical dependence between two spa-

tially separated random variables is likely to be weaker than that of two spatially connected variables. In practice this assumption is often found to be a reasonable approximation, therefore MRF has been successfully applied to many applications. In Figure 1.1(a) we show an example for using MRF to model the traffic flows in a computer network, in which the graphical model is constructed from the physical connectivity graph of the computer network. We refer the readers to Wainwright & Jordan (2008); Koller & Friedman (2009) for more examples.

The graph associated with a MRF is often an undirected graph, representing a certain symmetric relation (such as conditional dependence) between the nodes. Bayesian networks, on the other hand, use directed and often "hierarchical" graphs to represent the asymmetric relations between nodes, which could represent either observed or latent variables. The introduction of latent variables is a crucial part of creating a sparse representation, since the observed variables are otherwise densely connected. Another advantage of introducing latent variables that are connected to many observed variables is to allow for the sharing of statistical strength in the inference. In Figure 1.1(b), a hierarchical Bayesian network, called *latent Dirichlet allocation* (LDA) model is shown. The LDA model is widely used for modeling collection of text documents, where each leaf node represents a word, and all words in the same document share a parent node which represent certain (latent) document-specific characteristics. All these document-level nodes, again, share a parent node which captures certain corpus-level features. The "plate" notation is an efficient representation of this sharing of variables or parameters, as shown in Figure 1.1(b), where each box denotes multiple repetitions of the same structure in the "unrolled" network.

Given a graphical model with known structure and parameters, the classical task is to infer for the marginal distribution of a subset of random variables (or their conditional distribution conditioning on some evidence). For small or loop-free graphs

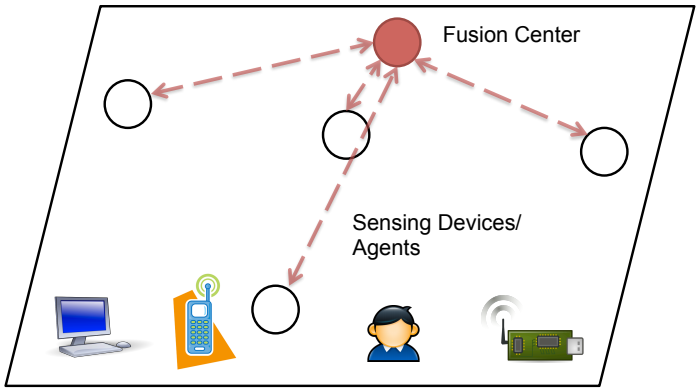
(such as tree-structured MRF or simple Bayesian networks), this inference can be done through sequential marginalization. However, for loopy graphs, an exact marginal inference is in general intractable even for graphs with moderate size, due to the exponentially increasing computational complexity. A distributed message passing algorithm (also known as belief propagation) has been developed for approximate inference in loopy graphs, which makes PGM especially appealing to large and decentralized systems. For Bayesian networks, when exact inference is intractable, methods such as the variational inference and Markov Chain Monte Carlo sampling techniques are often used for approximate inference. For more details on the inference of PGM, we refer the readers again to Wainwright & Jordan (2008); Koller & Friedman (2009).

1.2 Dissertation outline and contributions

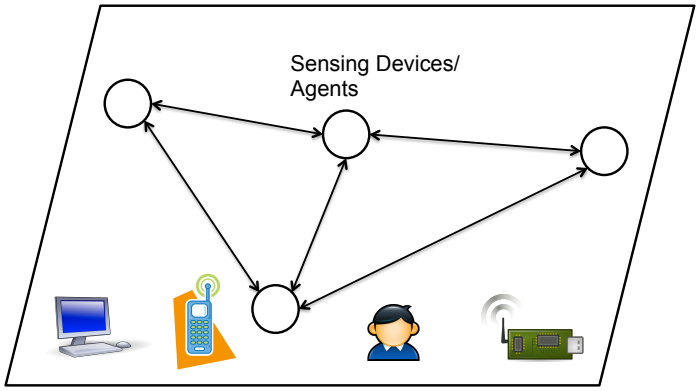
This dissertation focuses on three problems related to PGM.

1.2.1 Distributed estimation in Gaussian graphical models

In the first part of the dissertation, we consider distributed estimation in PGM, focusing on the Gaussian MRF model, also known as the Gaussian graphical model. This general problem is motivated by real-world applications, where the data is collected from networks of decentralized sensing devices and agents (see Figure 1.2 for an illustration). Due to the spatial correlations among the sensing devices, the joint distribution of the collected data often contains certain structure that can be naturally captured by MRF. Statistical estimation is often performed in a *centralized* framework (Figure 1.2(a)), in which all the data samples are collected at a fusion center for the estimation. Unfortunately, for large network applications, this approach is often constrained by limited computational capability of local devices and limited communication bandwidth. The huge volume of data generated by large networks



(a) The centralized estimation framework



(b) The distributed estimation framework

Figure 1.2: Illustration of the centralized and distributed estimation frameworks for data collected from networks.

creates computational challenges even for the fusion center. A distributed estimation framework (Figure 1.2(b)), on the other hand, utilizes in-network and local computation, and hence is advantageous in many aspects. It often requires negligible or no communication with the fusion center, which significantly reduces the communication cost by lowering bandwidth requirements. The computational cost for localized estimation may also be significantly reduced. The key to designing statistical estimation algorithms in a distributed framework is to find a good tradeoff between the computational complexity and the statistical accuracy achieved by the system.

In Chapter 2, we consider the problem of estimating parameters in a Gaussian MRF from data samples, which is a prerequisite for performing statistical inference

using the MRF. Since the classical global maximum likelihood estimation is centralized, we propose a general distributed framework for parameter estimation, in which local parameter estimates are obtained by maximizing the *marginal* likelihoods of local neighborhoods. Due to the non-convexity of this problem, we propose to solve a convex surrogate. The local estimates are then combined into a global estimate without any iterative message-passing between neighborhoods. For this proposed estimator, we derive asymptotic properties such as consistency and monotonicity of the variance (in terms of local neighborhood sizes) under Gaussian distribution assumption. Further high-dimensional analysis also provides a convergence rate for its estimation error, which is comparable to the global maximum likelihood estimator that requires centralized computation. This asymptotic analysis establishes the statistical efficiency of the proposed estimator. Since there is no need for message passing, the proposed estimation framework is naturally distributed and parallelizable, making it well adapted to decentralized and large-scale problems. We illustrate the performance of our proposed estimator in both synthetic and real-world data sets.

In Chapter 3, we focus on directed Gaussian graphical models, and introduce a distributed algorithm for performing principal component analysis (PCA) to estimate a low-dimensional subspace. Note that the graphical model structure characterizes the conditional dependencies arise in the data distribution, the low-dimensional subspace we are interested in often captures other application-specific aspects of the data, such as partially hidden latent effects. By leveraging structured sparsity in the Cholesky factor of the inverse covariance matrix, our proposed algorithm estimates the global principal subspace through local computation and message passing. At the core of this algorithm is a distributed matrix-vector multiplier, which is shown to coincide with a Cholesky-parameterized sum-product algorithm specialized to exploit the structure of the problem. We illustrate the estimation efficiency and the reduced computation and communication complexities of the proposed algorithm on both synthetic and

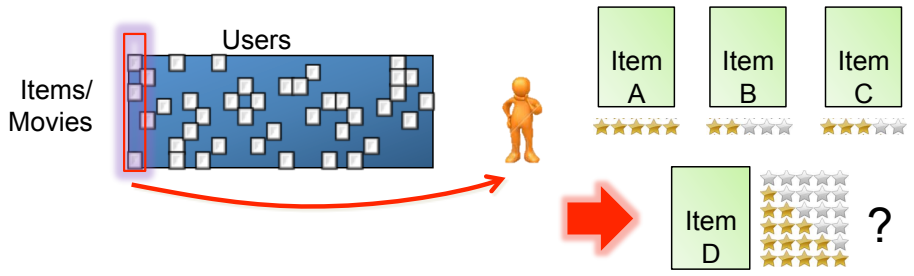


Figure 1.3: Missing-value prediction in recommender systems.

real-world network data sets. In particular, in the real-world problem, we propose two statistical directed graphical models based on the flow routing pattern of a computer network, and demonstrate the computational advantages using the proposed algorithm.

1.2.2 Learning latent variable Gaussian graphical models with application to recommender systems

In Chapter 4, we consider learning Gaussian graphical models with latent variables. This problem is also motivated by real-world applications. In particular, we focus on the missing-value prediction problem, also known as collaborative filtering, in recommender systems (Figure 1.3).

In recommender systems, available is an incomplete data sample matrix where each element denotes the rating score that a user gives to an item. For example, we consider rating movies in a database. The goal is to predict the unobserved or missing ratings based on the observed ones. Fundamental to missing-value prediction is to choose and learn a model that governs the joint distribution of the rating variables. In this chapter, we propose to use *latent variable Gaussian graphical models* (LVGGM). See Figure 1.4 for an illustration. Specifically, we introduce latent variables to model certain *global* factors that affect a large portion of the movie ratings (such as the movie genres). Conditioning on these latent factors, the conditional distribution of

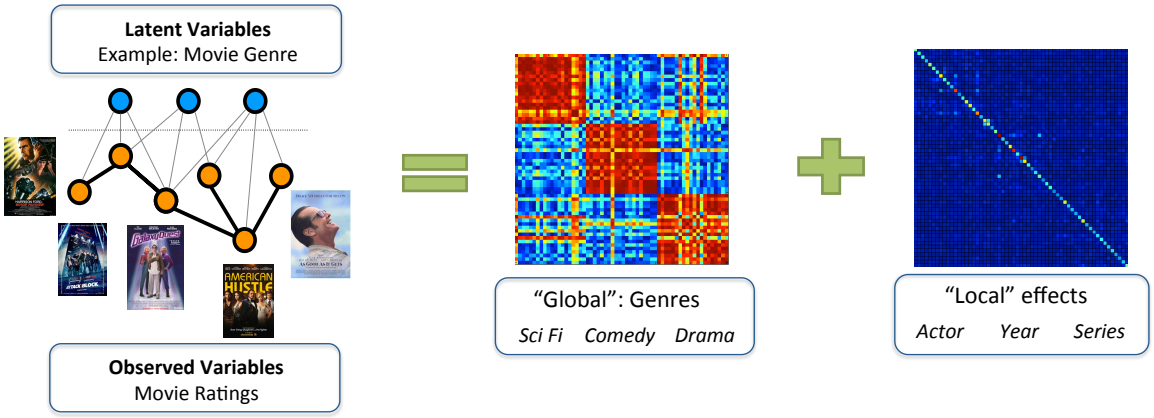


Figure 1.4: Illustration of the latent variable Gaussian graphical model for movie ratings. The marginal inverse covariance matrix of the rating variables has a "low-rank plus sparse" decomposition, corresponding to certain "global" and "local" effects, respectively.

the (observed) movie rating variables is captured by a sparse graphical model. Unlike standard sparse Gaussian graphical models that do not incorporate latent variables, our proposed model does not have a sparse marginal inverse covariance matrix for the observed variables. Instead, due to the property of Gaussian distribution, it can be shown to have a "low-rank plus sparse" decomposition, which allows for efficient learning as described below.

In practice, the number of items in a recommender system could be large. Therefore, following Chandrasekaran et al. (2012), we consider regularized maximum likelihood estimation as a convex surrogate for learning the marginal inverses covariance matrix of the LVGGM. Furthermore, by leveraging some recent advances in high-dimensional asymptotic statistics, we derive the parameter learning error bounds and sample complexity for the proposed LVGGM estimator under mild conditions in high-dimensional setting. These results provide insights valuable to understanding the statistical accuracy and efficiency of the proposed algorithm. The developed theory is validated by numerical simulation on synthetic data, and the superior missing-value prediction performance of LVGGM is illustrated through experiments on a real-world

movie rating data set.

1.2.3 Detecting emerging topics in topic models

In Chapter 5, we consider the problem of detecting emerging topics in text document corpora, which are modeled by one specific type of topic model, the latent Dirichlet allocation (LDA) model.

Topic detection has many practical real-world applications, such as discovering breaking news from online articles, and identifying anomalous activities from social or behavioral data. Built upon the LDA model, we propose a hypothesis testing framework for detecting emerging topics in a new text corpus (see Figure 1.5 for illustrations). Under the null hypothesis, the testing corpus is assumed to be generated from the old LDA model learned from the historical data, while under the alternative hypothesis, the documents are assumed to be generated from a new LDA model whose topic collection consists of the old and some emerging ones. The test is performed through calculating a certain test statistic followed by thresholding it to achieve a particular level of false positive error control.

For the proposed hypothesis test, the standard generalized likelihood-ratio test is computational expensive and lacks theoretical guarantees. We propose a surrogate test which estimates a low-dimensional LDA model and calculates certain geometric distance between the estimated and old "topic polytopes" (the convex hull of the topic vectors) for the thresholding decision. By exploiting the concentration behavior of the empirical processes, we establish asymptotic consistency of the proposed surrogate test. Under certain additional conditions, we further show a set of exponential bounds for the detection errors. The theory is validated through numerical simulations on synthetic data. The performance guarantees provide confidence for using the proposed surrogate test for detecting emerging topics.

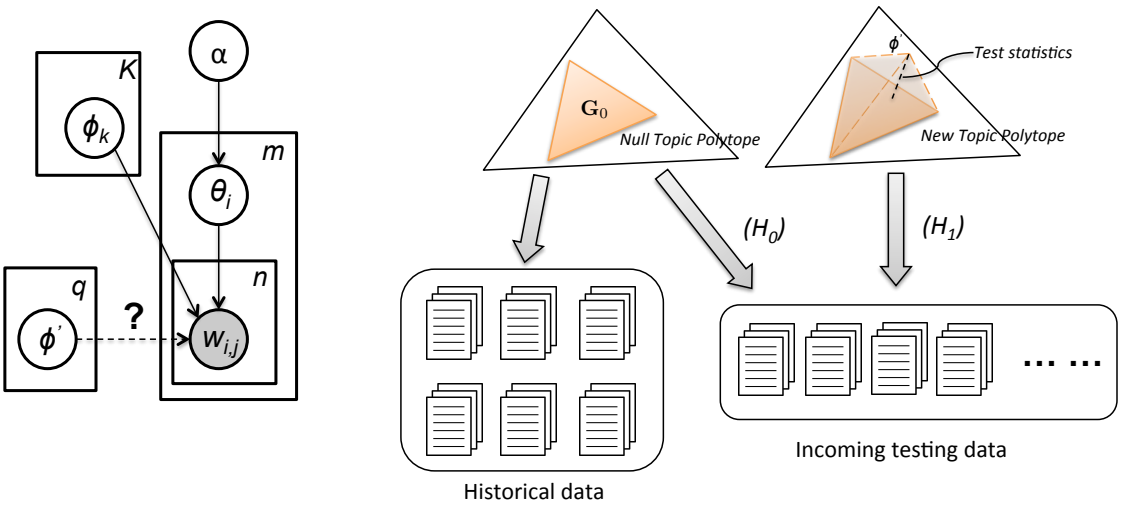


Figure 1.5: Illustrations of the emerging topic detection problem: the graphical model representation (left) and the proposed hypothesis testing framework based on topic polytopes (right).

Finally, in Chapter 6, I conclude the dissertation with a summary of the contributions and a discussion of future directions.

List of relevant publications:

1. Z. Meng, X. Nguyen, A.O. Hero III, Q. Mei, "Detecting Emerging Topics in Topic Models with Confidence", Technical Report, 2014.
2. Z. Meng, D. Wei, A. Wiesel, A.O. Hero III, "Marginal Likelihoods for Distributed Parameter Estimation of Gaussian Graphical Models", IEEE Transactions on Signal Processing, 2014.
3. Z. Meng, B. Eriksson, A.O. Hero III, "Learning Latent Variable Gaussian Graphical Models", International Conference on Machine Learning (ICML), 2014.
4. J. Tang, Z. Meng, X. Nguyen, Q. Mei, M. Zhang, "Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis", International Conference on Machine Learning (ICML), 2014. *Best Paper Award*
5. Z. Meng, D. Wei, A. Wiesel, A.O. Hero III, "Marginal Likelihoods for Distributed Estimation of Graphical Model Parameters", IEEE International

Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2013. *Best Student Paper Award*

6. Z. Meng, D. Wei, A. Wiesel, A.O. Hero III, "Distributed Learning of Gaussian Graphical Models via Marginal Likelihoods." International Conference on Artificial Intelligence and Statistics (AISTATS), 2013. *Notable Paper Award*

7. Z. Meng, A. Wiesel, A.O. Hero III, "Distributed Principal Component Analysis on networks via directed graphical models", IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2012.

CHAPTER 2

Distributed Parameter Estimation in Graphical Models via Marginal Likelihoods

In this chapter, we consider the problem of estimating the parameters of the inverse covariance matrix, also known as the concentration or precision matrix, in Gaussian graphical models.

Traditional centralized estimation often requires global inference of the covariance matrix, which can be computationally intensive in large dimensions. Approximate inference based on message-passing algorithms, on the other hand, can lead to unstable and biased estimation in loopy graphical models. In this chapter, we propose a general distributed estimation framework based on a maximum marginal likelihood (MML) approach. This approach computes local parameter estimates by maximizing marginal likelihoods defined with respect to data collected from local neighborhoods. Due to the non-convexity of the MML problem, we introduce and solve a convex relaxation. The local estimates are then combined into a global estimate without the need for iterative message-passing between neighborhoods. The proposed algorithm is naturally parallelizable and computationally efficient, thereby making it suitable for high-dimensional problems.

We provide asymptotic analyses for the proposed distributed estimator, which show

that it is asymptotically consistent, improves monotonically as the local neighborhood size increases, and has a high-dimensional error convergence rate that is comparable to the centralized maximum likelihood estimation. Extensive numerical experiments demonstrate the improved performance of the two-hop version of the proposed estimator, which suffices to almost close the gap to the centralized maximum likelihood estimator at a reduced computational cost.

2.1 Introduction

Graphical models provide a principled framework for compactly characterizing dependencies among many random variables, represented as nodes in a network (Lauritzen, 1996; Wainwright & Jordan, 2008). Their sparse structure allows for efficient and distributed inference using message-passing algorithms such as loopy belief propagation (LBP), which makes them especially well-suited to large networks, such as sensor, social, and biological networks (Liu & Ihler, 2012; Wiesel & Hero, 2012; Meng et al., 2012). Less well-studied, however, is the equally important task of distributed estimation of the parameters of a graphical model from data. The goal of this work is to develop and analyze distributed methods for model parameter estimation.

In this chapter we focus on Gaussian graphical models (GGM) with known graph structure, i.e, the pattern of edges is known. Our approach can also be extended to more general graphical models, including discrete distributions. For GGMs, parameter estimation essentially reduces to (inverse) covariance estimation, and knowledge of the edge pattern imposes sparsity constraints on the inverse covariance matrix, also known as the concentration or precision matrix. While the resulting GGM maximum likelihood (ML) parameter estimation problem is a convex optimization, solving it exactly for generally structured networks using centralized algorithms as in Banerjee et al. (2006); Dahl et al. (2008); Friedman et al. (2008) becomes impractical in large

real-world networks where data collection and computational resources are limited.

A natural approach toward distributed parameter estimation is to leverage methods for distributed marginal inference, such as LBP and its extensions. The idea is to replace the objective function and its gradient in the ML estimation problem with approximations that can be computed through iterative message-passing. However, in many cases LBP may fail to converge or give good marginal estimates, and when it does converge, the resulting parameter estimates may be biased (Malioutov et al., 2006; Heinemann & Globerson, 2012).

Another direction for distributed estimation is to consider a surrogate objective that decomposes into smaller problems that are locally parameterized. Then a distributed ML algorithm estimates the local parameters by processing local data with limited message passing. Some recent efforts along this direction (Wiesel & Hero, 2012; Liu & Ihler, 2012) have considered a pseudo-likelihood framework for exponential family distributions.

In this chapter, we proposes a general framework for distributed estimation based on *marginal* likelihoods, as contrasted with pseudo-likelihoods. Each node collects data within its extended neighborhood and independently forms a local estimate by maximizing a marginal likelihood. To deal with the non-convexity of the maximum marginal likelihood (MML) estimation problem, we formulate a convex relaxation of the problem. The resulting distributed estimator is computationally efficient, and involves minimal message passing.

We analyze the mean squared error (MSE) of the proposed distributed estimator in both the classical asymptotic regime (fixed number of parameters p and increasing number of samples $T \to \infty$), and also the high-dimensional regime where both p and T increase to infinity ($p, T \to \infty$). In the classical regime, the distributed estimator is shown to be asymptotically consistent. Furthermore, the asymptotic error improves monotonically as the local neighborhood size increases. In the high-

dimensional regime, we show that under certain conditions and proper scaling between p and T, the proposed estimator achieves a comparable statistical convergence rate to the (more expensive) global ML estimator.

Our analytical results are supported by extensive numerical experiments on both synthetic and real-world data sets. In particular, we show that two-hop local information is sufficient for the proposed distributed estimator to match the performance of the more expensive centralized ML estimator. The proposed estimator also improves significantly upon existing distributed estimators (Liu & Ihler, 2012; Wiesel & Hero, 2012). In terms of computation, the complexity of our estimator increases at most linearly with p in most cases and can be further reduced through parallelization. In the case of a physical network implementation, the near-absence of message passing and long-distance communication is also an advantage.

We emphasize that the problem we consider is different from covariance selection (Ravikumar et al., 2011; Rothman et al., 2008; Johnson et al., 2011; Friedman et al., 2008), in which the graph topology is not known *a priori* and must be estimated in addition to the parameters. To test our assumption of known graph structure, we also study the robustness of the proposed estimators against small model (i.e. structure) mismatch. Both theoretical analysis and numerical results show that the proposed distributed estimator is as robust as the centralized ML estimator.

The outline of the chapter is as follows. In Section 2.2, we give a brief review of graphical models, centralized ML parameter estimation, and the difficulty of parameter estimation using traditional marginal inference techniques. In Section 2.3, we propose a general approach to distributed estimation based on marginal likelihoods. In Section 2.4, we provide extensive analysis of the convergence rates and robustness of the proposed estimator. Section 2.5 discusses the computational complexity and implementation advantages of the estimator. Numerical experiments are presented in Section 2.6 and we conclude the chapter with a summary in Section

2.7.

Notation for Chapter 2. Boldface upper case letters denote matrices and boldface lower case letters denote column vectors. Sets of single indices are denoted by calligraphic upper case letters. The cardinality of a set $\mathcal{A}$ is denoted by $|\mathcal{A}|$ and the difference of two sets is denoted as $\mathcal{A}\backslash\mathcal{B}$. Following common notation, $\mathbf{A}_{\mathcal{M},\mathcal{N}}$ represents a submatrix of $\mathbf{A}$ with rows indexed by $\mathcal{M}$ and columns indexed by $\mathcal{N}$. We also make reference to irregular sets of index pairs such as the edge set E of a graph, for which we use standard upper case letters. $\mathbf{A}_E$ then refers to the vector of entries of $\mathbf{A}$ indexed by E. The standard inner product between two symmetric matrices is denoted as $\langle \mathbf{A}, \mathbf{B} \rangle$, i.e., $\langle \mathbf{A}, \mathbf{B} \rangle = \text{trace}(\mathbf{AB}) = \sum_{i,j} \mathbf{A}_{i,j}\mathbf{B}_{i,j}$. We distinguish the following two norms for matrices: the induced ℓ_∞/ℓ_∞ norm $|||\mathbf{A}|||_\infty := \max_{i=1,\ldots,p} \sum_{j=1}^{p} |\mathbf{A}_{i,j}|$, and the element-wise ℓ_∞ norm $\|\mathbf{A}\|_\infty := \max_{i,j=1,\ldots,p} |\mathbf{A}_{i,j}|$. $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ denote the maximum and minimum eigenvalues of matrix $\mathbf{A}$, respectively.

2.2 Background

We begin by providing background on graphical models and their statistical inference. We refer the reader to Lauritzen (1996); Wainwright & Jordan (2008) for a detailed treatment.

2.2.1 Gaussian Graphical Models

Consider a p-dimensional random vector $\mathbf{x}$ following a graphical model with respect to an undirected graph $\mathcal{G} = (V, E)$, where $V = \{1, \ldots, p\}$ is a set of nodes corresponding to elements of $\mathbf{x}$ and E is a set of edges connecting nodes. The vector $\mathbf{x}$ satisfies the Markov property with respect to $\mathcal{G}$ if for any pair of nonadjacent nodes in $\mathcal{G}$, the corresponding pair of variables in $\mathbf{x}$ are conditionally independent given the remaining

variables.

If the vector $\mathbf{x}$ follows a multivariate Gaussian distribution, the corresponding model is called a Gaussian graphical model (GGM). We assume without loss of generality that $\mathbf{x}$ has zero mean. Then the probability density function can be written in canonical form in terms of the concentration matrix $\mathbf{J}$ as follows:

$$p(\mathbf{x}; \mathbf{J}) = (2\pi)^{-p/2} (\det \mathbf{J})^{1/2} \exp\left(-\frac{1}{2}\mathbf{x}^T \mathbf{J} \mathbf{x}\right). \tag{2.1}$$

The Markov property manifests itself in a simple way through the sparsity pattern of $\mathbf{J}$:

$$\mathbf{J}_{i,j} = 0 \text{ for all } i \neq j, (i,j) \notin E. \tag{2.2}$$

2.2.2 Maximum Likelihood Parameter Estimation for GGMs

Estimating the parameters of a graphical model from sample data is the first step for many applications. For Gaussian graphical models this reduces to estimating the non-zero elements of the concentration matrix $\mathbf{J}$ (including the diagonal elements). Defining

$$\widetilde{E} := E \cup \{(i,i)\}_{i=1}^{p} \tag{2.3}$$

as the index set for these non-zero elements, the centralized global maximum likelihood (GML) estimation problem can be formulated as (Lauritzen, 1996):

$$\begin{aligned} \widehat{\mathbf{J}}^{\text{GML}} &= \arg\min_{\mathbf{J}} \ \langle \widehat{\mathbf{\Sigma}}, \mathbf{J} \rangle - \log\det \mathbf{J} \\ \text{s.t.} \quad & \mathbf{J}_{j,k} = 0 \quad \forall\, (j,k) \notin \widetilde{E} \\ & \mathbf{J} \succeq \mathbf{0}, \end{aligned} \tag{2.4}$$

where

$$\widehat{\mathbf{\Sigma}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}(t)\mathbf{x}(t)^T$$

is the sample covariance matrix and $\mathbf{x}(1), \ldots, \mathbf{x}(T)$ are i.i.d. samples of $\mathbf{x}$.

The GML problem (2.4) is a convex log-determinant-regularized semidefinite program ($\log\det$-SDP) with respect to $\mathbf{J}_{\widetilde{E}}$ and various gradient-based algorithms can be applied to solve this problem, many of which have specialized implementations on graphs, *e.g.* iterative proportional fitting (IPF) (Wainwright & Jordan, 2008), chordally-embedded Newton's method (Dahl et al., 2008), etc. The standard gradient descent algorithm for solving problem (2.4) has the following update rule at each iteration:

$$\begin{aligned}\widehat{\mathbf{J}}_{i,j}^{(t+1)} &\leftarrow \widehat{\mathbf{J}}_{i,j}^{(t)} - \gamma\cdot\nabla\ell(\widehat{\mathbf{J}}^{(t)})_{i,j} \\ &= \begin{cases}\widehat{\mathbf{J}}_{i,j}^{(t)} - \gamma\cdot\left(2\widehat{\mathbf{\Sigma}}_{i,j} - 2(\widehat{\mathbf{J}}^{(t)})_{i,j}^{-1}\right), & i\neq j \\ \widehat{\mathbf{J}}_{i,j}^{(t)} - \gamma\cdot\left(\widehat{\mathbf{\Sigma}}_{i,j} - (\widehat{\mathbf{J}}^{(t)})_{i,j}^{-1}\right), & i = j\end{cases}\end{aligned} \tag{2.5}$$

where $\ell(\mathbf{J})$ is the GML objective function and $\nabla\ell(\mathbf{J})$ denotes its gradient, γ is the step-size, and we have used the facts $\frac{\partial\log\det\mathbf{X}}{\partial\mathbf{X}_{i,j}} = 2(\mathbf{X}^{-1})_{i,j}$ for $i\neq j$ and $\frac{\partial\log\det\mathbf{X}}{\partial\mathbf{X}_{i,i}} = (\mathbf{X}^{-1})_{i,i}$ for symmetric matrices (Petersen et al., 2006). The obvious difficulty is the global matrix inversion involved in computing the gradient at each step, whose computational cost is cubic in the number of variables for generally structured models.

Given the expense of the matrix inversion in (2.5), an alternative is to consider distributed message-passing algorithms, such as loopy belief propagation (LBP), an iterative message-passing algorithm for inference of marginal distributions. When applied to tree-structured graphs, LBP yields exact marginals. Unfortunately, this does not hold for loopy graphs in general (Murphy et al., 1999). For Gaussian models, many *sufficient* conditions exist for Gaussian LBP to converge, such as diagonal dominance, walk-summablility, pairwise normalizability, etc. (Malioutov et al., 2006).

However, when these sufficient conditions do not hold, Gaussian LBP can be divergent, or it may converge to degenerate, unnormalized marginal distributions. A recent work by Pacheco & Sudderth (2012) uses the method of multipliers to improve the convergence behavior of Gaussian LBP for some less ill-conditioned models. However, even if LBP converges, its final estimate is not guaranteed to be consistent. For discrete graphical models, Heinemann & Globerson (2012) show that many models are in principle not learnable through LBP, which implies that an estimator based on LBP inference is inevitably biased for a subset of models. Similar drawbacks also hold when using other approximate inference techniques, for example, tree-reweighted BP (Wainwright, 2006). The above difficulties of parameter estimation using traditional marginal inference techniques motivate us to consider a different distributed framework for parameter estimation, as introduced in the next section.

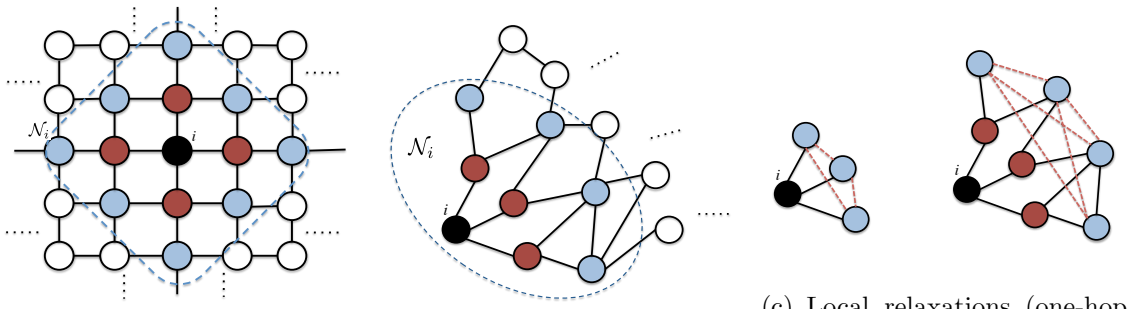
2.3 Distributed Estimation in GGMs

Our framework avoids the weakness of LBP and other message passing approaches to distributed estimation of GGMs. The proposed distributed algorithm collects all the data samples from within each neighborhood and computes a local parameter estimate. A global estimate of the parameter (e.g. precision matrix $\mathbf{J}$) is then formed by combining these local estimates with a simple, single pass aggregation rule.

2.3.1 Marginal Likelihood Maximization

We consider estimating local parameters by maximizing *marginal likelihood* functions in neighborhoods around each node. Define the index set for *immediate neighbors* of node i as

$$\mathcal{I}_i := \{j \mid (i,j) \in E\}, \tag{2.6}$$



(a) 2D lattice and two-hop neighborhood $\mathcal{N}_i$

(b) A general graph and two-hop neighborhood $\mathcal{N}_i$

(c) Local relaxations (one-hop (left) and two-hop (right)). Dotted lines denote fill-in edges.

Figure 2.1: Illustration of defined sets in the proposed local relaxation of MML. In (a) and (b) we show two different graphs, in which the two-hop neighborhood $\mathcal{N}$ for node i is indicated with dashed contours. The buffer set variables $\mathbf{x}_{\mathcal{B}}$ and the protected set variables $\mathbf{x}_{\mathcal{P}}$ (excluding node i itself) are colored blue and red, respectively. For the graph in (b), we illustrate the one-hop and two-hop local relaxations in (c). The dashed red lines in (c) denote the fill-in edges due to relaxation.

and consider a neighborhood indexed by a set $\mathcal{N}_i$ containing at least the node i itself and its immediate neighbors $\mathcal{I}_i$. Let $\mathbf{K}$ denote the concentration matrix corresponding to the marginal distribution over the variables $\{\mathbf{x}_j, j \in \mathcal{N}_i\}$ in the neighborhood, and let $\widehat{\mathbf{\Sigma}}_{\mathcal{N}_i,\mathcal{N}_i} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{\mathcal{N}_i}(t)\mathbf{x}_{\mathcal{N}_i}(t)^T$ be the marginal sample covariance matrix. The maximum marginal likelihood (MML) estimation problem in neighborhood $\mathcal{N}_i$ can be formulated as:

$$\begin{aligned} \widehat{\mathbf{K}}^{i,\mathrm{MML}} = \underset{\mathbf{K},\mathbf{J}}{\arg\min}\ & \langle \widehat{\mathbf{\Sigma}}_{\mathcal{N}_i,\mathcal{N}_i}, \mathbf{K} \rangle - \log\det \mathbf{K} \\ \text{s.t.}\quad & \mathbf{K} = \left[\left(\mathbf{J}^{-1} \right)_{\mathcal{N}_i,\mathcal{N}_i} \right]^{-1}, \\ & \mathbf{J}_{j,k} = 0 \quad \forall\, (j,k) \notin \widetilde{E}, \\ & \mathbf{J} \succeq \mathbf{0}, \end{aligned} \tag{2.7}$$

where the first constraint represents the marginalization relationship between $\mathbf{K}$ and the global precision matrix $\mathbf{J}$, and the second line of constraints reflects the global sparsity constraints. We index the nodes in the MML problem (2.7) in the same way as in the GML problem (2.4). (For example, if $\mathcal{N}_1 = \{1, 3, 6\}$, the rows and columns

of $\mathbf{K}$ are indexed by $\{1, 3, 6\}$ and not re-indexed to $\{1, 2, 3\}$.)

The difficulty with direct application of MML is that problem (2.7) is in general a non-convex optimization with respect to $\mathbf{K}$ and $\mathbf{J}$. The non-convexity arises from the coupling of the nonlinear marginalization constraint linking $\mathbf{K}$ to $\mathbf{J}$ and the sparsity constraints on $\mathbf{J}$. As a surrogate, we derive next a convex relaxation of the MML estimation problem.

2.3.2 Convex Relaxation of MML

We apply the Schur complement identity to the marginalization constraint in (2.7), yielding

$$\mathbf{K} = \mathbf{J}_{\mathcal{N},\mathcal{N}} - \mathbf{J}_{\mathcal{N},\mathcal{N}^C} \cdot \left[\mathbf{J}_{\mathcal{N}^C,\mathcal{N}^C}\right]^{-1} \cdot \mathbf{J}_{\mathcal{N}^C,\mathcal{N}}, \tag{2.8}$$

where $\mathcal{N}^C$ is the complementary set to $\mathcal{N}$, and we have dropped the subscript i to simplify notation. Define the *buffer set* $\mathcal{B} \subset \mathcal{N}$ as the set of all variables in $\mathcal{N}$ that have immediate neighbors in the complement $\mathcal{N}^C$,

$$\mathcal{B} := \{j \mid j \in \mathcal{N} \text{ and } \mathcal{I}_j \cap \mathcal{N}^C \neq \emptyset\}. \tag{2.9}$$

The difference set between $\mathcal{N}$ and $\mathcal{B}$ is referred to as the *protected set* $\mathcal{P} := \mathcal{N} \backslash \mathcal{B}$. The buffer and protected sets are illustrated in Figure 2.1(a) and 2.1(b). Due to the Markov property, we have $\mathbf{J}_{\mathcal{P},\mathcal{N}^C} = \mathbf{0}$. Decomposing $\mathcal{N}$ into $\mathcal{B}$ and $\mathcal{P}$ then reveals the

sparsity pattern of $\mathbf{K}$ using (2.8):

$$\begin{aligned}\mathbf{K} &= \mathbf{J}_{\mathcal{N},\mathcal{N}} - \begin{bmatrix} \mathbf{0} \\ \mathbf{J}_{\mathcal{B},\mathcal{N}^C} \end{bmatrix} \left[\mathbf{J}_{\mathcal{N}^C,\mathcal{N}^C}\right]^{-1} \begin{bmatrix} \mathbf{0}, & \mathbf{J}_{\mathcal{N}^C,\mathcal{B}} \end{bmatrix}, \\ &= \mathbf{J}_{\mathcal{N},\mathcal{N}} - \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_{\mathcal{B},\mathcal{N}^C} \left[\mathbf{J}_{\mathcal{N}^C,\mathcal{N}^C}\right]^{-1} \mathbf{J}_{\mathcal{N}^C,\mathcal{B}} \end{bmatrix}\end{aligned}$$

and hence

$$\mathbf{K}_{\mathcal{P},\mathcal{P}} = \mathbf{J}_{\mathcal{P},\mathcal{P}}, \ \mathbf{K}_{\mathcal{P},\mathcal{B}} = \mathbf{J}_{\mathcal{P},\mathcal{B}}, \tag{2.10}$$

$$\mathbf{K}_{\mathcal{B},\mathcal{B}} = \mathbf{J}_{\mathcal{B},\mathcal{B}} - \mathbf{J}_{\mathcal{B},\mathcal{N}^C} \left[\mathbf{J}_{\mathcal{N}^C,\mathcal{N}^C}\right]^{-1} \mathbf{J}_{\mathcal{N}^C,\mathcal{B}}. \tag{2.11}$$

An important observation from (2.10) is that in the rows and columns indexed by the protected set $\mathcal{P}$, the sparsity pattern of $\mathbf{J}_{\mathcal{N},\mathcal{N}}$ is entirely preserved and the local parameters are equal to the global ones. On the other hand, the sparsity pattern in the "buffer submatrix" $\mathbf{K}_{\mathcal{B},\mathcal{B}}$ is in general modified from $\mathbf{J}_{\mathcal{B},\mathcal{B}}$ due to the fill-in term, *i.e.*, the second term in (2.11).

Based on these observations, we now specify a relaxed set of constraints on the marginal concentration matrix $\mathbf{K}$. First denote the set of all local edges that are not affected by the fill-in term in (2.11) as

$$E^{\text{Prot}} := \widetilde{E} \cap \{\{\mathcal{P} \times \mathcal{P}\} \cup \{\mathcal{P} \times \mathcal{B}\} \cup \{\mathcal{B} \times \mathcal{P}\}\}, \tag{2.12}$$

where the superscript stands for "protected". We then add to E^{Prot} all index pairs $\mathcal{B} \times \mathcal{B}$ that could potentially be affected by fill-in in (2.11), resulting in a *relaxed edge set* R (see Figure 2.1(c) for illustrations):

$$R = E^{\text{Prot}} \cup \{\mathcal{B} \times \mathcal{B}\}. \tag{2.13}$$

In light of (2.10) and (2.11), any feasible marginal concentration matrix $\mathbf{K}$ for the MML estimation problem (2.7) is guaranteed to be supported only on the set R. Therefore we can relax the feasible set and formulate the following relaxation of (2.7) at each node i, called the relaxed MML (RMML) problem:

$$\begin{aligned} \widehat{\mathbf{K}}^{i,\text{Relax}} = \arg\min_{\mathbf{K}} \ & \langle \widehat{\mathbf{\Sigma}}_{\mathcal{N}_i,\mathcal{N}_i}, \mathbf{K} \rangle - \log\det \mathbf{K} \\ \text{s.t.} \ \ & \mathbf{K}_{j,k} = 0 \quad \forall\, (j,k) \notin R \\ & \mathbf{K} \succeq \mathbf{0}. \end{aligned} \tag{2.14}$$

The above RMML problem is a convex optimization with respect to $\mathbf{K}$ and has the same form as the global MLE problem (2.4) but over matrices of much lower dimension.

After solving the RMML estimation problems as surrogates to estimate local parameters, a global estimate of the concentration matrix can then be constructed by extracting a subset of parameters from each local estimate and concatenating them. Specifically, we extract the local parameter estimates indexed by

$$L_i := \{(j,k) \in \widetilde{E} \mid j = i\}, \tag{2.15}$$

i.e., the non-zero entries in the ith row of $\mathbf{J}$. We refer to the parameters indexed by L_i as the *row parameters* for node i. From (2.10), when there are no sampling errors, i.e. $T \to \infty$, the marginal and global concentration matrices are guaranteed to share the same parameters in L_i. Therefore our global estimate of $\mathbf{J}$ is formed by concatenating local solutions of (2.14):

$$\widehat{\mathbf{J}}^{\text{Relax}}_{L_i} \leftarrow \widehat{\mathbf{K}}^{i,\text{Relax}}_{L_i}, \quad \text{for } i = 1, \ldots, p. \tag{2.16}$$

The proposed RMML framework is very general and applies to many possible

choices of local neighborhoods, which include, e.g., nearest neighbors, second-order nearest neighbors, or, in general, k-th order nearest neighbors of a node i. In the following subsections, we consider one- and two-hop neighborhoods. The absence of sampling errors is still assumed, i.e. $T \to \infty$.

2.3.3 Case I: One-hop Estimator

We first consider a first-order (i.e., one-hop) neighborhood consisting of node i and its immediate neighbors $\mathcal{I}_i$, i.e., $\mathcal{N}_i = \{i\} \cup \mathcal{I}_i$. Generically in the worst case where the immediate neighbors are all buffer nodes, we have $\mathcal{B}_i = \mathcal{I}_i$, and $\mathcal{P}_i = \{i\}$. The fill-in term in (2.11) affects the submatrix $\mathbf{K}_{\mathcal{I}_i,\mathcal{I}_i}$, leaving only the first row and column untouched. In this case, since i is by definition connected to all elements in $\mathcal{I}_i$, the relaxed edge set R_i defined in (2.13) includes all possible pairs (see leftmost graph of Figure 2.1(c) for an illustration): $R_i^{\text{1hop}} = \mathcal{N}_i \times \mathcal{N}_i$.

The solution to the relaxed MML problem (2.14) for the first-order neighborhood is simply the inverse of the local sample covariance,

$$\widehat{\mathbf{K}}^{i,\text{1hop}} = \left(\widehat{\mathbf{\Sigma}}_{\mathcal{N}_i,\mathcal{N}_i}\right)^{-1}. \tag{2.17}$$

The global estimate is obtained by combining the local one-hop estimates as in (2.16).

In the one-hop case, the proposed relaxed MML estimator reduces to the LOC estimator in Wiesel & Hero (2012). As shown in Wiesel & Hero (2012), this estimator is also equivalent to the pseudolikelihood estimator (Liang & Jordan, 2008) without symmetry constraints, and the covariance selection procedure in Friedman et al. (2008) when the graph is known.

2.3.4 Case II: Two-hop Estimator

We next consider a second-order neighborhood (two-hop), $\mathcal{N}_i$ that includes nodes that are reachable from node i within two hops. In this setting, the worst-case protected set is given by $\mathcal{P}_i = \{i\} \cup \mathcal{I}_i$ and the buffer set $\mathcal{B}_i = \mathcal{N}_i \backslash \mathcal{P}_i$ consists of all nodes that are exactly two hops away from the ith node. Hence $\mathcal{B}_i$ can be thought of as the set of *second-hop* nodes. In the two-hop case, the protected edge set E^{Prot} includes not only edges between node i and its immediate first-hop neighbors, but also edges between first-hop neighbors and between first- and second-hop neighbors (see Figure 2.1(c) for an illustration).

Unlike in the one-hop case, the two-hop problem (2.14) does not admit a general closed-form solution. However, as mentioned before, Eq. (2.14) can be solved using efficient algorithms for semidefinite programming. A global estimate is obtained as before by combining row parameter estimates (2.16).

2.3.5 Symmetrization of RMML Estimator

When $\widehat{\mathbf{\Sigma}}$ is estimated from finite sample sizes, the local estimates from the relaxed MML problems are not perfectly consistent with each other. For example, $\widehat{\mathbf{J}}^{\text{Relax}}_{i,j}$, which comes from node i's local estimate, may not agree with $\widehat{\mathbf{J}}^{\text{Relax}}_{j,i}$, which comes from node j's local estimate. Therefore the resulting global estimate $\widehat{\mathbf{J}}^{\text{Relax}}$ in (2.16) is not guaranteed to be symmetric.

A common way of addressing these discrepancies is to use iterative consensus methods as in Wiesel & Hero (2012); Liu & Ihler (2012). In this work however, we find that a single round of naive local averaging along edges is sufficient to ensure convergence to the true parameters, and also to yield a good approximation to the

global MLE. Specifically, the local average is given by

$$\widehat{\mathbf{J}}_{i,j}^{\text{Relax}} \leftarrow \frac{1}{2}(\widehat{\mathbf{J}}_{i,j}^{\text{Relax}} + \widehat{\mathbf{J}}_{j,i}^{\text{Relax}}), \quad (i,j) \in E, \tag{2.18}$$

which is the only message passing required. This message passing is single pass, unlike LBP which requires several iterations (if it converges at all). In the one-hop case, the resulting symmetric estimator coincides with the AVE estimator proposed in Wiesel & Hero (2012).

2.4 Analysis

2.4.1 Asymptotic Analysis: Classical Fixed-Dimensional Regime

First we analyze the proposed distributed RMML estimator in the classical asymptotic regime, where the number of variables p is fixed while the number of samples T goes to infinity. Let $\mathbf{J}^*$ and $\mathbf{\Sigma}^*$ denote the true precision and covariance matrices, respectively. The following theorem states the asymptotic consistency of the RMML estimator $\widehat{\mathbf{J}}^{\text{Relax}}$ and characterizes its asymptotic mean squared error:

Theorem 2.1 (Asymptotic MSE)**.** *The relaxed MML estimator $\widehat{\mathbf{J}}^{Relax}$ is asymptotically consistent, and its mean squared (Frobenius) error satifies*

$$T \cdot \mathbb{E}\|\widehat{\mathbf{J}}^{Relax} - \mathbf{J}^*\|_F^2 \overset{T\to\infty}{\longrightarrow} \sum_{i=1}^{p} \sum_{j \in L_i} [diag\left(\mathbf{F}_i^{-1}\right)]_j, \tag{2.19}$$

where T is the number of samples, $diag(\cdot)$ denotes the diagonal of a matrix, and $\mathbf{F}_i$ is the Fisher information matrix of the relaxed MML problem in the ith neighbor-

hood (2.14)*, which takes the following form:*

$$(\mathbf{F}_i)_{(m,n),(l,k)} = \begin{cases} 2{\boldsymbol{\Sigma}^*_{m,l}}^2, & m = n \text{ and } l = k \\ 2\boldsymbol{\Sigma}^*_{m,k}\boldsymbol{\Sigma}^*_{l,n}, & m = n, l \neq k \text{ or } m \neq n, l = k \\ \boldsymbol{\Sigma}^*_{m,k}\boldsymbol{\Sigma}^*_{n,l}, & \text{otherwise.} \end{cases} \tag{2.20}$$

The above result can be derived by applying classical asymptotic theory (Van der Vaart, 2000) to each local RMML problem (2.14), which is a well-defined M-estimation problem. Then the asymptotic behavior of the global RMML estimate follows by aggregation. The detailed proof of Theorem 2.1 is provided in Section 2.8.1.

While Theorem 2.1 ensures the consistency of RMML estimators with arbitrary local neighborhoods (as long as the row parameters are included), the following theorem guarantees that, in the asymptotic limit, larger neighborhoods always yield reduced estimation variance:

Theorem 2.2 (Monotonicity of Asymptotic MSE)**.** *Let* $\widehat{\mathbf{J}}^{\textit{Relax, k-hop}}$ *be the RMML estimate obtained from k-hop local neighborhoods. When the number of samples* $T \to \infty$*, for* $k = 1, 2, \ldots$*, we have*

$$\mathbb{E}\|\widehat{\mathbf{J}}^{\textit{Relax, k-hop}} - \mathbf{J}^*\|_F^2 \geq \mathbb{E}\|\widehat{\mathbf{J}}^{\textit{Relax, (k+1)-hop}} - \mathbf{J}^*\|_F^2 \tag{2.21}$$

$$\geq \mathbb{E}\|\widehat{\mathbf{J}}^{GML} - \mathbf{J}^*\|_F^2. \tag{2.22}$$

While Theorem 2.2 is stated for Gaussian graphical models, it was first proven for the case of discrete graphical models by Massam and Wang in Massam & Wang (2013). As pointed out by Massam & Wang (2013), their proof can be easily extended to the Gaussian case. For completeness, we include our own proof of Theorem 2.2 in Section 2.8.2. The two proofs follow parallel lines of argument.

In Section 2.6, we present numerical results that verify Theorem 2.2 not only in

the large-sample regime but also when the sample size T is comparable to or smaller than p. In particular, it will be seen that the difference between $k = 1$ and $k = 2$ hops is most significant while the difference between $k = 2$ and the GML estimator (and by extension $k > 2$ and GML) is much smaller.

2.4.2 Asymptotic Analysis: High-Dimensional Regime

Theorems 2.1 and 2.2 characterize the classical asymptotic behavior of the RMML estimator. In this subsection we analyze the high-dimensional convergence rate of the RMML estimator, which can be applied to settings where both the number of variables p and the number of samples T increase to infinity, i.e. $p, T \to \infty$. Such problems arise in high-dimensional applications, and have attracted much attention in modern statistics (Ravikumar et al., 2011; Friedman et al., 2008; Rothman et al., 2008). We will show that under very mild conditions, the proposed RMML estimator enjoys a sharp MSE convergence rate to the true parameter, which is almost the same as the more expensive global ML estimator.

Similar to Rothman et al. (2008); Ravikumar et al. (2011), we first assume that the maximum eigenvalue of $\mathbf{J}^*$ is bounded from above:

$$\lambda_{\max}(\mathbf{J}^*) \leq \overline{\kappa} < \infty. \tag{2.23}$$

Recall that R_i defines the relaxed edge set in the i^{th} local neighborhood. Let $\overline{R}$ denote the maximum cardinality among all local relaxed edge sets, i.e.

$$\overline{R} := \max_{i=1,\ldots,p} |R_i|, \tag{2.24}$$

and let r denote the sum of the cardinalities of all local relaxed edge sets:

$$r := \sum_{i=1}^{p} |R_i|. \tag{2.25}$$

Also denote $\overline{\sigma} := \max_{i=1,\ldots,p} \mathbf{\Sigma}^*_{i,i}$ as the maximum variance.

The following theorem states an upper bound on the estimation error rate in the high-dimensional regime.

Theorem 2.3 (High-dimensional MSE)**.** *Assume the number of samples T satisfies*

$$T \geq C^2 c_1 \log p, \tag{2.26}$$

for $c_1 = 6400\overline{\sigma}^2/\min^2\{\frac{1}{9\overline{\kappa}\sqrt{\overline{R}}}, 40\overline{\sigma}\}$ and an arbitrary constant $C \geq 1$. Then

$$\|\widehat{\mathbf{J}}^{Relax} - \mathbf{J}^*\|_F \leq 720C \cdot \overline{\kappa}^2\overline{\sigma}\sqrt{\frac{r \log p}{T}}, \tag{2.27}$$

with probability greater than $1 - 4/p^{2(C^2-1)}$.

Proof of Theorem 2.3 can be found in Section 2.8.3.

Remarks:

1) It is interesting to compare the result in Theorem 2.3 with the standard convergence rate for the GML estimator (e.g., Rothman et al. (2008); Ravikumar et al. (2011); Wainwright (2009)). Theorem 2.3 assumes a very mild condition (Eq. (2.26)) on the sample size, which is less restrictive than the requirement $O(p \log p)$ shown in Rothman et al. (2008) in the high dimensional regime, and is comparable to those obtained in Ravikumar et al. (2011); Wainwright (2009) when the local neighborhood size increases more slowly than p, i.e. $\overline{R} = o(p)$. However, we emphasize that, unlike some of the literature, we assume the graph structure is known.

2) The error bound in Theorem 2.3 is (up to a constant) slightly more pessimistic

than the rate $O(\sqrt{p \log p/T})$ shown in Rothman et al. (2008); Ravikumar et al. (2011) by the additional factor of $r/p = \frac{\sum_{i=1}^{p} |R_i|}{p}$, which is roughly the average cardinality of local neighborhoods. Again, when the local neighborhood size increases more slowly than p in the high-dimensional regime, this additional factor becomes relatively insignificant.

3) The mild sample size requirement is partly due to our distributed framework, under which the stochastic deviation is smaller since a smaller set of parameters needs to be considered for each local RMML problem. However, the additional parameters introduced by convex relaxation and the aggregation of local estimation errors result in the additional factor r/p mentioned above. This demonstrates the trade-off due to the desire for distributed, convex optimization in the proposed framework.

2.4.3 Robustness Against Model Mismatch

One of the premises of the estimation framework we consider in this chapter is that the true structure of the graph is known. However, this assumption could be violated in practice. In this section, we investigate the robustness of the estimators against small structure mismatch. Our specific interest is in the bias due to model mismatch and hence we focus on the infinite sample regime.

We first consider the GML problem. The GML estimator effectively provides a mapping from the edge elements of moment (covariance) parameters $\widehat{\mathbf{\Sigma}}_{\widetilde{E}}$ to the canonical (concentration) parameters $\widehat{\mathbf{J}}^{\mathrm{GML}}_{\widetilde{E}}$. We denote this mapping as $\mathcal{M}(\cdot; \widetilde{E})$, i.e., $\widehat{\mathbf{J}}^{\mathrm{GML}}_{\widetilde{E}} = \mathcal{M}(\widehat{\mathbf{\Sigma}}; \widetilde{E})$. This mapping is specified implicitly by the optimality condition:

$$\widehat{\mathbf{\Sigma}}_{\widetilde{E}} - \left(\left(\widehat{\mathbf{J}}^{\mathrm{GML}} \right)^{-1} \right)_{\widetilde{E}} = \mathbf{0}. \tag{2.28}$$

Due to a property of minimal exponential families, $\mathcal{M}(\cdot; \widetilde{E})$ exists and is unique provided that covariance matrix $\widehat{\mathbf{\Sigma}}$ is positive definite (Wainwright & Jordan, 2008).

Also by the implicit function theorem, $\mathcal{M}(\cdot;\widetilde{E})$ is differentiable and thus continuous.

Consider a perturbed concentration matrix $\widetilde{\mathbf{J}}^*$ which has uniformly bounded perturbations on the non-edge entries with respect to the nominal parameter $\mathbf{J}^*$:

$$\widetilde{\mathbf{J}}^* = \mathbf{J}^* + \Delta\mathbf{J}, \tag{2.29}$$

where $\Delta\mathbf{J}$ is supported only on $\widetilde{E}^C$. We assume the perturbation is small enough, such that the perturbed matrix is still positive definite. Denote the corresponding covariance matrix as $\widetilde{\mathbf{\Sigma}} = (\widetilde{\mathbf{J}}^*)^{-1}$. Then the bias of the GML estimator due to model perturbation can be obtained by a first-order perturbation analysis of the GML mapping defined above.

Let $\Gamma = \mathbf{\Sigma} \otimes \mathbf{\Sigma}$ denote the Hessian of the GML problem (2.4) with no sparsity constraints, which is also related to the Jacobian of (2.28) with respect to $\widehat{\mathbf{J}}^{\mathrm{GML}}_{\widetilde{E}}$. We have

$$\begin{aligned}
\widetilde{\mathbf{J}}^{\mathrm{GML}}_{\widetilde{E}} &= \mathcal{M}(\widetilde{\mathbf{\Sigma}};\widetilde{E}) \\
&= \mathcal{M}((\widetilde{\mathbf{J}}^*)^{-1};\widetilde{E}) \\
&= \mathcal{M}((\mathbf{J}^* + \Delta\mathbf{J})^{-1};\widetilde{E}) \\
&= \mathcal{M}(\mathbf{J}^{*-1} + \Gamma_{\widetilde{E},\widetilde{E}^C}\Delta\mathbf{J}_{\widetilde{E}^C} + O(\|\Delta\mathbf{J}\|^2);\widetilde{E}) \\
&= \mathcal{M}(\mathbf{\Sigma}^*;\widetilde{E}) + (\Gamma_{\widetilde{E},\widetilde{E}})^{-1}\Gamma_{\widetilde{E},\widetilde{E}^C}\Delta\mathbf{J}_{\widetilde{E}^C} + O(\|\Delta\mathbf{J}\|^2),
\end{aligned}$$

where in the second-to-last relation we have used the first-order approximation of matrix inversion, and the last identity is due to the implicit function theorem applied to the optimality condition (2.28). Also note that $\mathbf{J}^* = \mathcal{M}(\mathbf{\Sigma}^*;\widetilde{E})$ due to consistency of the GML estimator.

Therefore the maximum element-wise bias with respect to the new model can be

bounded as follows (disregarding higher-order terms):

$$\begin{aligned}
&\|\widetilde{\mathbf{J}}^{\mathrm{GML}} - \widetilde{\mathbf{J}}^*\|_\infty \\
&\leq \|\widetilde{\mathbf{J}}^{\mathrm{GML}} - \mathbf{J}^*\|_\infty + \|\widetilde{\mathbf{J}}^* - \mathbf{J}^*\|_\infty \\
&\lesssim \left|\!\left|\!\left|(\Gamma_{\widetilde{E},\widetilde{E}})^{-1}\Gamma_{\widetilde{E},\widetilde{E}^C}\right|\!\right|\!\right|_\infty \|\Delta\mathbf{J}_{\widetilde{E}^C}\|_\infty + \|\Delta\mathbf{J}_{\widetilde{E}^C}\|_\infty,
\end{aligned} \tag{2.30}$$

where we recall $|||\cdot|||$ is the induced ∞/∞ matrix norm.

The second term in the last display is the inevitable bias due to model mismatch, while the first term captures the additional bias attributable to the GML estimator under model perturbation. The additional bias depends on $\left|\!\left|\!\left|(\Gamma_{\widetilde{E},\widetilde{E}})^{-1}\Gamma_{\widetilde{E},\widetilde{E}^C}\right|\!\right|\!\right|_\infty$, which is intuitively related to the level of *incoherence* between the edge and non-edge elements in the Hessian of the GML problem (2.4). Similar incoherence quantities have been shown to play a crucial role in the literature on variable selection (Meinshausen & Bühlmann, 2006) (e.g. Lasso) and structure estimation in Gaussian graphical models (Ravikumar et al., 2011). Therefore the smaller this incoherence parameter is, the more robust the GML estimator will be.

Since each local problem in RMML estimation has the same structure as the GML problem, we can apply similar analysis to each local neighborhood. The resulting bound on the bias of the RMML estimator is dependent on similar incoherence parameters but defined with respect to relaxed edge sets in the local neighborhoods. We conjecture that these local incoherence parameters are comparable to, if not smaller than, the global incoherence. Hence the robustness of the distributed RMML estimator is expected to be comparable to the GML estimator. While our conjecture is not formally proven in this chapter, it is positively supported by the numerical experiments in Sec. 2.6.

2.5 Computational Complexity and Implementation

In this section we discuss the computational complexity of the proposed RMML approach and some implementation issues. First we note that each local RMML problem has the same structure as the centralized ML problem, which is a log det-regularized semidefinite program (log det-SDP). Therefore many well-developed solvers and efficient specialized algorithms can be used. Furthermore, due to the distributed nature of the RMML approach, the local problems can all be solved in parallel before the final one-step averaging. The combination of lower dimensionality in the local problems and parallelization can significantly reduce the total run time compared with centralized algorithms.

In terms of algorithms, we find the iterative regression method introduced in Friedman et al. (2009) is very efficient for sparse graphs. This algorithm iteratively performs linear regressions of each node variable against its immediate neighbors until global convergence. However, the major drawback of this algorithm is the need to maintain global parameters, which prevents direct parallelization and also makes implementation difficult in distributed networks (as discussed below).

The computational advantage of the proposed RMML algorithm becomes more obvious when the number of variables p increases to large numbers. Assuming that the local neighborhood dimensions increase more slowly than p, such as with K-NN graphs and lattice graphs, the total complexity of the RMML estimator scales linearly in p, independent of the algorithm used to solve the local problems. The run time increases even more slowly if the overall algorithm can be parallelized. In contrast, for the centralized algorithms, the dependence of complexity on p is at least linear and is much faster for denser graphs and/or if generic log det-SDP solvers are used.

Another advantage of the proposed RMML algorithm is that it is highly suitable

for network applications due to its minimal requirement for message passing which reduces communication cost. In sharp contrast, many centralized algorithms, such as the iterative regression algorithm mentioned above, require centralized storage and iterative updating of a large number of variables, which in turn requires expensive communication among non-adjacent nodes in the network.

2.6 Experiments

In this section, we evaluate the proposed RMML estimator and compare it with the centralized and other distributed estimators in the literature. All methods have been coded in Matlab routines that will be available at the reproducible research web page [1]. We focus on the one-hop and two-hop versions of the RMML estimator (denoted as `RelaxMML-1hop` and `RelaxMML-2hop`, respectively). Other estimators considered in this section are:

- The centralized GML estimator, denoted as `GML` in the legends;
- The *LOCAL* and *AVE* estimators from Wiesel & Hero (2012), denoted as `LOC` and `AVE`. They coincide with the asymmetric and symmetric versions respectively of the *one-hop* relaxed MML estimator;
- The weighted maximum pseudo-likelihood estimator using Alternating Direction Method of Multipliers (ADMM) consensus, proposed in Wiesel & Hero (2012) and Liu & Ihler (2012) and denoted as `PML-ADMM`. We use the weights $\left[\hat{\mathbf{J}}_{i,i}^{LOC}\right]^2$ as in Wiesel & Hero (2012).

We first verify the classical asymptotic rates for the proposed estimators predicted by Theorems 2.1 and 2.2 (see Fig. 2.2) using 10,000 randomized runs sampled from a four-nearest-neighbor Gaussian graphical model with $p = 20$ nodes distributed

[1] http://tbayes.eecs.umich.edu/rrpapers

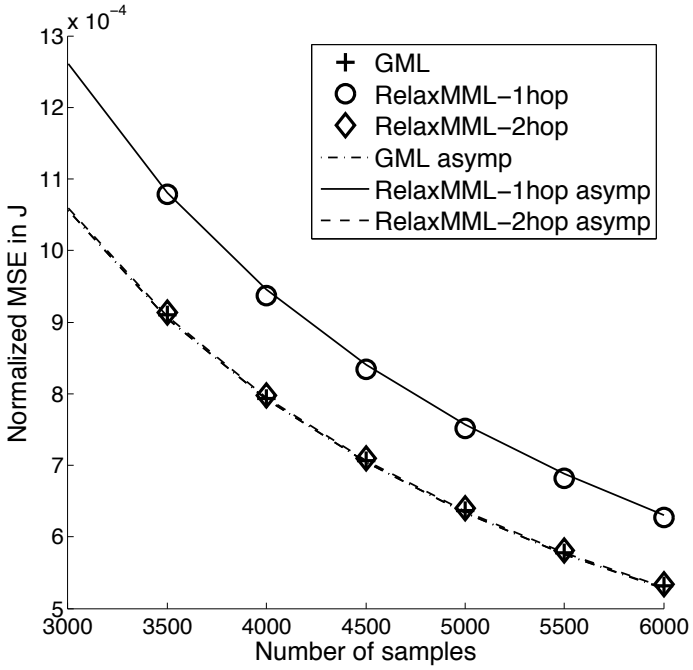


Figure 2.2: Asymptotic normalized MSE for K-NN graphs ($p = 20, K = 4$). The curves denote the theoretical asymptotic limits, whereas the symbols denote the empirical normalized MSE over 10,000 runs.

uniformly in space over the unit square. The concentration matrix is initialized as $\mathbf{J}_{i,j} = \pm \exp(-0.5 \cdot d_{i,j})$ with random sign, where $d_{i,j}$ is the Euclidean distance between the ith and jth nodes. The empirical normalized mean squared errors (MSE), defined as $\frac{\|\widehat{\mathbf{J}} - \mathbf{J}\|_F^2}{\|\mathbf{J}\|_F^2}$, are computed from Monte Carlo samples, and they are compared with the theoretical bounds predicted by Theorem 2.1. Fig. 2.2 illustrates the tightness of these bounds. It is also worth noting that the bound for the two-hop RMML estimator is much lower than that of the one-hop estimator, as predicted by Theorem 2.2. The two-hop bound approximates the bound for the GML estimator closely, suggesting that RMML estimators are nearly asymptotically efficient. The asymptotic bounds for RMML estimators with larger neighborhoods follow the monotonicity relation in Theorem 2.2, however the differences are too small to visually identify, and hence are omitted from the plot.

Next we evaluate the non-asymptotic MSE performance of the proposed estimator, and compare it with the other estimators on both synthetic and real-world data sets. For synthetic data sets, we consider three classes of graphs that are motivated by real-world applications. For each class we follow similar experiment settings as in Wiesel & Hero (2012). Specifically, we randomly generate 20 topologies and associated sparse concentration matrices $\mathbf{J}$, and for each $\mathbf{J}$, we perform 10 experiments in which random samples are drawn from the distribution and the concentration matrix is estimated from the samples. The normalized MSEs are averaged over all 200 experiments, and are reported in Figure 2.3. An illustration of the graph topology is shown in the top-right corner of each plot. The classes of graphs we consider are:

- **K-NN graphs** (Figure 2.3(a)): A K-nearest neighbor graph is a straightforward model for real-world networks whose measurements have correlations that depend on pairwise Euclidean distances, *e.g.*, sensor networks. For these experiments, we randomly generate $p = 500$ nodes uniformly over the unit square. Each node is then connected to its K-nearest neighbors, where $K = 4$. The concentration matrix is initialized as $\mathbf{J}_{i,j} = \pm \exp(-0.5 \cdot d_{i,j})$ with random sign, where $d_{i,j}$ is the Euclidean distance between the ith and jth nodes. Finally we add a small value to the diagonal to ensure positive definiteness.

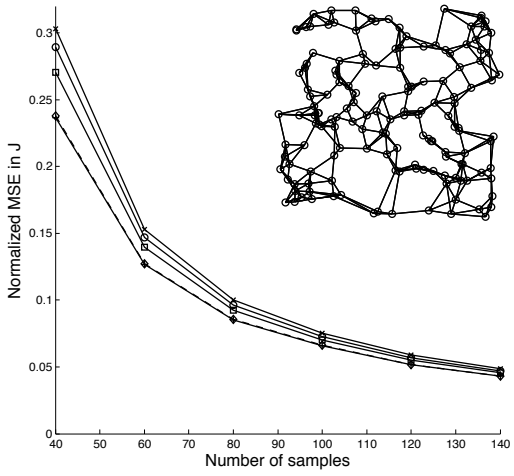
- **Lattice graphs** (Figure 2.3(b)): A lattice graph is appropriate for networks with regular spatial correlations, e.g., images that are Markov random fields. We generate a square lattice graph with $p = 20 \times 20 = 400$ nodes and edge weights generated as $\mathbf{J}_{i.j} = \min\{w, 1\}$, where w is a normally distributed random variable with mean 0.5 and variance 0.2. A small value is added to the diagonal to ensure positive definiteness.

- **Small-world graphs** (Figure 2.3(c)): Small-world graphs have been proposed for social networks, biological networks, etc., where most nodes have few imme-

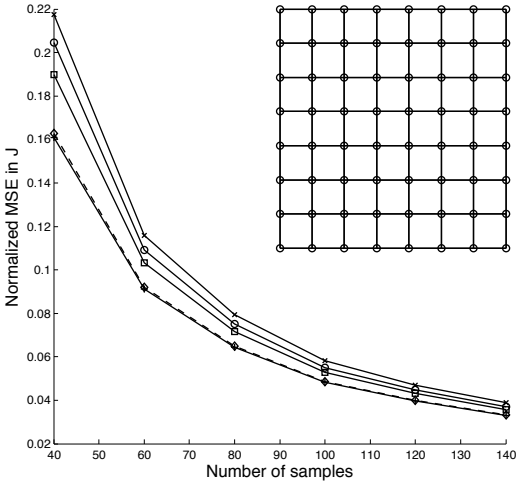
diate neighbors but can be reached from any other node through a small number of hops (Watts & Strogatz, 1998). We generate graphical models structured as random small-world networks using the Watts-Strogatz mechanism (Watts & Strogatz, 1998) with $p = 100$, K(mean degree) $= 20$, and parameter $\beta = 0.5$. Under this particular setting, a large fraction of nodes have large second-hop neighborhoods with dimension close to p. In general we expect the second-hop neighborhood to scale linearly with respect to p. We choose the edge weights to be uniformly distributed and also add a small diagonal loading to ensure that $\mathbf{J}$ is positive definite.

The MSE curves shown in Figure 2.3 match our theoretical predictions in Section 2.4.2, and they also demonstrate the superior performance of the proposed RMML estimator. In particular, for the graphs that have relatively small two-hop neighborhoods, namely the K-NN graphs and the lattice grids, the MSE of the proposed two-hop relaxed MML estimator almost coincides with the MSE of the global MLE. On the other hand, for small-world networks, the dimensions of the two-hop neighborhoods grow as fast as p. In this case, a noticeable gap emerges between the global MLE and the two-hop relaxed MML estimator. These graphs are known to be harder to learn through distributed algorithms. The two-hop relaxed MML estimator still outperforms the other distributed algorithms by a large margin.

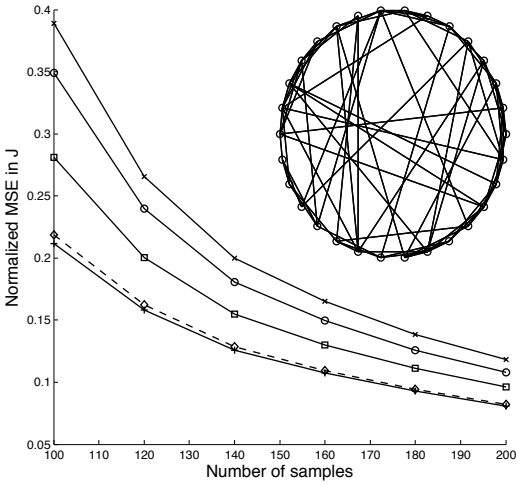
Next, we apply the estimators to a real-world sensor network. The IntelLab dataset (Guestrin et al., 2004) contains temperature information from a sensor network of 54 nodes deployed in the Intel Berkeley Research lab between February 28 and April 5, 2004. This dataset is known to be very difficult with missing data, noise and failed sensors. We select 50 sensors with relatively stable and regular measurements. To obtain a target concentration matrix, we use 1800 consecutive samples per sensor, interpolate the missing or failed readings and de-trend the data using a local rectangular window of 10 samples. Next, we compute the sample covariance



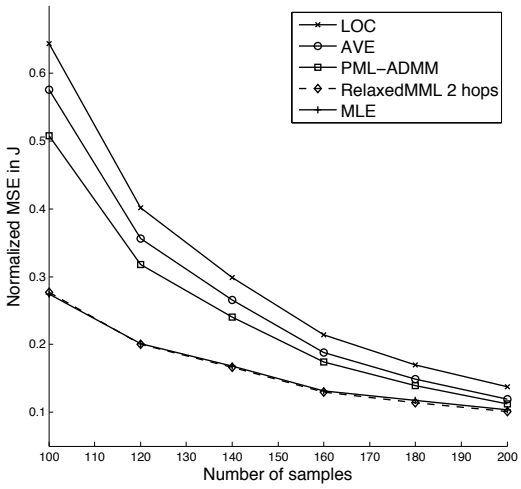
(a) Normalized MSE for K-NN graphs ($p = 500, K = 4$)



(b) Normalized MSE for lattice graphs ($p = 20 \times 20 = 400$, $\mu = 0.5$, $\sigma^2 = 0.2$)



(c) Normalized MSE for small-world graphs ($p = 100$, $K = 20$, $\beta = 0.5$)



(d) Normalized MSE for IntelLab Sensor Network Data set Guestrin et al. (2004) ($p = 50$)

Figure 2.3: Normalized MSE in the concentration matrix estimates for different graphical models. The legend in Figure 2.3(d) applies to all plots. The proposed 2-hop relaxed maximum marginal likelihood (RMML) estimator clearly improves upon existing distributed estimators and nearly closes the gap to the centralized maximum likelihood estimator.

and invert it to obtain a sample concentration matrix. This concentration matrix is then thresholded to yield a ground truth graphical model with a sparsity level of 70% zeros. Using knowledge of the sparsity and sampling from the original 1800 samples, we estimate the concentration matrix using the same estimators as before. As shown in Figure 2.3(d), the proposed two-hop relaxed MML estimator still gives a very tight approximation to the centralized GML estimator and its advantage over other distributed estimators is obvious.

We investigate the robustness of the centralized and distributed estimators in the presence of model mismatch. The nominal precision matrix $\mathbf{J}$ corresponds to a four-nearest-neighbor graphical model with $p = 200$ as in the previous experiments. We add ± 0.1 random perturbations to the non-edge components of the nominal precision matrix (also with minimal diagonal loading to ensure positive definiteness), then generate samples from the perturbed model. The different estimation algorithms are applied assuming the nominal graph structure and the resulting MSEs are plotted with respect to the nominal model. The MSEs of all estimators (using samples from both the original and perturbed models, respectively) are reported in Figure 2.4. All errors are averaged across 50 randomized experiments. As can be seen, the model mismatch leads to estimation bias for both centralized and distributed estimators. The magnitudes of the model mismatch bias for all estimators are comparable, as predicted by the theoretical analysis in Sec. 2.4.3. These experiment results confirm the robustness of the proposed distributed algorithm.

We next turn to computational comparisons. In the following experiments, we illustrate the computational gain of our distributed estimator over the centralized one through two runtime comparisons performed in Matlab. Our main focus is on the relative scaling of the runtime with respect to the number of nodes p for different estimators. We consider two algorithms for solving both the centralized GML problem and the local RMML problems. The first is an interior point algorithm implemented

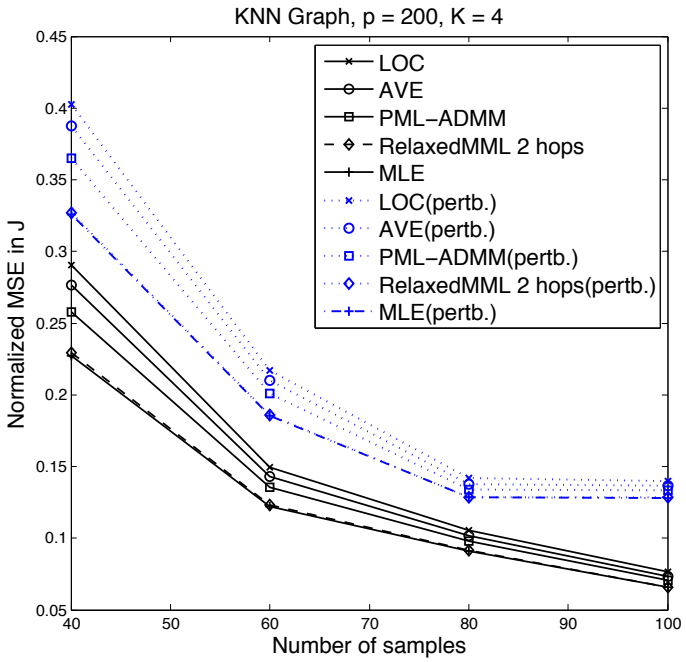


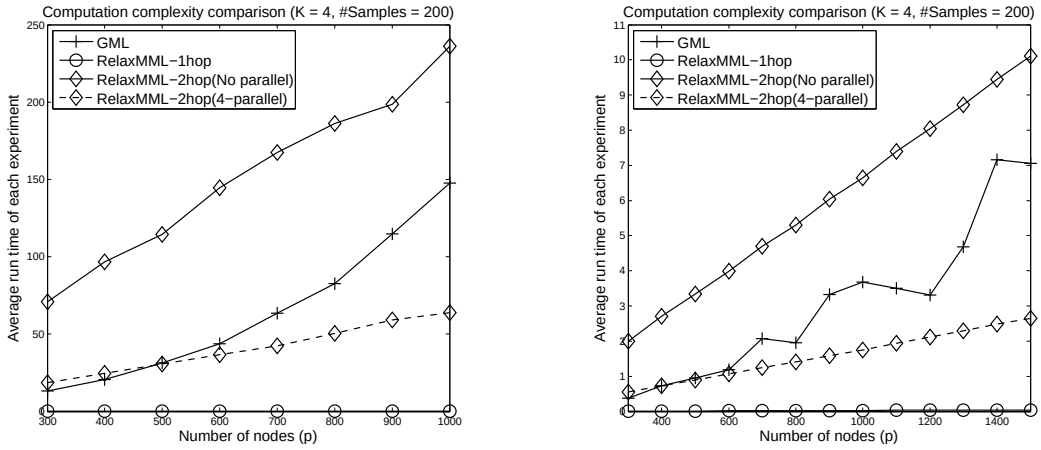
Figure 2.4: Robustness of estimators under model mismatch. All errors are obtained from K-NN ($p = 200, K = 4$) graphs and averaged over 50 experiments. For the perturbed models, ± 0.1 is added to the non-edge components of the nominal precision matrix. The proposed distributed RMML estimator is as robust as the GML estimator.

in the solver `logdetPPA` (Wang et al., 2010), which is specially designed for solving log det-SDPs. The second algorithm is the iterative regression approach in Friedman et al. (2009) for solving the covariance selection problem (Friedman et al., 2008) with known structure. In both experiments, the graphical model is a four-NN graph with similar parameter settings as before. We compare the total runtime of the GML estimator and that of different versions of RMML estimators. For the RMML estimators, we implement a sequential and a parallel version using the `parfor` function in Matlab. The results are reported in Figure 2.5. As expected, the runtime of the GML estimator is at least linear in p and the generic solver appears to be much more expensive than the iterative regression algorithm for this particular task. The total cost of the RMML estimator without parallelization is also linear in p, and is slightly higher than the GML estimator. However, when four-core parallelization is used, the run time is approximately reduced by a factor of four, resulting in lower computational complexity after $p > 500$.

It is expected that with a higher degree of parallelization, the run time of the proposed RMML estimator will continue to decrease almost linearly with the number of cores. As discussed in Section 2.5, all local RMML problems can be solved in parallel without the need for any iterative message-passing. Therefore the communication overhead is minimal, consisting of the final concatenation and symmetrization steps (2.16) and (2.18).

2.7 Summary of Chapter 2

We have proposed a distributed MML framework for estimating the concentration matrix in Gaussian graphical models. The proposed method solves a convex relaxation of a marginal likelihood maximization problem independently in each local neighborhood. A global estimate is then obtained by combining the local estimates



(a) Run time comparison using `logdetPPA` solver

(b) Run time comparison using iterative regression algorithm

Figure 2.5: Run time comparisons for GML and RMML estimators. In panel (a) the `logdetPPA` solver is used, and in panel (b) the iterative regression algorithm is used. In both figures, solid lines denote the runtime scaling of the sequential version of the algorithm, while the dashed lines denote runtime scaling for a parallelized version with four cores. In both figures, the runtime of the GML estimator is super-linear in p, while the RMML estimator exhibits linear scaling in p, and the runtime is further reduced by a factor approximately equal to the number of cores used. All experiments are implemented in a Matlab environment.

via a single round of local averaging. The proposed estimator is shown to be statistically consistent and computationally efficient. In particular, we have shown that the statistical convergence rate of our estimator is comparable to that of the more expensive centralized maximum likelihood estimator. Likewise in numerical experiments, a two-hop version of the distributed estimator is seen to be sufficient to attain centralized performance. Its improved performance relative to existing distributed estimators is also illustrated.

2.8 Proofs for Chapter 2

2.8.1 Proof of Theorem 2.1

Proof. Consider the following set of sparse positive semidefinite matrices with respect to a non-zero element set R:

$$\mathcal{K}^R := \{\mathbf{K} \mid \mathbf{K} \succeq 0, \mathbf{K}_{(j,k)} = 0, \forall (j,k) \notin R\}.$$

We first note that, when R is taken to be the relaxed edge set of a neighborhood as defined in (2.13), then the true marginal concentration matrix corresponding to the neighborhood, $\mathbf{K}^* = (\mathbf{\Sigma}^*_{\mathcal{N},\mathcal{N}})^{-1}$, must belong to the set $\mathcal{K}^R$. This can be seen from the fact that the true global concentration matrix $\mathbf{J}^*$ conforms to the sparsity pattern specified by $\widetilde{E}$ and from relations (2.10) and (2.11). Therefore the proposed relaxed MML problem (2.14) is equivalent to a standard ML problem with respect to a GGM distribution parameterized by matrix $\mathbf{K} \in \mathcal{K}^R$, with $\mathbf{K}^*$ being the population parameter. Then the asymptotic consistency, normality and efficiency of the proposed relaxed MML estimator (with respect to the local problem) all follow from the standard asymptotic analysis of the ML estimator Van der Vaart (2000). In particular, the variances of the errors achieve the diagonal elements of the inverse Fisher

information matrix $\mathbf{F}$ defined in Eq. (2.20) (see Johnson (2006) for the derivation). Finally by extracting and summing the variances corresponding to the row parameters, we obtain the expression for the asymptotic mean squared Frobenius error of the proposed global estimator $\widehat{\mathbf{J}}^{\text{Relax}}$. □

2.8.2 Proof of Theorem 2.2

Proof. We first consider the case of $k = 1$, i.e., we compare the asymptotic variances of the one-hop and two-hop RMML estimators. Subsequently we generalize the arguments to $k > 1$ and to the global ML estimator. Suppressing the index i for local neighborhoods, let $\mathcal{B}^j, \mathcal{N}^j$ be the sets of buffer and all nodes (i.e. variables) with respect to the j-hop neighborhood, respectively ($j = 1, 2$).

Next we define some set notation for edge parameters. Let $E^j = \widetilde{E} \cap (\mathcal{N}^j \times \mathcal{N}^j)$ denote the subset of edges in $\widetilde{E}$ with both endpoints in $\mathcal{N}^j$. Let B^j be the set of all possible edges connecting j-hop buffer nodes, i.e. $B^j := \mathcal{B}^j \times \mathcal{B}^j$. Recall from (2.15) that L denotes the set of row parameters, which is defined as $L = E^1 \backslash B^1$. Finally note that the (j-hop) relaxed edge sets defined in Eq. (2.13) are related to the above two sets as $R^j := E^j \cup B^j$, $j = 1, 2$.

We augment the two-hop neighborhood graph by adding all edges among one-hop buffer nodes and among two-hop buffer nodes that are not already in E^2 (see Figure 2.6 for an illustration). This augmented edge set is denoted as $\overline{E^2} := E^2 \cup B^1 \cup B^2$. After this augmentation, the one-hop buffer clique B^1 separates the two-hop neighborhood graph into two components and a non-overlapping decomposition follows:

$$\overline{E^2} = [\underbrace{L, B^1}_{C_1 = R^1}, \overbrace{(E^2 \backslash E^1) \cup B^2}^{C_2}], \tag{2.31}$$

where we define two subsets C_1 and C_2. The augmented two-hop neighborhood graph

is therefore decomposed by $(C_1 \backslash B^1, B^1, C_2 \backslash B^1)$ (Lauritzen, 1996, Def. 2.1).

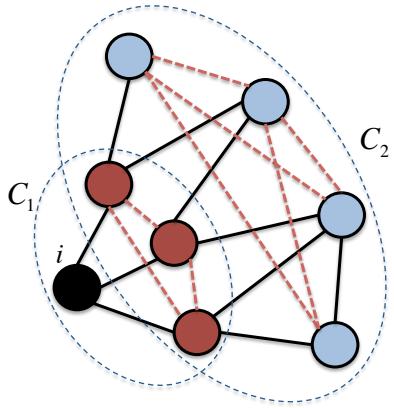


Figure 2.6: Illustration of the graph augmentation in the proof of Theorem 2.2. Dashed red lines indicate the added edges, and dashed blue contours indicate the sets C_1 and C_2, which intersect at the one-hop separator clique formed by red nodes.

Similar to Theorem 2.1, the asymptotic error covariance matrix of the RMML estimator for the augmented two-hop neighborhood is the inverse of corresponding Fisher information matrix (FIM), denoted as $\overline{\mathbf{F}}$. By Proposition 5.8 in Lauritzen (1996), the decomposability of the augmented graph leads to the following decomposition of the inverse of FIM:

$$\overline{\mathbf{F}}^{-1} = \left[(\overline{\mathbf{F}}_{C_1,C_1})^{-1}\right]^0 + \left[(\overline{\mathbf{F}}_{C_2,C_2})^{-1}\right]^0 - \left[(\overline{\mathbf{F}}_{B^1,B^1})^{-1}\right]^0,$$

where $[\cdot]^0$ appropriately zero-pads its argument to conform to the dimensions of $\overline{\mathbf{F}}^{-1}$.

Restricting this relation to the row parameters L, we have

$$\overline{\mathbf{F}}_{L,L}^{-1} = (\overline{\mathbf{F}}_{C_1,C_1})_{L,L}^{-1}, \tag{2.32}$$

since the row parameters are only contained in C_1. Noting that set C_1 is equivalent to the one-hop relaxed edge set R^1, then

$$\overline{\mathbf{F}}_{L,L}^{-1} = (\overline{\mathbf{F}}_{C_1,C_1})_{L,L}^{-1} = (\overline{\mathbf{F}}_{R^1,R^1})_{L,L}^{-1}. \tag{2.33}$$

Therefore, from Theorem 2.1 we have that the asymptotic mean squared error of the RMML estimator using the augmented graph is the same as that of the one-hop RMML estimator.

On the other hand, the augmented edge set $\overline{E^2}$ is different from the relaxed edge set R^2 only in the one-hop buffer clique B^1. Therefore another possible decomposition of the augmented edge set is (after re-ordering):

$$\overline{E^2} = [\underbrace{L, E^1 \backslash L, (E^2 \backslash E^1) \cup B^2}_{R^2}, \underbrace{B^1 \backslash E^1}_{D}], \tag{2.34}$$

where we define the difference set as D. Then using a property of Schur complements of positive semidefinite matrices, the variance matrix corresponding to R^2 (i.e. the non-zero pattern of the two-hop RMML estimator) satisfies

$$\begin{aligned} \overline{\mathbf{F}}^{-1}_{R^2,R^2} &= \left(\overline{\mathbf{F}}_{R^2,R^2} - \overline{\mathbf{F}}_{R^2,D}(\overline{\mathbf{F}}_{D,D})^{-1}\overline{\mathbf{F}}_{D,R^2}\right)^{-1} \\ &\succeq \left(\overline{\mathbf{F}}_{R^2,R^2}\right)^{-1}. \end{aligned} \tag{2.35}$$

Restricting this relation to the submatrix indexed by set L, we have

$$\overline{\mathbf{F}}^{-1}_{L,L} \succeq \left(\overline{\mathbf{F}}_{R^2,R^2}\right)^{-1}_{L,L}. \tag{2.36}$$

Now combining Eq. (2.33), Eq. (2.36) and Theorem 2.1, we can conclude that the asymptotic variance of the one-hop RMML estimator (i.e. the mean squared error) is larger than that of the two-hop estimator.

Similar arguments can be established for comparing the asymptotic variances of the two-hop RMML and the GML estimators, which shows that the asymptotic variance of RMML estimator is larger than that of the GML estimator. The above proof can be easily generalized to arbitrary k-hop neighborhoods.

□

2.8.3 Proof of Theorem 2.3

The key ingredient in proving Theorem 2.3 is the following lemma, which provides a bound for the error of the RMML estimator $\widehat{\mathbf{K}}^{\text{Relax}}$ (2.14) in a given local neighborhood (the neighborhood index i is suppressed). Let $\mathbf{\Sigma}^*$ be the true global covariance matrix, and $\mathbf{K}^*$ be the true marginal precision matrix corresponding to the given neighborhood.

Lemma 2.1. *For a given local neighborhood $\mathcal{N}$, if*

$$\left\| \left(\widehat{\mathbf{\Sigma}}^{(T)} - \mathbf{\Sigma}^* \right)_R \right\|_\infty \leq \epsilon_\Sigma \leq \frac{1}{9\overline{\kappa}\sqrt{|R|}}, \tag{2.37}$$

we have

$$\left\| \widehat{\mathbf{K}}^{Relax} - \mathbf{K}^* \right\|_F \leq 9\overline{\kappa}^2 \epsilon_\Sigma \sqrt{|R|}. \tag{2.38}$$

The proof of Lemma 2.1 is given in Section 2.8.4. The above lemma is deterministic in nature. To ensure that assumption (2.37) is satisfied with high probability when the sample covariance $\widehat{\mathbf{\Sigma}}^{(T)}$ is random, we make use of the following concentration result for Gaussian random variables by Ravikumar et al. Ravikumar et al. (2011):

Lemma 2.2. *For a p-dimensional Gaussian random vector with covariance matrix $\mathbf{\Sigma}^*$, the sample covariance matrix obtained from T samples $\widehat{\mathbf{\Sigma}}^{(T)}$ satisfies*

$$P\left\{ |\widehat{\mathbf{\Sigma}}^{(T)}_{i,j} - \mathbf{\Sigma}^*_{i,j}| > \epsilon \right\} \leq 4 \exp\left(-\frac{T \cdot \epsilon^2}{3200\overline{\sigma}^2} \right), \tag{2.39}$$

for all $\epsilon \in (0, 40\overline{\sigma})$, where $\overline{\sigma} := \max_{i=1,\ldots,p} \mathbf{\Sigma}^*_{i,i}$.

Now we are ready to prove Theorem 2.3.

Proof. (Theorem 2.3) Given the condition (2.26) on T, we have

$$C\sqrt{\frac{3200\overline{\sigma}^2 \log p^2}{T}} \leq 40\overline{\sigma}. \tag{2.40}$$

Then applying Lemma 2.2 and the union bound, we have

$$\begin{aligned} P\left\{ \left\| \left(\widehat{\mathbf{\Sigma}}^{(T)} - \mathbf{\Sigma}^* \right)_{R_i} \right\|_\infty \leq C\sqrt{\frac{3200\overline{\sigma}^2 \log p^2}{T}} \right\} \\ \geq P\left\{ \left\| \widehat{\mathbf{\Sigma}}^{(T)} - \mathbf{\Sigma}^* \right\|_\infty \leq C\sqrt{\frac{3200\overline{\sigma}^2 \log p^2}{T}} \right\} \\ \geq 1 - \frac{4}{p^{2(C^2-1)}}. \end{aligned} \tag{2.41}$$

Conditioned on the event in (2.41), condition (2.26) also guarantees that (2.37) holds for all local neighborhoods. Then the total Frobenius error in the global estimate $\widehat{\mathbf{J}}^{\text{Relax}}$ can be bounded by Lemma 2.1:

$$\begin{aligned} \|\widehat{\mathbf{J}}^{\text{Relax}} - \mathbf{J}^*\|_F &\overset{(i)}{=} \left(\sum_{i=1}^{p} \|(\widehat{\mathbf{J}}^{\text{Relax}} - \mathbf{J}^*)_{L_i}\|_F^2 \right)^{1/2} \\ &\overset{(ii)}{=} \left(\sum_{i=1}^{p} \|(\widehat{\mathbf{K}}^{\text{Relax}} - \mathbf{K}^*)_{L_i}\|_F^2 \right)^{1/2} \\ &\overset{\text{(Lem. 2.1)}}{\leq} \left(\sum_{i=1}^{p} \left(9\overline{\kappa}^2 C\sqrt{\frac{3200\overline{\sigma}^2 |R_i| \log p^2}{T}} \right)^2 \right)^{1/2} \\ &\leq 720C \cdot \overline{\kappa}^2 \overline{\sigma} \sqrt{\frac{r \log p}{T}}, \end{aligned}$$

where identity (i) is due to the fact that the global estimator is a concatenation of non-overlapping row parameter sets (see Eq. (2.15) for definition of L_i's), equality (ii) is due to our construction of $\widehat{\mathbf{J}}^{\text{Relax}}$ from $\widehat{\mathbf{K}}^{\text{Relax}}$ (see Eq. (2.16)), and the fact that row parameters are always protected. □

2.8.4 Proof of Lemma 2.1

Proof. The main idea of this proof is inspired by Rothman et al. (2008). The difference is that we focus on the local RMML problem, rather than the global ML problem (which is studied in Rothman et al. (2008)). Define the marginal likelihood function for a local neighborhood $\mathcal{N}$ as $\mathcal{L}(\mathbf{K}) = \langle \widehat{\mathbf{\Sigma}}^{(T)}_{\mathcal{N},\mathcal{N}}, \mathbf{K} \rangle - \log\det(\mathbf{K})$, where we super-script the sample covariance to emphasize that it is obtained from T samples.

Recall $\mathbf{K}^* := \left(\mathbf{\Sigma}^*_{\mathcal{N},\mathcal{N}}\right)^{-1}$ is the local marginal precision matrix. Define the shorthand notation for the local RMML estimate as $\widehat{\mathbf{K}} := \widehat{\mathbf{K}}^{\text{Relax}}$.

Consider the function $\mathcal{Q}(\Delta) := \mathcal{L}(\mathbf{K}^* + \Delta) - \mathcal{L}(\mathbf{K}^*)$, where Δ respects the sparsity structure of the RMML problem, i.e. $\Delta_{R^C} = \mathbf{0}$ and $\Delta = \Delta^T$. Let $0 < \delta \leq \overline{\kappa}$ be a given radius, define the following set

$$\mathcal{C}(\delta) := \{\Delta \mid \Delta_{R^C} = \mathbf{0}, \Delta = \Delta^T, \|\Delta\|_F = \delta\}, \tag{2.42}$$

where R is the local relaxed edge set. Note that $\mathcal{C}(\delta)$ defines a sphere, not a ball.

Note that $\mathcal{Q}(\Delta)$ is a convex function of Δ. By construction we have $\mathcal{Q}(\mathbf{0}) = 0$, and the optimality of $\widehat{\mathbf{K}}^{\text{Relax}}$ implies that $\mathcal{Q}(\widehat{\Delta}) \leq \mathcal{Q}(\mathbf{0}) = 0$, where we define $\widehat{\Delta} := \widehat{\mathbf{K}} - \mathbf{K}^*$. Then if we can establish that

$$\inf_{\Delta \in \mathcal{C}(\delta)} \mathcal{Q}(\Delta) > 0,$$

then the optimal error matrix $\widehat{\Delta}$ must lie inside the sphere defined by $\mathcal{C}(\delta)$ by convexity of $\mathcal{Q}$, implying that $\|\widehat{\Delta}\|_F \leq \delta$. Now it suffices to find a suitable radius $\delta > 0$ such that $\mathcal{Q}(\Delta)$ is lower-bounded from zero for all $\Delta \in \mathcal{C}(\delta)$.

Since

$$\begin{aligned}
\mathcal{Q}(\Delta) &= \mathcal{L}(\mathbf{K}^* + \Delta) - \mathcal{L}(\mathbf{K}^*) \\
&= \langle \widehat{\mathbf{\Sigma}}^{(T)}_{\mathcal{N},\mathcal{N}}, \Delta \rangle - (\log\det(\mathbf{K}^* + \Delta) - \log\det(\mathbf{K}^*)).
\end{aligned}$$

Similar to Rothman et al. (2008), we make use of the Taylor's theorem for the $\log\det(\cdot)$ function

$$\log\det(\mathbf{K}^* + \Delta) - \log\det(\mathbf{K}^*) = \langle(\mathbf{K}^*)^{-1}, \Delta\rangle - \overrightarrow{\Delta}^T \left[\int_0^1 (1-t)(\mathbf{K}^* + t\Delta)^{-1} \otimes (\mathbf{K}^* + t\Delta)^{-1} dt\right] \overrightarrow{\Delta}, \tag{2.43}$$

where $\otimes$ denotes the Kronecker product, and $\overrightarrow{\Delta}$ is the properly vectorized form of matrix Δ.

Using this identity, we have

$$\mathcal{Q}(\Delta) = \underbrace{\langle \widehat{\mathbf{\Sigma}}^{(T)}_{\mathcal{N},\mathcal{N}} - (\mathbf{K}^*)^{-1}, \Delta\rangle}_{T_1} + \underbrace{\overrightarrow{\Delta}^T \left[\int_0^1 (1-t)(\mathbf{K}^* + t\Delta)^{-1} \otimes (\mathbf{K}^* + t\Delta)^{-1} dt\right] \overrightarrow{\Delta}}_{T_2}. \tag{2.44}$$

Next we bound T_1 and T_2 defined above separately.

For T_1, notice that the difference matrix Δ is non-zero only in a restricted set R, therefore it reduces to a lower-dimensional inner product:

$$|T_1| = |\langle(\widehat{\mathbf{\Sigma}}^{(T)} - \mathbf{\Sigma}^*)_R, \Delta_R\rangle| \overset{(i)}{\leq} \|(\widehat{\mathbf{\Sigma}}^{(T)} - \mathbf{\Sigma}^*)_R\|_\infty \cdot \|\Delta_R\|_1 \overset{\text{Eq.(2.37)}}{\leq} \epsilon_\Sigma \cdot \sqrt{|R|} \cdot \|\Delta\|_F, \tag{2.45}$$

where (i) is due to the duality between norms $\|\cdot\|_\infty$ and $\|\cdot\|_1$.

For T_2, we follow similar derivations as in Rothman et al. (2008):

$$\begin{aligned}
T_2 &\geq \|\Delta\|_F^2 \cdot \lambda_{\min}\left(\int_0^1 (1-t)(\mathbf{K}^* + t\Delta)^{-1} \otimes (\mathbf{K}^* + t\Delta)^{-1} dt\right) \\
&\overset{(i)}{\geq} \|\Delta\|_F^2 \int_0^1 (1-t)\lambda_{\min}^2\left((\mathbf{K}^* + t\Delta)^{-1}\right) dt \\
&\geq \frac{1}{2}\|\Delta\|_F^2 \min_{0\leq t\leq 1} \lambda_{\min}^2\left((\mathbf{K}^* + t\Delta)^{-1}\right) \\
&\overset{(ii)}{\geq} \frac{1}{2}\|\Delta\|_F^2 \min_{\|\widetilde{\Delta}\|_F\leq\delta} \lambda_{\min}^2\left((\mathbf{K}^* + \widetilde{\Delta})^{-1}\right) \\
&\geq \frac{1}{2}\|\Delta\|_F^2 \min_{\|\widetilde{\Delta}\|_F\leq\delta} \|\mathbf{K}^* + \widetilde{\Delta}\|_2^{-2} \\
&\geq \frac{1}{2}\|\Delta\|_F^2 \min_{\|\widetilde{\Delta}\|_F\leq\delta} (\|\mathbf{K}^*\|_2 + \|\widetilde{\Delta}\|_2)^{-2} \\
&\overset{(iii)}{\geq} \frac{1}{2}\|\Delta\|_F^2 \min_{\|\widetilde{\Delta}\|_F\leq\delta} (\overline{\kappa} + \|\widetilde{\Delta}\|_F)^{-2} = \frac{1}{8\overline{\kappa}^2}\|\Delta\|_F^2,
\end{aligned} \tag{2.46}$$

where (i) follows the eigenvalue property of Kronecker product, (ii) is due to the fact that $\Delta \in \mathcal{C}(\delta)$, (iii) is due to the interlacing property of eigenvalues of sub-matrices

$$\|\mathbf{K}^*\|_2 = \frac{1}{\lambda_{\min}(\mathbf{\Sigma}^*_{\mathcal{N},\mathcal{N}})} \leq \frac{1}{\lambda_{\min}(\mathbf{\Sigma}^*)} = \|\mathbf{J}^*\|_2 = \overline{\kappa}, \tag{2.47}$$

The last inequality is due to construction, i.e. $\delta \leq \overline{\kappa}$.

Now $\mathcal{Q}(\Delta)$ can be bounded by

$$\mathcal{Q}(\Delta) \geq -\epsilon_\Sigma \cdot \sqrt{|R|} \cdot \|\Delta\|_F + \frac{1}{8\overline{\kappa}^2}\|\Delta\|_F^2 \tag{2.48}$$

$$= \|\Delta\|_F \left(\frac{1}{8\overline{\kappa}^2}\|\Delta\|_F - \epsilon_\Sigma \cdot \sqrt{|R|}\right). \tag{2.49}$$

The proof is complete if the RHS can be lower bounded away from zero. It can be verified that with the choice of ϵ_Σ as in (2.37), letting $\delta = 9\overline{\kappa}^2\epsilon_\Sigma\sqrt{|R|}$ suffices. Therefore $\|\widehat{\Delta}\|_F \leq \delta = 9\overline{\kappa}^2\epsilon_\Sigma\sqrt{|R|}$.

□

CHAPTER 3

Distributed Principal Component Analysis in Directed Graphical Models

In this chapter, we continue to develop the distributed estimation framework for graphical models, but focus on directed Gaussian graphical models. We introduce an efficient algorithm for performing distributed principal component analysis (PCA) in directed Gaussian graphical models. By exploiting structured sparsity in the Cholesky factor of the inverse covariance (concentration) matrix, our proposed algorithm accomplishes global principal subspace estimation through local computation and message passing. We illustrate the computation/communication advantages of the proposed algorithm through a simulated online estimation example and an anomaly detection example in a real-world network.

3.1 Introduction

We consider distributed principal component analysis (PCA) in directed Gaussian graphical models. As a widely used dimensionality reduction technique, PCA estimates the principal subspace from data samples. The principal subspace is the low-dimensional subspace spanned by the leading eigenvectors of the data covariance matrix. The projected data onto this subspace preserves most of the covariation of the data, thus the principal components are often used as compressed representation

of the original data (Anderson, 1958). The difficulties in implementation of PCA on large-scale and networked data include the heavy computation burden of eigenvalue decomposition (EVD) and the massive communication cost required for processing data at a fusion center. Aiming at overcoming these bottlenecks, distributed and decentralized implementations of PCA have been proposed for applications such as distributed estimation and tracking in wireless sensor network (Gastpar et al., 2006; Li et al., 2011), computer vision (Tron & Vidal, 2011) and anomaly detection in computer networks (Huang et al., 2006b; Wiesel et al., 2010).

In this work we consider the case that the variables are governed by a graphical model. Such models represent conditional dependencies between variables by edges in a graph (Lauritzen, 1996). When the graph is sparse and the variables are jointly Gaussian, the graphical model imposes sparsity on the inverse covariance, variously called the information, concentration or precision matrix. Such a representation enables distributed and efficient inference algorithms, such as the well-known junction tree algorithm (Jordan & Bishop, 2001) and belief propagation (BP) (Pearl, 1988). When the topology (local dependency) of the graphical model matches the topology (local data passage) of internode communication, superior performance can be achieved at significantly reduced computational cost as compared to the global centralized approach. In applications such as those mentioned above, often there is a good match between local dependency and local data passage, e.g., in geographically distributed networks of sensing devices. The premise of this chaper is that the model topology and communication topology are matched. Such an assumption is common in other decentralized formulations of networked estimation, e.g, BP via message passing in imaging and networks. Even in the case of approximate modeling, it is an approach to tradeoff precision for complexity through leveraging the computationally efficient inference algorithm.

Since PCA can be interpreted as maximum likelihood (ML) estimation of the

covariance followed by its EVD, the first goal of distributed PCA is therefore to perform distributed covariance estimation. However, it is well-known that the distributed maximum likelihood covariance estimation problem has no closed form solution (Wiesel et al., 2010). In the special case of chordal graphs (also known as decomposable or triangulated graphs), a closed-form distributed covariance estimator exists, and a distributed PCA algorithm called DPCA can be implemented for exploiting sparsity in concentration domain (Wiesel & Hero, 2009).

This chapter extends the DPCA framework to *directed* graphical models, and we call this extension Distributed Directed PCA (DDPCA). Instead of assuming sparsity in the concentration matrix, it assumes sparsity in Cholesky factor of the concentration matrix. This assumption leads to a more parsimonious representation when the Cholesky factor specifies a generative model for the observations in terms of a white noise process (Smith & Kohn, 2002). This proposed approach results in closed-form distributed covariance estimation and reduced inference complexity in terms of computation/communication. The DDPCA algorithm can equally be applied to non-directed decomposable graphical models by using a sparsity-preserving Markov-equivalent conversion. More specifically, DDPCA first performs distributed regressions to estimate a rank reduced covariance matrix. Then a distributed iterative EVD algorithm is implemented based on an efficient matrix-vector multiplier that fully exploits sparsity structure. We also show that this matrix-vector multiplier coincides with a specific Cholesky-parametrized sum-product algorithm on Gaussian DAG models, which is provably convergent and exact for inference of marginal means and also for solving DAG-sparse linear systems.

The outline of the chapter is as follows. In Section 3.2 we briefly review directed Gaussian graphical models and state our problem. A two-step DDPCA algorithm is introduced in Section 3.3. In Section 3.4, we interpret the proposed matrix-vector multiplier as a specific Cholesky-parametrized sum-product algorithm and then compare

DDPCA with DPCA algorithm in Section 3.5. We illustrate the improved properties of DDPCA in an online subspace estimation problem in Section 3.6, and apply it to the task of distributed anomaly detection in a real-world computer network in Section 3.7. We conclude the chapter with a summary in Section 3.8.

The following notation is used in Chapter 3. Boldface upper case letters denote matrices, boldface lower case letters denote column vectors, and standard lower case letters denote scalars. The superscripts $(\cdot)^T$ and $(\cdot)^{-1}$ denote the transpose and matrix inverse, respectively; and $(\cdot)^{-T}$ means the transpose of inverse. The cardinality of a set a is denoted by $|a|$, the difference of two sets is denoted as $a \backslash b$. We use indices in the subscript $\mathbf{x}_a$ or $\mathbf{X}_{a,b}$ to denote sub-vectors or sub-matrices, respectively.

3.2 Problem Formulation

Graphical models are compact representations of the conditional independencies between subsets of the random variables of multivariate distributions. Two common classes of graphical models are directed graphical models (also known as Bayesian networks), and undirected graphical models (also known as Markov Random Field). Our focus is on directed models, and first we give a brief review. We refer the readers to Lauritzen (1996) for more detailed introduction to graphical models.

A directed graphical model is denoted by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes (or vertices) and all edges in the edge set $\mathcal{E}$ are directed. Two vertices i and j are adjacent if there is an edge between them. If $i \to j$, then i is called a parent of j and j is called a child of i. The set of parents of i in $\mathcal{G}$ is denoted as $pa(i)$, and $ch(i)$ denotes the set of its children. We assume no directed cycle exist in the graph, therefore the model is also called *directed acyclic graph* (DAG) model. However, loops or cycles in the underlying undirected skeleton are allowed.

A p-dimensional random vector $\mathbf{x}$ satisfies the Markov property with respect to a

DAG model $\mathcal{G}$, if the conditional independencies between the variable $\mathbf{x}_i$'s are encoded in $\mathcal{G}$ through the notion of *d-separation* Lauritzen (1996). Applying the Markov property on the DAG results in the following factorization of joint distribution of $\mathbf{x}$,

$$f(\mathbf{x}) = \prod_{i \in \mathcal{V}} f(x_i \,|\, \mathbf{x}_{pa(i)}). \tag{3.1}$$

The DAG corresponds to a certain partial ordering of the variables, therefore is suitable for modeling spatial or temporal dependent random variables. Each variable only depends on a subset of its previous variables, denoted as its parents nodes in the DAG model, i.e.

$$x_i = f(x_1, \ldots, x_{i-1}) + \epsilon_i = f(\mathbf{x}_{pa(i)}) + \epsilon_i, \tag{3.2}$$

where ϵ_i's are uncorrelated residuals.

Assuming the random variables are jointly Gaussian distributed with zero mean, the dependency reduces to the following recursive linear system Wermuth (1980):

$$x_i = \sum_{k \in pa(i)} \lambda_{ik} x_k + \epsilon_i, \quad i = 1, ..., p. \tag{3.3}$$

Define a lower triangular matrix $\mathbf{\Lambda}$ with zero diagonal and

$$\mathbf{\Lambda}_{ik} = \begin{cases} 0 & \text{if } k \notin pa(i) \\ \lambda_{ik} & \text{if } k \in pa(i), \end{cases} \tag{3.4}$$

then the above linear system (3.3) can be written as

$$\mathbf{x} = \mathbf{\Lambda}\mathbf{x} + \epsilon. \tag{3.5}$$

This linear equation leads to the following well-known diagonalization of the co-

variance matrix $\mathbf{\Sigma} = \mathbb{E}(\mathbf{x}\mathbf{x}^T)$ Smith & Kohn (2002); Rütimann & Bühlmann (2009):

$$(\mathbf{I} - \mathbf{\Lambda})\mathbf{\Sigma}(\mathbf{I} - \mathbf{\Lambda})^T = \mathbf{\Omega}, \tag{3.6}$$

where $\mathbf{\Omega}$ is a diagonal matrix denoting the variances of residual ϵ_j's.

It is convenient to define the *modified* Cholesky factor (also called the Backward Cholesky factor) of the concentration matrix $\mathbf{J} = \mathbf{\Sigma}^{-1}$, denoted as $\mathbf{L}$ and defined as:

$$\mathbf{L}^T\mathbf{L} = \mathbf{J} = \mathbf{\Sigma}^{-1}. \tag{3.7}$$

From (3.6), we have

$$\mathbf{L} = \mathbf{\Omega}^{-1/2}(\mathbf{I} - \mathbf{\Lambda}). \tag{3.8}$$

It is easy to see that $\mathbf{L}$ shares the same lower-diagonal nonzero pattern as $\mathbf{\Lambda}$, which characterizes the topology of the DAG $\mathcal{G}$ through (3.4). This fact indicates that, for directed Gaussian graphical models, the conditional independencies between variables translate to zeros in the Cholesky factor of the concentration matrix $\mathbf{J}$.

Now we can formulate our problem. The input to our algorithm is a set of T realizations of a random vector $\mathbf{x}$ modeled by a known Gaussian DAG model $\mathcal{G}$: $\{\mathbf{x}[t]\}_{t=1}^T$. We assume that there are p units, or agents, that collect and process the data in a distributed manner. Each agent i only collects and processes all the T samples of the i-th component (or sub-vector) $\{\mathbf{x}_i[t]\}_{t=1}^T$. Agents can perform computation (called local computation) and communicate with their neighbors (called local communication) defined by $\mathcal{G}$. Using local data samples and local computation/communication, our goal is to perform global estimation of the principal subspace spanned by the first r leading eigenvectors of the covariance matrix. In other words, our algorithm searches for the the linear combination $X = \mathbf{V}^T\mathbf{x}$ having maximal variance, where $\mathbf{V} \in \mathbb{R}^{p\times r}$.

We emphasize that, although we sometimes user scalar notation x_i for simplicity, the proposed framework and algorithm can be easily generalized to multivariate case, where each node of the graphical model (i.e. each agent) corresponds to a of random sub-vector $\mathbf{x}_i$.

3.3 Algorithm

As discussed above, the classical centralized PCA algorithms are not scalable for general high-dimensional problem or networked data, due to the intensive computation, communication and storage costs of covariance estimation and EVD algorithms applied to a potentially dense covariance matrix (or its inverse). The proposed Directed Distributed PCA (DDPCA) algorithm enables decentralized computation by exploiting sparsity in the Cholesky factor of the concentration matrix, and the complexities and costs only scale with respect to local dimensions.

DDPCA is a two-step algorithm. In the first stage, the Cholesky factor of the concentration matrix is estimated through local regressions. Then in the second stage the leading eigenvectors are estimated through an iterative EVD algorithm based on a distributed matrix-vector multiplier that exploits the estimated Cholesky factor.

3.3.1 Step I: Distributed Covariance Estimation

As mentioned before, given a random vector $\mathbf{x}$ that satisfies a DAG model, the conditional independencies among its components translate to zeros in the Cholesky factor $\mathbf{L}$ of the concentration matrix $\mathbf{J}$. Since the inverse covariance matrix is fully characterized by $\mathbf{L}$ (or equivalently, $\mathbf{\Lambda}$ and $\mathbf{\Omega}$), a simple approach for inverse covariance estimation on such DAGs is to perform distributed regression (3.3) of each node on its parents to estimate the parameters in $\mathbf{L}$. The procedure is illustrated in Algorithm

1, where we define the local covariance matrices as:

$$\mathbf{S}_{pa(i),pa(i)} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{pa(i)}[t]\mathbf{x}_{pa(i)}^T[t], \tag{3.9}$$

$$\mathbf{S}_{pa(i),i} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{pa(i)}[t]\mathbf{x}_{i}^T[t]. \tag{3.10}$$

Algorithm 1: Distributed covariance estimation on DAGs

Input: Gaussian distributed samples $\{\mathbf{x}[t]\}_{t=1}^T$, DAG $\mathcal{G}$
Output: Estimated Cholesky factor $\hat{\mathbf{L}}$
for *all agent* $i \in \{1, 2, \ldots, p\}$ **do**
  Collect $\{\mathbf{x}_{pa(i)}[t]\}_{t=1}^T$ from parents
  Calculate local sample covariance (3.9) and (3.10)
  $\hat{\mathbf{\Lambda}}_{i,pa(i)}^T = \mathbf{S}_{pa(i),pa(i)}^{-1}\mathbf{S}_{pa(i),i}$
  $\hat{\epsilon}_i[t] = x_i[t] - \hat{\mathbf{\Lambda}}_{i,pa(i)}\mathbf{x}_{pa(i)}[t]$
  $\hat{\mathbf{\Omega}}_{ii} = \frac{1}{T}\sum_{t=1}^{T} \hat{\epsilon}_i[t]\hat{\epsilon}_i^T[t]$
  $\hat{\mathbf{L}}_{ij} = -\hat{\mathbf{\Omega}}_{ii}^{-1/2} \cdot \hat{\mathbf{\Lambda}}_{ij}$, for $j \in pa(i)$
  $\hat{\mathbf{L}}_{ii} = \hat{\mathbf{\Omega}}_{ii}^{-1/2}$

In fact, this procedure results in the maximum likelihood estimation Wermuth (1980). Upon completion, each agent has its corresponding local component of the matrix $\hat{\mathbf{L}}$, i.e. the i_{th} row, $[\hat{\mathbf{L}}]_{i,:}$.

This distributed regression-based covariance estimator is feasible for temporally recursive estimation, e.g., by using recursive least squares (RLS), which can enable online estimation or fast distributed principal subspace tracking in time-varying scenarios. The covariance estimation procedures can also be easily extended to the scenario of unknown graph structure by adopting penalized regression for regularization and variable selection Shojaie & Michailidis (2010); Huang et al. (2006a).

3.3.2 Step II: Distributed Principal Subspace Estimation

The second stage of DDPCA is a distributed implementation of an iterative EVD algorithm to estimate the leading eigenvectors of the covariance matrix. Among many EVD algorithms designed for symmetric matrices, in this chapter we focus on the most simple yet prevailing one, the Power Iteration. For searching the leading eigenvector of the covariance matrix $\mathbf{\Sigma}$, in the t-th iteration of Power Iteration, current estimate of the eigenvector $\mathbf{v}^{(t)}$ is multiplied by $\mathbf{\Sigma}$, the result is then normalized as the input of next iteration, i.e.

$$\mathbf{u}^{(t)} = \mathbf{\Sigma}\mathbf{v}^{(t)}, \quad \mathbf{v}^{(t+1)} = \frac{\mathbf{u}^{(t)}}{\|\mathbf{u}^t\|}. \tag{3.11}$$

Under mild assumptions (such as separation of the true leading eigenvalues), the estimate converges to the true leading eigenvector at a linear rate (Golub & Van Loan, 1996).

The most expensive cost for performing the Power Iteration is the repeated matrix-vector multiplication, especially for large-scale or network data. The proposed DDPCA decentralizes this computation through the use of an efficient distributed matrix-vector multiplier that exploiting sparsity pattern of the estimated Cholesky factor $\hat{\mathbf{L}}$. We will elaborate this multiplier in the next section.

The norm calculation and scaling operation in (3.11) could be done in a distributed manner via average consensus (Li et al., 2011) or message-passing on the DAG under serial scheduling (e.g. from root to leaves then back to root). The cost is negligible comparing with the matrix-vector multiplication.

For estimating the principal subspace, a extension of Power Iteration, namely Orthogonal Iteration could be used for searching multiple leading eigenvectors. In each iteration, matrix-vector multiplication is performed simultaneously for each estimated vector, and QR factorization is used for orthonormalization. The QR factor-

ization could be implemented in an efficient and distributed form with small additional cost (O'Leary & Whitman, 1990), when the number of eigenvectors we are interested in is much smaller than the total dimension (this is often the case in dimensionality reduction applications).

Similarly, many more advanced EVD algorithms, such as the Lanczos Iteration (Golub & Van Loan, 1996), are also based on repeated matrix-vector multiplication. Therefore they can be easily adopted in our DDPCA framework. These algorithms provide a good way to trade-off implementation complexity for better convergence property.

3.3.3 Cholesky-based Distributed Matrix-Vector Multiplier

In this section we describe the distributed matrix-vector multiplier (MVM) designed for directed graphs. Recall in the factorization in (3.7), the multiplication in (3.11) can be structured as

$$\mathbf{u} = \mathbf{L}^{-1}\mathbf{L}^{-T}\mathbf{v}, \tag{3.12}$$

which is efficiently performed by introducing an auxiliary vector $\mathbf{y}$ and sequentially solving the following two triangular systems:

$$\mathbf{L}^T\mathbf{y} = \mathbf{v}, \quad \mathbf{L}\mathbf{u} = \mathbf{y}, \tag{3.13}$$

through *backward substitution* and *forward substitution* respectively. The procedures are detailed in Algorithm 2. We emphasize that, since the non-zero pattern of $\mathbf{L}$ matches the graph structure of $\mathcal{G}$, in the substitution algorithms, solving a given component of the solution vector only requires *linear* message-passing from its parents

or children. For example, the j-th component of $\mathbf{y}$ is calculated by

$$y_j = \mathbf{L}_{jj}^{-1}(v_j - \sum_{m \in \mathrm{ch}(j)} \mathbf{L}_{mj}^T y_m). \tag{3.14}$$

The proposed MVM is highly scalable by distributing the challenging centralized computation cost to local computation and efficient message-passing on the edges.

Algorithm 2: Distributed Matrix-Vector Multiplier

Input: Estimated Cholesky factor $\hat{\mathbf{L}}$ with matching sparsity pattern to DAG $\mathcal{G}$, vector $\mathbf{v}$
Output: Vector $\mathbf{u} = \hat{\mathbf{L}}^{-1}\hat{\mathbf{L}}^{-T}\mathbf{v}$
// Solve $\hat{\mathbf{L}}^T\mathbf{y} = \mathbf{v}$ *through Backward Substitution:*
for $j = p, ..., 1$ **do**
  $y_j = \hat{\mathbf{L}}_{jj}^{-1}(v_j - \sum_{m \in \mathrm{ch}(j)} \mathbf{M}_{m \to j})$
  Bottom-up message calculation: $\mathbf{M}_{j \to k} = \hat{\mathbf{L}}_{jk}^T y_j$, for all $k \in \mathrm{pa}(j)$
// Solve $\hat{\mathbf{L}}\mathbf{u} = \mathbf{y}$ *through Forward Substitution:*
for $j = 1, \ldots, p$ **do**
  $u_j = \hat{\mathbf{L}}_{jj}^{-1}(y_j - \sum_{m \in \mathrm{pa}(j)} \hat{\mathbf{L}}_{jm}\mathbf{M}_{m \to j})$
  Top-down message calculation:
  $\mathbf{M}_{j \to k} = u_j$, for all $k \in \mathrm{ch}(j)$

With this MVM we can easily implement the above-mentioned Orthogonal Iteration in a distributed manner. The standard convergence properties also hold for this distributed version since it is an exact reformulation.

3.4 A Sum-Product Perspective

In this section, we discuss the relation between the proposed MVM (Algorithm 2) and classical inference methods in Gaussian DAGs. The traditional approach for efficient inference in graphical models is via message passing techniques, as Gaussian belief propagation (BP) (Weiss & Freeman, 2001) and the sum-product algorithm (Pearl, 1988). Recently, it was shown that such techniques can be used as distributed solutions to sparse linear systems $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A}$ is symmetric and represents a sparse

undirected graph (Shental et al., 2008). Roughly speaking, the solutions are exact for undirected tree like structures and approximate for loopy undirected graphs (Weiss & Freeman, 2001; Malioutov et al., 2006). On the other hand, the standard approach to exact inference in DAGs is via graph conversion to undirected graphical models (e.g. a possibly inefficient junction tree[1] via moralization), and then apply inference algorithms such as loopy belief propagation. In general, inference on DAGs leads to inexact solutions. In what follows, we prove that the proposed MVM (Algorithm 2) in fact coincides with a specific Cholesky-parametrized sum-product algorithm in Gaussian DAG models, which is provably convergent and exact for inferring the marginal means and solving DAG-sparse linear systems. Thus, we exploit the DAG properties without resorting to inefficient conversion and inexact inference (i.e. loopy belief propagation). However, we emphasize that our task, namely solving a single linear system, is a much easier than the general inference tasks, namely solving multiple linear systems simultaneously by marginalizing over all the nodes together.

DAGs can be conveniently represented by *factor graphs* Kschischang et al. (2001). Factor graphs are bipartite graphs that express the structure of factorization. Assuming a factorization of the joint probability density function of $\mathbf{x}$,

$$f(\mathbf{x}) = \prod_{j \in J} f_j(X_j), \tag{3.15}$$

where the j-th factor is also called a *local function*, and X_j denotes the set of variables related to f_j. A factor graph has a *variable node* for each variable x_i, a *factor node* for each local function f_j, and edges connecting factor node with all related variable nodes.

Recall (3.7), the probability density function of a Gaussian distributed random vector $\mathbf{x}$ defined on a DAG model $\mathcal{G}$ can be parametrized by the Cholesky factor $\mathbf{L}$

[1] A junction tree is basically a tree whose nodes are sets of variables which satisfy a special ordering

as

$$f(\mathbf{x}) \propto \exp\big(\sum_{i=1}^{p} h_i x_i - \frac{1}{2}\sum_{i=1}^{p}\big(\mathbf{L}_{ii}x_i + \sum_{j\in pa(i)} \mathbf{L}_{ij}x_j\big)^2\big), \tag{3.16}$$

then it is natural to define the following factorization

$$f(\mathbf{x}) = \prod_{i=1}^{p} f_i(x_i, \mathbf{x}_{pa(i)}), \tag{3.17}$$

where the i-th factor is

$$f_i(x_i, \mathbf{x}_{pa(i)}) \triangleq \exp\big(h_i x_i - \frac{1}{2}(\mathbf{L}_{i,i}x_i + \mathbf{L}_{i,pa(i)}\mathbf{x}_{pa(i)})^2\big). \tag{3.18}$$

With these defined factors, each variable node has an associated factor node connecting it and all its parent variable nodes. Unlike conversion to undirected graphs, no additional loops are introduced in the resulting factor graphs.

Sum product algorithm is an algorithm for computing marginal functions on factor graphs via message passing (between variable nodes and factor nodes) and local marginalization. More specifically, sum-product algorithm computes the following two types of messages:

Variable node to factor node:

$$m_{x\to f}(x) = \prod_{h\in n(x)\setminus f} m_{h\to x}(x). \tag{3.19}$$

Factor node to variable node:

$$m_{f\to x}(x) = \int_{\sim x}\left(f(X)\prod_{y\in n(f)\setminus x} m_{y\to f}(y)\right), \tag{3.20}$$

where notation $\int_{\sim x}$ means integral with respect to all the variables but x, as used in Kschischang et al. (2001). In Gaussian case, the calculations of (3.19) and (3.20)

reduce to computing the parameters $\mathbf{h}$ and $\mathbf{J}$ of the messages. Finally the marginal functions can be obtained by computing the product of all incoming messages at variable nodes.

Since the resulting factor graphs defined by DAGs potentially contain cycles, in general the convergence and exactness of sum-product algorithm is not guaranteed. However, an iterative version of sum-product algorithm can be performed in loopy graphs under pre-defined message scheduling and initialization. Define a serial scheduling $\mathcal{M}$ as $\{p, p-1, \ldots, 2, 1, 2, \ldots, p\}$, i.e. a bottom-up pass from leaf variable nodes to root variable node (routing via corresponding factor nodes) followed by a top-down pass back to the leaves, and initialize all the messages as constant 1 (Kschischang et al., 2001). Interestingly, we have the following equivalence result.

Theorem 3.1. *The sum-product algorithm on the factor graph defined by a Gaussian DAG model under scheduling $\mathcal{M}$ and with constant initialization is equivalent to Algorithm 2 with input vector equals to $\mathbf{h}$, which results in exact inference of the marginal means of the Gaussian DAG.*

Proof. See Section 3.9. □

In the proof, it is shown that each pass of the sum-product algorithm is equivalent to a solving a linear triangular system with the Cholesky factor $\mathbf{L}$ via distributed substitution methods. This result resembles (Shental et al., 2008, Proposition 6), where the authors show the equivalence between Gaussian belief propagation (also a specific sum-product algorithm) and solving a linear system using Jacobi method. In contrast to Shental et al. (2008), our approach is especially efficient for Gaussian DAG models and DAG-sparse systems.

3.5 Comparison between DDPCA and DPCA

The introduced DDPCA framework is designed for *directed* Gaussian graphical models. In this section, we will compare DDPCA with a previously proposed distributed PCA framework, called DPCA (Wiesel & Hero, 2009), which is designed for decomposable Gaussian graphical models (DGGM).

Unlike DAG models, the Markov property on undirected graphs states that random variables measured at two subsets are conditionally independent given the separating subset[2]. Under Gaussian assumption, Markov property in undirected models imposes zeros in the information matrix $\mathbf{J}$. More specifically, conditional independence between $\mathbf{x}_i$ and $\mathbf{x}_j$ given all the other random variables implies $\mathbf{J}_{ij} = \mathbf{0}$ (Lauritzen, 1996).

Decomposable models (also known as chordal or triangulated models) are a special type of undirected graphical model in which the conditional independencies satisfy an appealing structure. A decomposable graph can be divided into an ordered sequence of fully connected subgraphs known as cliques and denoted by $C_1, \ldots, C_K$. These ordered cliques are coupled through separators

$$S_j = (C_1 \cup C_2 \cup \cdots \cup C_{j-1}) \cap C_j \tag{3.21}$$

for $j = 2, \ldots, , K$, and satisfy the *running intersection property*: for all $j \geq 2$ there is a $k < j$ such that $S_j \subseteq C_k$. Assuming Gaussian distribution, it is readily seen from the above running intersection property that the sparsity patterns of the information matrix for DGGMs matches its decomposable structures. This sparsity pattern can be exploited for efficient inference (Wiesel & Hero, 2009; Wiesel et al., 2010; Dahl et al., 2008).

DGGMs are closely related to Gaussian DAG models. Given a DGGM, we can

[2]We say that set c *separates* sets a and b, if there is no path between any node in a and node in b that does not include a node in c.

construct a DAG model by connecting nodes in the same clique with directed acyclic edges. The resulting model preserves all the conditional independencies and the sparsity pattern in its information domain, and is called a *Markov equivalence*. The opposite conversion from DAG models to DGGMs is also straightforward through moralization, but in most cases additional edges are required which might reduce the sparsity level. The worst-case scenario is the "star" graph, where all but one nodes have an edge pointing to the last one. The corresponding Cholesky factor of the information matrix is very sparse, whereas the information matrix itself is completely dense. Therefore DAG models are more parsimonious than DGGMs.

Due to this immediate equivalence between DAG models and DGGMs, we can compare the proposed DDPCA with DPCA (Wiesel & Hero, 2009). DPCA performs local computation within overlapping cliques of a DGGM, and global estimation is achieved through message-passing within the separators. DPCA has cubic local computation complexity in the clique dimensions (which solves for the local minimum eigenvalue) and quadratic inter-clique communication cost in the separator dimensions, therefore is especially efficient for DDGMs with very small separators and relatively large cliques. However, the proposed DDPCA only requires quadratic computation cost in the total local dimension for solving a triangular linear system (Equation (3.14)), and the communication cost is linear in the clique dimension. Therefore for general graphs DDPCA enjoys reduced computation and communication costs than DPCA for Markov equivalent graphical modeling.

We emphasize that PCA algorithms on Markov equivalent Gaussian graphical models have identical estimation results, since the models have the same second order moment. But the algorithms have potentially different implementation costs. In the following sections, we only consider Markov-equivalent Gaussian graphical models for comparing different PCA algorithms.

3.6 An Illustrative Example: Online subspace estimation

We first illustrate the propose DDPCA algorithm using a synthetic example, where our task is an online estimation of the first principal component from incremental Gaussian samples. The samples are generated from a 80-dimensional multivariate normal random vector using a four-node DAG model plus Gaussian noise. Each node corresponds to a 20-dimensional random sub-vector, and three non-adjacent nodes pointing to the last one (the "star" graph). The Markov equivalent undirected graph of this sparse DAG via moralization is a fully-connected graph, which means that the concentration matrix has no sparsity. Therefore DPCA does not work for distributing the computation. However, the proposed DDPCA is a good candidate for taking advantage of this sparsity in the Cholesky factor.

Using random initialization, we perform DDPCA at each time step with 200 incremental samples to update the previous estimation. Very few number of iterations (specifically 1, 2 and 8) are performed for the updating. For comparison, we implement a centralized PCA using the full sample covariance (not our sparse covariance estimator), which is also updated with incremental samples. Again we emphasize that DPCA coincides with the centralized PCA on this example.

Fig. 3.1 shows the subspace estimation error, i.e., the subspace distance[3] between the estimated and the true one with respect to the number of iterations (which indicates the total number of samples used). It can be seen that even two iterations in each update are enough for DDPCA; whereas centralized PCA performs poorly using 20 iterations, and requires 50 iterations for similar performance to DDPCA. The better performance and faster convergence time of DDPCA is due to the matching covariance model and the better statistical accuracy of DDPCA in low sample

[3]defined as $\|P_{est} - P_{true}\|_2$, where P is the projector of the subspace.

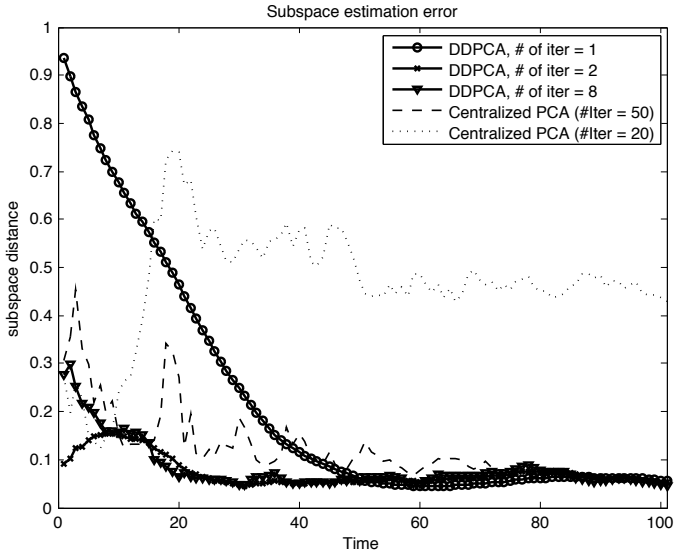


Figure 3.1: Online estimation example (DPCA coincides with centralized PCA)

number regime. The decentralized framework also makes DDPCA computationally advantageous.

3.7 Experiments on Real-world Networks

3.7.1 Network Model

In this section we introduce a widely used network model and describe two specific models that match our graphical model assumptions. Consider a computer network whose topology is a graph $\mathcal{G}_n$, consisting of N nodes and L links (adjacent nodes are connected by two links). The network carries traffic flows from origin nodes to destination nodes, known as OD (Origin-Destination) flows, through routing over a predetermined ordered subset of links (called a *path*) of the network. On each link of the network, the accumulation of all OD flows that pass through this link is measured. Therefore link flows and the OD flows can be related by a linear equation

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \tag{3.22}$$

where $\mathbf{y} \in \mathbb{R}^L$ is the link-level flow measurement vector, $\mathbf{x} \in \mathbb{R}^P$ is the OD flow vector, and P is the total number of OD paths. The *routing matrix* $\mathbf{A} = (a_{lp})_{L \times P}$ is defined as:

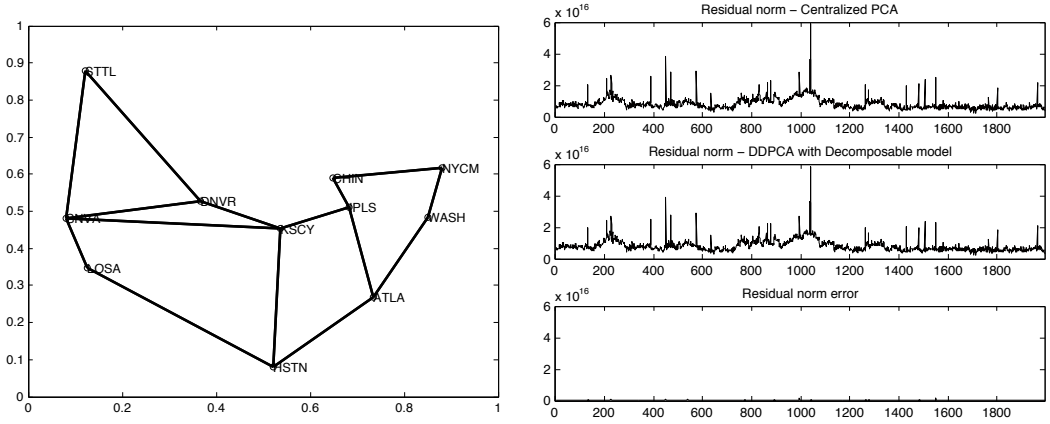
$$a_{lp} = \begin{cases} 1, & \text{if link } l \text{ is on path } p \\ 0, & \text{otherwise.} \end{cases} \tag{3.23}$$

As noted in Stoev et al. (2010), under uncongested network conditions, the OD traffic flows x_p's can be well approximated as statistically independent. Thus the statistical dependence between components of $\mathbf{y}$ are determined by the structure of routing matrix $\mathbf{A}$. We consider the following two scenarios for modeling the link flow vector $\mathbf{y}$:

Scenario A (Decomposable model). In this scenario, we consider the natural intuition that two distantly separated link flow variables are approximately conditionally independent given the separator flow variables between them. Therefore if the network topology satisfies decomposable separation as defined in (3.21), then $\mathbf{y}$ is readily modeled by a DGGM, which also can be converted into a DAG model as described in Section 3.5. This model is equivalent to the model assumed in Wiesel & Hero (2009) that specifies a decomposable sparsity structure on the concentration matrix.

Scenario B (Single-source model). In this scenario, we consider network flows consisting of OD flows originating from a single source node (See Fig. 3.3(a)). Due to the tree-structured routing, "parent" link carries accumulated flows of all its descendants. Thus its corresponding flow random variable statistically depends only on its descendant flows variables. Therefore there exists a natural ordering of all the link flow variables as in (3.2), and $\mathbf{y}$ naturally follows a DAG model. Note that the parent-child dependence relationships of the DAG are opposite to the directions of the flows.

This single-source model is useful in the context of network spoofing detec-



(a) Connectivity of Abilene network

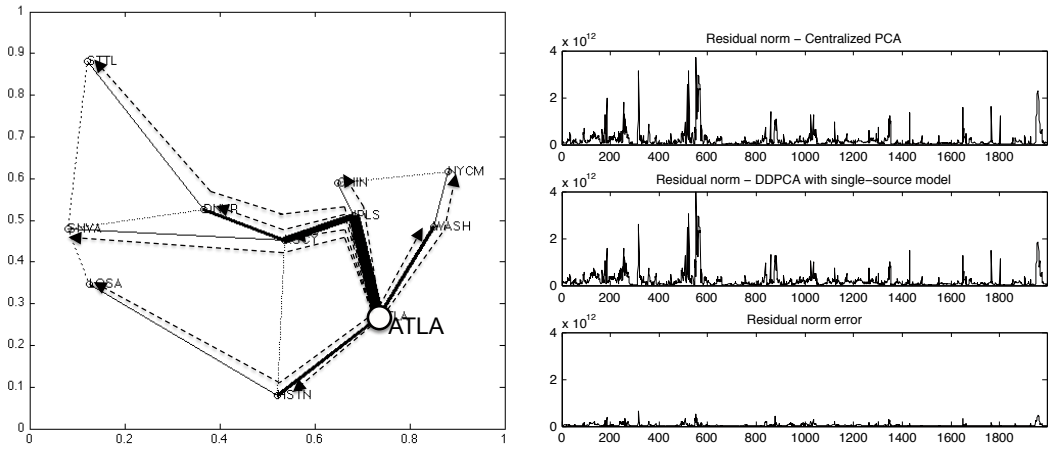
(b) Residual norm and error (decomposable model)

Figure 3.2: Anomaly detection in Abilene with decomposable model

tion (Kemmerer & Vigna, 2002), where an attacker who impersonates the identity of one of the routers making independent and unauthorized connections and injects phony packets into its OD flow tree. The attack could be detected by origin authentication at each receiver but this requires a trusted certification authority, i.e., transport layer security (TLS). A distributed flow-based anomaly detection algorithm would allow network administrator to detect the intrusion without access to private endpoint authentication information (Lakhina et al., 2004).

3.7.2 Distributed Anomaly Detection in Abilene Network

Finally, we apply the proposed DDPCA to anomaly detection of a real-world network named Abilene (Lakhina et al., 2004) based on a DAG network model. Abilene is the Internet2 backbone network which carries traffic between universities in the United States. Fig. 3.2(a) shows its connectivity map consisting of 11 nodes and 30 links. Measurements of link flow traffic data satisfy the network model (3.22), where the routing matrix $\mathbf{A}$ is known. Our goal is to detect the anomalies occurring in the OD flow vector $\mathbf{x}$ with observations of $\mathbf{y}$.



(a) OD tree from ATLA

(b) Residual norm and error (single-source model)

Figure 3.3: Anomaly detection in Abilene with single-source model

In this context, PCA is used for estimating a low dimensional principal subspace containing the nominal flow traffic. The test data are projected into the nullspace (the residual space) and the norm of the projected data is thresholded to indicate potential anomalies (Lakhina et al., 2004). Compared with centralized PCA, DDPCA enables distributed estimation of the principal subspace through decentralized computation and communication over the network.

We consider the two models as described in Section 3.7.1 for implementing the PCA algorithms:

(A) Decomposable model. We consider all the OD flows and links. From the topology of the network (See Fig. 3.2(a)), it can be seen that there exist two separators: S_1 = {DNVR-KSCY, SNVA-KSCY, LOSA-HSTN}, S_2 = {KSCY-IPLS, HSTN-ATLA} which physically separate the other three subsets of links. Then we can model the link traffic variables by a three-clique DGGM, which can be equivalently converted into a five-node DAG model.

Note that in this model, the dimensions of the separating sets ($|S_1| = 6, |S_2| = 4$) are not significantly smaller than the clique dimensions (14, 12, and 14, respectively).

Therefore the computation cost ($\sim 14^2$) and communication cost (~ 14) for implementing the DAG-based DDPCA is less expensive than the DGGM-based DPCA, which requires cubic local computation cost ($\sim 14^3$) and quadratic communication cost ($\sim 6^2$). The advantage would be more significant for large-scale networks.

(B) Single-source model. In this scenario, we consider all the OD flows originating from node ATLA, as shown in Fig. 3.3(a). The dotted lines with arrows indicate OD flows, and the thickness of solid links are proportional to the number of OD flows passing through it. As described above we can construct a 11-node sparse DAG model for the link traffic. Note that in this example, the equivalent DGGM has two additional edges and is thus less sparse than the DAG model. Again, the improvement due to sparsity level would be more significant for large-scale networks.

In our experiments on both models, we used two weeks of real-world flows data[4]. We learn the normal subspace from the first week's data using centralized PCA and DDPCA, respectively. Then for anomaly detection on the second week's link traffic, we project it on the learned residual subspace. Fig. 3.2 and Fig. 3.3 shows the norm of residual signal using centralized PCA (first row), using DDPCA (second row) and their difference (third row). We can clearly see that DDPCA successfully approximates the centralized PCA in computing the normal subspace and obtaining the residual signal under both scenarios and models. Anomaly detection can be performed by thresholding the residual norm and locating the peaks. As explained in Section 3.5, DPCA can also be applied on corresponding Markov equivalent DGGMs to achieve the same performance, but with more expensive computation and communication costs.

[4]The data are available on http://www.cs.bu.edu/~crovella/links.html

3.8 Summary of Chapter 3

We have presented a distributed PCA algorithm on directed Gaussian graphical models, called DDPCA, designed for exploiting structured sparsity arising from the Cholesky factors of the concentration matrix. We have illustrated DDPCA's computational and communication advantages in sparse directed models, and showed its application to distributed anomaly detection using link traffic of a real-world computer network.

3.9 Proof of Theorem 3.1

Proof. First we consider the first pass of bottom-up message-passing. We claim that all the bottom-up messages have zero information parameter $\mathbf{J} = 0$, and we prove it by induction. For leaf nodes, it could be easily verified. Now consider the message passed from factor node f_i to the variable node x_j, who is a parent of x_i,

$$\begin{aligned} & m_{f_i \to x_j}(x_j) \\ &= \int \left(f_i(x_i, \mathbf{x}_{pa(i)}) \prod_{y \in \{x_i\} \cup \mathbf{x}_{pa(i)\setminus j}} m_{y \to f_i}(y) \right) dx_i d\mathbf{x}_{pa(i)\setminus j}. \end{aligned} \tag{3.24}$$

Since all the top-down messages are initialized to 1, only messages involved with x_i will contribute to the integration. Also, due to the induction, the information parameter of message $m_{x_i \to f_i}$ is zero, we only have to include its potential parameter

$\mathbf{h}_{x_i \to f_i}$ for performing the integration,

$$
\begin{aligned}
& m_{f_i \to x_j}(x_j) \\
& \propto \int_{x_i} \exp\left((h_i + \mathbf{h}_{x_i \to f_i}) x_i - \frac{1}{2} (\mathbf{L}_{i,i} x_i + \mathbf{L}_{ij} x_j)^2 \right) dx_i && (3.25) \\
& \propto \exp\left(-\mathbf{L}_{ij}^T \mathbf{L}_{ii}^{-1} (h_i + \mathbf{h}_{x_i \to f_i}) x_j \right) && (3.26) \\
& \triangleq \exp\left(\mathbf{h}_{f_i \to x_j} x_j + \frac{1}{2} \mathbf{J}_{f_i \to x_j} x_j^2 \right), && (3.27)
\end{aligned}
$$

where the following formula is used

$$
\int \exp(-\frac{1}{2} a x^2 + bx) dx = C \cdot \exp(\frac{b^2}{2a}), \tag{3.28}
$$

and the updated parameters of message $m_{f_i \to x_j}$ are

$$
\begin{aligned}
\mathbf{h}_{f_i \to x_j} &= -\mathbf{L}_{ij}^T \mathbf{L}_{ii}^{-1} (h_i + \mathbf{h}_{x_i \to f_i}), && (3.29) \\
\mathbf{J}_{f_i \to x_j} &= 0. && (3.30)
\end{aligned}
$$

The resulting information parameter is also zero, which completes the induction.

The next bottom-up message-passing from variable x_j to factor f_j is simple, as

$$
m_{x_j \to f_j}(x_j) = \prod_{i \in ch(j)} m_{f_i \to x_j}(x_j), \tag{3.31}
$$

which results in the following parameter updates

$$
\begin{aligned}
\mathbf{h}_{x_j \to f_j} &= \sum_{i \in ch(j)} -\mathbf{L}_{ij}^T \mathbf{L}_{ii}^{-1} (h_i + \mathbf{h}_{x_i \to f_i}), && (3.32) \\
\mathbf{J}_{x_j \to f_j} &= 0. && (3.33)
\end{aligned}
$$

Now define a vector $\mathbf{y}$ whose j-th component is

$$y_j \triangleq \mathbf{L}_{jj}^{-1}\big(h_j + \mathbf{h}_{x_j \to f_j}\big). \tag{3.34}$$

Then it can be easily seen from (3.32) that $\mathbf{y}$ satisfies the following recursion

$$y_j = \mathbf{L}_{jj}^{-1}\big(h_j - \sum_{i \in ch(j)} \mathbf{L}_{ij}^T y_i\big), \tag{3.35}$$

the equivalence between vector $\mathbf{y}$ and the auxiliary vector introduced in Algorithm 2 with $\mathbf{h}$ as input is obvious. Therefore, the first bottom-up pass of sum-product algorithm is equivalent to the first part of Algorithm 2, which solves a linear upper triangular system via distributed backward substitution. Similarly, the second pass of top-down message-passing of sum-product algorithm can also be shown to be equivalent to the second part of Algorithm 2, namely solving a linear lower triangular system via distributed forward substitution. The resulting inferred mean vector $\mu = \mathbf{L}^{-1}\mathbf{L}^{-T}\mathbf{h} = \mathbf{J}^{-1}\mathbf{h}$ is thus exact inference. □

CHAPTER 4

Learning Latent Variable Gaussian Graphical Models with Application to Recommender Systems

In the previous two chapters, we have focused on Gaussian graphical models, for which sparsity plays a central role both statistically and computationally. Unfortunately, real-world data often does not fit well to sparse Gaussian graphical models. In this chapter, motivated by the missing-value prediction in recommender systems, we consider a family of latent variable Gaussian graphical models (LVGGM). In LVGGM, the model is conditionally sparse given a set of latent variables, but is marginally non-sparse. In particular, the inverse covariance matrix has a low-rank plus sparse structure, and can be learned in a regularized maximum likelihood framework. We derive novel parameter estimation error bounds for LVGGM under mild conditions in the high-dimensional setting, which complements the existing theory on the structural learning. Our theory is validated through numerical simulations, and the effectiveness of LVGGM is illustrated on a real-world movie rating data set.

4.1 Introduction

Critical to many statistical inference tasks in complex real-world systems, such as prediction and detection, is the ability to extract and estimate distributional characteristics from the observations. Unfortunately, in the high-dimensional regime such model estimation often leads to ill-posed problems, particularly when the number of observations n (or sample size) is comparable to or fewer than the ambient dimensionality p of the model (*i.e.,* the "large p, small n" problem). This challenge arises in many modern real-world applications ranging from recommender systems, gene microarray data, and financial data, to name a few. To perform accurate model parameter estimation and subsequent statistical inference, low dimensional structure is often imposed for regularization (Negahban et al., 2012).

For Gaussian-distributed data, the central problem is often to estimate the inverse covariance matrix (alternatively known as the precision, concentration or information matrix). Gaussian graphical models (GGM) provide an efficient representation of the precision matrix through a graph that represents non-zeros in the matrix (Lauritzen, 1996). In high-dimensional regimes, this graph can be forced to be sparse, imposing a low-dimensional structure on the GGM. For sufficiently sparse GGM, statistically consistent estimates of the model structure (*i.e.,* sparsistency) can be achieved (*e.g.,* Ravikumar et al. (2011)). On the computational side, sparsity also leads to reduced complexity of the estimator (Hsieh et al., 2013). However, when the true distribution can not be well-approximated by a sparse GGM, the standard learning paradigm suffers from either large estimation bias due to enforcing a overly sparse model, or degraded computation time for a dense model. Both result in suboptimal performance in the subsequent inference tasks.

In this chapter, we consider a new class of high-dimensional GGM for extending the standard sparse GGM. The proposed model is motivated by many real-world applications, where there exist certain exogenous and often latent factors affecting a

large portion of the variables. Examples are the price of oil on the airlines' stock price variables (Choi et al., 2010), and the genres on movie rating variables. Conditioning on these *global* effects, the variables are assumed to have highly *localized* interactions, which can be well-fitted by a sparse GGM. However, due to the marginalization over global effects, the observed (marginal) GGM, and its corresponding precision matrix, is not sparse. Unfortunately, in this regime, existing theoretical results and computational tools for sparse GGM are not applicable.

To address this problem, we propose to use latent variable Gaussian graphical models (LVGGM) for modeling and statistical inference. LVGGM introduce latent variables to capture the correlations due to the global effects, and the remaining effects are captured by a conditionally sparse graphical model. The resulting marginal precision matrix of the LVGGM has a sparse plus low-rank structure, therefore we consider a regularized maximum likelihood (ML) approach for parameter estimation (previously considered by Chandrasekaran et al. (2012)). By utilizing the *almost strong convexity* (Kakade et al., 2010) of the log-likelihood, we derive a non-asymptotic parameter error bound for the regularized ML estimator. Our derived bounds apply to the high-dimensional setting of $p \gg n$ due to restricted strong convexity (Negahban et al., 2012) and certain structural incoherence between the sparse and low-rank components of the precision matrix (Yang & Ravikumar, 2013).

We show that for sufficiently large n, the Frobenius norm error of the precision matrix of LVGGM converges at the rate $\mathcal{O}(\sqrt{\frac{(s+r_{\text{eff}} \cdot r)\log p}{n}})$, where s is the number of non-zeros in the conditionally sparse precision matrix, r_{eff} is the effective rank of the covariance matrix and r is the number of latent variables. This rate is in general significantly faster than the standard convergence rate of $\mathcal{O}(\sqrt{\frac{p^2 \log p}{n}})$ for an unstructured dense GGM. This result offers a compelling argument for using LVGGM over sparse GGM for many inference problems.

The chapter is structured as follows. In Section 4.2 we review the relevant prior

literature. In Section 4.3 we formulate the LVGGM estimation problem. In Section 4.4 the main theoretical results are presented. Experimental results are shown in Section 4.5 and we conclude the chapter in Section 4.7. Throughout the chapter, we use boldface letters to denote vectors and matrices. $\|\cdot\|_1$, $\|\cdot\|_2$, $\|\cdot\|_F$, $\|\cdot\|_*$ denote the elementwise ℓ_1, spectral, Frobenius, and nuclear matrix norms, respectively.

4.2 Background and Related Work

The problem of learning GGM with sparse inverse covariance matrices using ℓ_1-regularized maximum likelihood estimation, often referred to as the graphical lasso (Glasso) problem, has been studied in Friedman et al. (2008); Ravikumar et al. (2011); Rothman et al. (2008). In particular, the authors of Ravikumar et al. (2011) study the model selection consistency (*i.e.,* "sparsistency") under certain incoherence condition. Beyond sparse GGM, Choi et al. (2010) propose a multi-resolution extension of a GGM augmented with sparse inter-level correlations, while in Choi et al. (2011) the authors consider latent tree-structured graphical models. Both models lead to computationally efficient inference and learning algorithms but restrict the latent structure to trees. Recently, Liu & Willsky (2013) consider a computationally efficient learning algorithm for a class of conditionally tree-structured LVGGM.

The work that is most relevant to ours is by Chandrasekaran et al. (2012), who study the LVGGM learning problem, but focus on the simultaneous model selection consistency of both the sparse and low-rank components. In contrast, in this chapter we focus on the Frobenius norm error bounds for estimating the precision matrix of LVGGM. Although structural consistency can be useful for deriving insights, parameter estimation error analysis is of equal or greater importance in practice. Since it provides additional, and usually more direct, insights into factors influencing the performance of the subsequent statistical inference tasks, such as prediction and de-

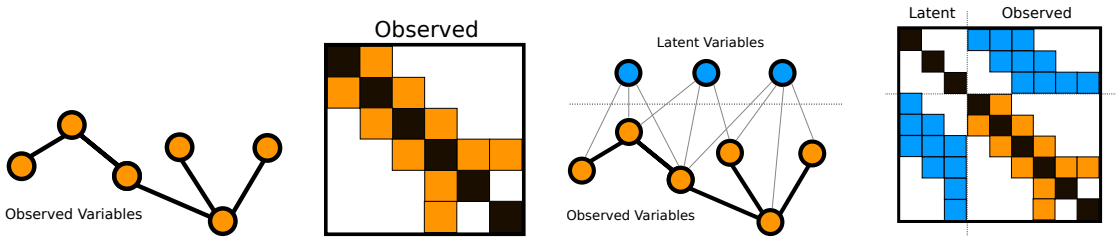


Figure 4.1: Illustrations of a sparse Gaussian graphical model (GGM) (left) and a latent variable Gaussian graphical model (LVGGM) (right). (A) Example of a sparse GGM with only observed variables, (B) Sparsity pattern of example sparse GGM's precision matrix, (C) Example of a LVGGM with both observed and latent variables, (D) Sparsity pattern of example LVGGM's precision matrix.

tection. Also, compared with Chandrasekaran et al. (2012), our Frobenius norm error bounds are derived under mild condition on the Fisher information of the distribution.

We note that there is a fundamentally different line of work on estimating models with a similar structural composition, known as *robust PCA* (Candès et al., 2011). In robust PCA, the data matrix is modeled as "low-rank plus sparse". This model has been applied to extracting the salient foreground from background in videos, and detecting malicious user ratings in recommender system data (Xu et al., 2012). In contrast, the equivalent covariance model of our LVGGM can be decomposed into a low-rank plus a dense matrix whose *inverse* is sparse. A similar covariance model has recently been studied by Kalaitzis & Lawrence (2012), in which an EM algorithm is proposed for estimation but no theoretical error bounds are derived. In this chapter, we instead focus on the precision matrix parameterization, which enables model estimation through a convex optimization. This formulation is of both theoretical and computational importance.

4.3 Problem Setup

In this section, we briefly review Gaussian graphical models and formulate the problem of latent variable Gaussian graphical model estimation via a regularized maximum

likelihood optimization.

4.3.1 Gaussian Graphical Models

Consider a p-dimensional random vector $\mathbf{x}$ associated with an undirected graph $\mathcal{G} = (V_G, E_G)$, where V_G is a set of nodes corresponding to elements of $\mathbf{x}$ and E_G is a set of edges connecting nodes (including self-edges for each node). Then $\mathbf{x}$ follows a graphical model distribution if it satisfies the Markov property with respect to $\mathcal{G}$: for any pair of nonadjacent nodes in $\mathcal{G}$, the corresponding pair of variables in $\mathbf{x}$ are conditionally independent given the remaining variables, *i.e.*, $x_i \perp x_j \mid \mathbf{x}_{\setminus i,j}$, for all $(i,j) \notin E_G$.

If $\mathbf{x}$ follows a multivariate Gaussian distribution, the corresponding graphical model is called a Gaussian graphical model (GGM). We assume without loss of generality that $\mathbf{x}$ has zero mean. The Markov property in GGM is manifested in the sparsity pattern of the inverse covariance matrix $\mathbf{J}$:

$$\mathbf{J}_{i,j} = 0 \text{ for all } i \neq j, (i,j) \notin E. \tag{4.1}$$

An example of this property for sparse GGM is shown in Figure 4.1(a) and 4.1(b).

The precision matrix parameterization arises in many statistical inference problems for Gaussian distributions, in areas such as belief propagation (Malioutov et al., 2006), linear prediction, portfolio selection in financial data (Ledoit & Wolf, 2003), and anomaly detection (Chen et al., 2011). Estimation of the precision matrix in GGM is the first step in these inference problems.

4.3.2 Latent Variable Gaussian Graphical Models

Unfortunately, due to the presence of global factors that destroy sparsity, real-world observations often do not conform exactly to a sparse GGM (Choi et al., 2010, 2011).

By introducing latent variables (denoted as a r-dimensional random vector $\mathbf{x}_L$) to capture global factors, we can generalize the GGM. Specifically, we construct a model that is conditionally a GGM, *i.e.,* one that has a sparse precision matrix given knowledge of latent variables, $\mathbf{x}_L$.

Defining the p observed variables as $\mathbf{x}_O$, we assume the joint distribution of the $(p+r)$-dimensional concatenated random vector $\mathbf{x} = (\mathbf{x}_O, \mathbf{x}_L)$ follows a Gaussian distribution with covariance matrix $\mathbf{\Omega}$ and precision matrix $\mathbf{J} = \mathbf{\Omega}^{-1}$. An example of this structure can be seen in Figure 4.1(c) and 4.1(d). Marginalizing over the latent variables $\mathbf{x}_L$, the distribution of the observed variables $\mathbf{x}_O$ remains Gaussian with observed covariance matrix, $\mathbf{\Sigma} = \mathbf{\Omega}_{O,O}$. The observed precision matrix $\Theta \in \mathbb{R}^{p \times p}$ satisfies:

$$\Theta = \mathbf{\Sigma}^{-1} = \underbrace{\mathbf{J}_{O,O}}_{\mathbf{S}} \underbrace{-\mathbf{J}_{O,L}\mathbf{J}_{L,L}^{-1}\mathbf{J}_{L,O}}_{\mathbf{L}}, \tag{4.2}$$

where we have defined $\mathbf{S} := \mathbf{J}_{O,O}$ and $\mathbf{L} := -\mathbf{J}_{O,L}\mathbf{J}_{L,L}^{-1}\mathbf{J}_{L,O}$. Thus, the marginal precision matrix can be written as $\Theta = \mathbf{S} + \mathbf{L}$, the sum of a sparse and a low-rank matrix. Similar to standard GGM, we parameterize the marginal distribution through the precision matrix. We refer to this model as the latent variable GGM, or LVGGM.

The LVGGM is a hierarchical model that generalizes the (sparse) GGM. Note that $\mathbf{S}^{-1} = \mathbf{J}_{O,O}^{-1} = \mathbf{\Omega}_{O,O} - \mathbf{\Omega}_{O,L}\mathbf{\Omega}_{L,L}^{-1}\mathbf{\Omega}_{L,O}$ is the covariance matrix of the conditional distribution of the observed variables. The matrix is not generally sparse, even though $\mathbf{S}$ is assumed to be sparse. We will also assume that the number of latent variables is much smaller than the number of observed variables, *i.e.,* $r \ll p$. We place no sparsity restrictions on the dependencies between the observed and latent variables – the submatrices $\mathbf{J}_{O,L}$ and $\mathbf{J}_{L,O}$ could be dense. As a result, the $p \times p$ matrix $\mathbf{L} = -\mathbf{J}_{O,L}\mathbf{J}_{L,L}^{-1}\mathbf{J}_{L,O}$ is low-rank and potentially dense. The sparse plus low-rank structure of the marginal precision matrix Θ is the key property of the precision

matrix that will be exploited for model estimation.

The structural assumptions on the precision matrix of the LVGGM can be validated on real-world recommender system data. See Section 4.6 for more details.

4.3.3 Effective Rank of Covariance Matrix

We introduce the *effective rank* of a matrix, which will be useful to derived high-dimensional error bounds. The effective rank of a matrix $\mathbf{\Sigma}$ is defined as (Vershynin, 2010):

$$r_{\text{eff}}(\mathbf{\Sigma}) := \text{tr}(\mathbf{\Sigma})/\|\mathbf{\Sigma}\|_2. \tag{4.3}$$

The effective rank can be considered a measure of the concentration level of the spectrum of $\mathbf{\Sigma}$. As we will show in Section 4.5.1, in many situations the effective rank of the covariance matrix corresponding to a LVGGM is much smaller than p. Under this condition, our theoretical results in the sequel provide a tight Frobenius norm estimation error bound, which is significantly improved upon the error bound derived without the effective rank assumption.

4.3.4 Regularized ML Estimation of LVGGM

Available are n samples $x_1, x_2, \ldots, x_n$ from a LVGGM model $\mathbf{x}_O$, concatenated into a data matrix $\mathbf{X} \in \mathbb{R}^{p\times n}$. The negative log-likelihood function is

$$\mathcal{L}(\Theta; \mathbf{X}) = \langle \widehat{\mathbf{\Sigma}}, \Theta \rangle - \log\det(\Theta), \tag{4.4}$$

where $\widehat{\mathbf{\Sigma}} := \frac{1}{n}\mathbf{X}^T\mathbf{X}$ is the sample covariance matrix. The regularized ML estimate minimizes the objective function $\mathcal{L}(\Theta; \mathbf{X}) + \lambda\mathcal{R}(\Theta)$, where the regularization parameter $\lambda > 0$, and the regularization function $\mathcal{R}(\Theta)$ is designed to enforce the sparse

plus low-rank structure on Θ.

Similar to Chandrasekaran et al. (2012), we consider the following regularized ML estimation problem:

$$\begin{aligned} \min_{\mathbf{S},\mathbf{L}} \quad & \mathcal{L}(\mathbf{S}+\mathbf{L};\mathbf{X}) + \lambda\|\mathbf{S}\|_1 + \mu\|\mathbf{L}\|_* \\ \text{s.t.} \quad & -\mathbf{L} \succeq \mathbf{0},\ \mathbf{S}+\mathbf{L} \succeq \mathbf{0}, \end{aligned} \tag{4.5}$$

where the corresponding regularization function is the sum of two regularizers: $\mathcal{R}(\Theta) = \|\mathbf{S}\|_1 + \frac{\mu}{\lambda}\|\mathbf{L}\|_*$, each of which has been shown to promote sparse (low-rank) structure in $\mathbf{S}$ ($\mathbf{L}$, respectively) (Negahban et al., 2012). Constants $\lambda, \mu > 0$ are regularization parameters corresponding to the two functions, respectively. The LVGGM estimator is defined as a solution to the above convex optimization problem (4.5). Efficient convex solver, such as Ma et al. (2013), can be used to solve.

Note that only the sample covariance matrix is needed as the input of the regularized ML estimation (4.5) (see (4.4) for the expression of likelihood). Therefore in the presence of missing observations, as in the case of recommender systems, an estimator of the fully-observed covariance matrix constructed from the incomplete observations can be used instead. See Kolar & Xing (2012); Loh & Wainwright (2012) for some recent discussions on the issue of consistent estimation in the presence of missing observations.

4.4 Error Bounds on ML LVGGM Estimation

We analyze the regularized ML estimation problem (4.5) and provide Frobenius norm error bounds for estimating the precision matrix in high-dimensional setting. We adopt the decomposable regularization framework of Negahban et al. (2012); Agarwal et al. (2012); Yang & Ravikumar (2013) to derive these bounds. In contrast to this prior work, here we focus on multiple decomposable regularizers interacting with the

non-quadratic log-likelihood loss function encountered in the LVGGM. Two important ingredients in the derivations are the restricted strong convexity of the loss function, and an incoherence condition between the two structured subspaces containing the sparse and low-rank components ($\mathbf{S}$ and $\mathbf{L}$). We show that under assumptions on the Fisher information these two conditions are verified.

In the following subsections, first we define some necessary notation, then we introduce the assumptions and place them in the context of prior literature, and finally we state the main results in Theorem 4.1 and Theorem 4.2.

4.4.1 Decomposable Regularizers and Subspace Notation

In this subsection we introduce the notion of decomposable regularizers and the corresponding subspace pairs. We refer the reader to Negahban et al. (2012) for more details.

Consider a pair of subspaces $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$, where $\mathcal{M} \subset \overline{\mathcal{M}} \subset \mathbb{R}^{p \times p}$. $\mathcal{R}(\cdot)$ is called a decomposable regularization function with respect to the subspace pair if, for any $u \in \mathcal{M}, v \in \overline{\mathcal{M}}^{\perp}$, we have $\mathcal{R}(u+v) = \mathcal{R}(u) + \mathcal{R}(v)$.

For the sparse and low-rank matrix-valued parameters, the following two subspace pairs and their corresponding decomposable regularizers are considered:

- *Sparse matrices.* Let $E \subseteq \{1, \ldots, p\} \times \{1, \ldots, p\}$ be a subset of index pairs (edges). Define $\mathcal{M}(E) = \overline{\mathcal{M}}(E)$ as the subspace of all sparse matrices in $\mathbb{R}^{p \times p}$ that are supported in subsets of E, *i.e.,* $\mathcal{P}_{\mathcal{M}(E)}(\mathbf{A}) = \mathbf{A}_E$. A decomposable regularizer is the ℓ_1 norm, since $\|\mathbf{A}\|_1 = \|\mathbf{A}_E\|_1 + \|\mathbf{A}_{E^C}\|_1$.
- *Low-rank PSD matrices.* Consider a class of low-rank and positive semi-definite matrices $\mathcal{A} \subset \mathbb{S}_+^{p \times p}$ which have rank $r \leq p$. For any given matrix $\mathbf{A} \in \mathcal{A}$, let $\mathrm{col}(\mathbf{A})$ denote its column space. Let $U \subset \mathbb{R}^n$ be a r-dimensional subspace and define the

subspace $\mathcal{M}(U)$ and the perturbation subspace $\overline{\mathcal{M}}^{\perp}(U)$ as

$$\mathcal{M}(U) := \{\mathbf{A} \in \mathbb{R}^{n \times p} \mid \mathrm{col}(\mathbf{A}) \subseteq U\},$$
$$\overline{\mathcal{M}}^{\perp}(U) := \{\mathbf{A} \in \mathbb{R}^{n \times p} \mid \mathrm{col}(\mathbf{A}) \subseteq U^{\perp}\}.$$

Then the nuclear norm $\mathcal{R}_L(\cdot) = \|\cdot\|_*$ is a decomposable regularization function with respect to the subspace pair $(\mathcal{M}(U), \overline{\mathcal{M}}^{\perp}(U))$.

For the true model parameter Θ^*, we define its associated *structural error set* with respect to a subspace $\mathcal{M}$ as (Negahban et al., 2012):

$$\mathbb{C}(\mathcal{M}, \overline{\mathcal{M}}^{\perp}; \Theta^*) := \left\{\Delta \in \mathbb{R}^{n \times p} \mid \mathcal{R}(\Delta_{\overline{\mathcal{M}}^{\perp}}) \leq 3\mathcal{R}(\Delta_{\overline{\mathcal{M}}}) + 4\mathcal{R}(\Theta^*_{\overline{\mathcal{M}}^{\perp}})\right\}.$$

By construction, if the norm of the projection of the true parameter Θ^* into $\overline{\mathcal{M}}^{\perp}$ is small, then elements Δ in this structural error set also have limited projection onto the perturbation subspace $\overline{\mathcal{M}}^{\perp}$.

Now let Θ^* be the true (marginal) precision matrix of the LVGGM, and let the sparse and low-rank components be $\mathbf{S}^*$ and $\mathbf{L}^*$, respectively. For the defined subspace pairs $(\mathcal{M}(E), \overline{\mathcal{M}}(E)^{\perp})$ and $(\mathcal{M}(U), \overline{\mathcal{M}}(U)^{\perp})$, we use $\mathbb{C}(E)$ and $\mathbb{C}(U)$ as the shorthand notations for the corresponding structural error sets centered at $\mathbf{S}^*$ and $\mathbf{L}^*$, *i.e.,* $\mathbb{C}(\mathcal{M}(E), \overline{\mathcal{M}}(E)^{\perp}; \mathbf{S}^*)$ and $\mathbb{C}(\mathcal{M}(U), \overline{\mathcal{M}}(U)^{\perp}; \mathbf{L}^*)$, respectively. Later, we will consider the perturbation of Θ^* along restricted directions in these two sets.

4.4.2 Assumptions on Fisher Information

We characterize the interaction between the elements in the two subspaces through their inner products using the Hessian of the loss function, also known as the *Fisher information* of the distribution. Denoting the Fisher information matrix of a Gaussian distribution as $\mathcal{F}^*$ (evaluated at Θ^*), we find that $\mathcal{F}^* = \Theta^{*-1} \otimes \Theta^{*-1}$, where $\otimes$ is the

Kronecker product. We define the *Fisher inner product* between two matrices Δ_A and Δ_B as

$$\langle \Delta_A, \Delta_B \rangle_{\mathcal{F}^*} := \mathrm{vec}(\Delta_A)^T \mathcal{F}^* \mathrm{vec}(\Delta_B) \tag{4.6}$$

$$= \mathrm{Tr}(\Theta^{*-1} \Delta_A \Theta^{*-1} \Delta_B), \tag{4.7}$$

where $\mathrm{vec}(\cdot)$ denotes the vectorization of a matrix.

Similar to prior work of Kakade et al. (2010), we define the induced *Fisher norm* of a matrix Δ as

$$\|\Delta\|_{\mathcal{F}^*}^2 := \mathrm{vec}(\Delta)^T \mathcal{F}^* \mathrm{vec}(\Delta) \tag{4.8}$$

$$= \mathrm{Tr}(\Theta^{*-1} \Delta \Theta^{*-1} \Delta). \tag{4.9}$$

The first assumption we make is the following *Restricted Fisher Eigenvalue* (RFE) condition on the true precision model with respect to the sparse and low-rank structural error sets.

Assumption 4.1 (**Restricted Fisher Eigenvalue**). *There exists some constant* $\kappa_{\min}^* > 0$*, such that for all* $\Delta \in \mathbb{C}(E) \cup \mathbb{C}(U)$*, the following holds:*

$$\|\Delta\|_{\mathcal{F}^*}^2 \geq \kappa_{\min}^* \|\Delta\|_F^2 \ . \tag{4.10}$$

This RFE condition generalizes the restricted eigenvalue (RE) condition for sparsity-promoting linear regression problems Bickel et al. (2009). It assumes that the minimum eigenvalue of the Fisher information is bounded away from zero along the directions $\mathbb{C}(E)$ and $\mathbb{C}(U)$. Due to the identity (4.8) and properties of the Kronecker product, a trivial lower bound for $\kappa_{\min}^*$ is $\lambda_{\min}^2(\Theta^*)$, where $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalue. In the high-dimensional setting, the RFE parameter $\kappa_{\min}^*$, which is defined only with respect to the above restricted set of directions, can be substantially

larger than $\lambda^2_{\min}(\Theta^*)$. As a result, the derived error bounds, which depend on $\kappa^*_{\min}$, are generally tighter than the bounds depending on $\lambda^2_{\min}(\Theta^*)$ (cf. Theorem 4.1).

Due to the sparse plus low-rank superpositioned structure, we impose a type of incoherence between the two structural error sets to ensure consistent estimation of the combined model. The incoherence condition will limit the interaction between elements from the two sets. For our problem, such interaction occurs through their inner products with the Fisher information, which motivates the following *Structural Fisher Incoherence* (SFI) assumption (which generalizes the *C-Linear* assumption proposed in Yang & Ravikumar (2013)).

Let $\mathcal{P}_E := \mathcal{P}_{\overline{\mathcal{M}}(E)}$ denote the projection operator corresponding to the subspace $\overline{\mathcal{M}}(E)$. Similarly define $\mathcal{P}_U := \mathcal{P}_{\overline{\mathcal{M}}(U)}$, $\mathcal{P}_{E^\perp} := \mathcal{P}_{\overline{\mathcal{M}}(E)^\perp}$, and $\mathcal{P}_{U^\perp} := \mathcal{P}_{\overline{\mathcal{M}}(U)^\perp}$. We assume the following condition on the Fisher information.

Assumption 4.2 (**Structural Fisher Incoherence**). *Given a constant $M > 6$, a set of regularization parameters (λ, μ), and the subspace pairs $(\mathcal{M}(E), \overline{\mathcal{M}}(E)^\perp)$ and $(\mathcal{M}(U), \overline{\mathcal{M}}(U)^\perp)$ as defined above, let $\Lambda = 2 + 3\max\left\{\frac{\lambda\sqrt{s}}{\mu\sqrt{r}}, \frac{\mu\sqrt{r}}{\lambda\sqrt{s}}\right\}$, where $s = |E|$ and $r = rank(U)$. Then the Fisher information $\mathcal{F}^*$ satisfies:*

$$\max\left\{\overline{\sigma}\left(\mathcal{P}_E\mathcal{F}^*\mathcal{P}_U\right), \overline{\sigma}\left(\mathcal{P}_{E^\perp}\mathcal{F}^*\mathcal{P}_U\right), \overline{\sigma}\left(\mathcal{P}_E\mathcal{F}^*\mathcal{P}_{U^\perp}\right), \overline{\sigma}\left(\mathcal{P}_{E^\perp}\mathcal{F}^*\mathcal{P}_{U^\perp}\right)\right\} \leq \frac{\kappa^*_{\min}}{c_1\Lambda^2},$$

where $\overline{\sigma}(\cdot)$ denotes the maximum singular value, and constant c_1 is defined as $c_1 = \frac{16M}{M-6}$.

The constant M is related to a "burn-in" period after which the likelihood loss function has desirable properties in a small neighborhood of the true parameter. In particular, when $M = 7$, the constant $c_1 = 112$ suffices for our theory to hold. See the main theorem and its proof for more discussion on this quantity.

It is interesting to compare our SFI assumption to other similar assumptions in the literature of GGM estimation. In Ravikumar et al. (2011), a form of irrepre-

sentability condition is assumed, which limits the induced ℓ_1 norm of a matrix that is similar to the projected Fisher information onto the sparse matrix subspace pair. In Chandrasekaran et al. (2012), the notion of irrepresentability is extended to two subspace pairs (*i.e.,* sparse and low-rank), but detailed behaviors of the projected Fisher information are controlled (see the main assumption on page 1949 of Chandrasekaran et al. (2012)). For model selection consistency, a more general form of irrepresentability has been shown to be necessary for model selection consistency, see Lee et al. (2013) for a recent discussion. In contrast to the above line of work, the SFI assumption we make only controls the maximum singular values of the projected Fisher information. This can be explained as we are interested in bounding a weaker quantity, the Frobenius norm of the parameter estimation error, instead of establishing the stronger model selection consistency of Ravikumar et al. (2011) or the algebraic consistency as in Chandrasekaran et al. (2012).

4.4.3 Error Bounds for LVGGM Estimation

We have the following bound on the parameter error of the estimated precision matrix of LVVGGM, $\widehat{\Theta} = \widehat{\mathbf{S}} + \widehat{\mathbf{L}}$, obtained by solving the regularized ML problem (4.5).

Theorem 4.1. *Suppose Assumption 4.1 and 4.2 hold for the true marginal precision matrix Θ^*, and the regularization parameters are chosen such that*

$$\lambda \geq 2\|\mathbf{\Sigma}^* - \widehat{\mathbf{\Sigma}}\|_\infty \quad \textit{and} \quad \mu \geq 2\|\mathbf{\Sigma}^* - \widehat{\mathbf{\Sigma}}\|_2. \tag{4.11}$$

Given a constant $M > 6$, if an optimal solution pair $(\widehat{\mathbf{S}}, \widehat{\mathbf{L}})$ to the convex program (4.5) satisfies

$$\max\{\|\widehat{\mathbf{S}} - \mathbf{S}^*\|_{\mathcal{F}^*}, \|\widehat{\mathbf{L}} - \mathbf{L}^*\|_{\mathcal{F}^*}\} \leq \frac{1}{6M^2}, \tag{4.12}$$

then we have the following error bound for the estimated precision matrix $\widehat{\Theta} = \widehat{\mathbf{S}} + \widehat{\mathbf{L}}$*:*

$$\|\widehat{\Theta} - \Theta^*\|_F \leq \frac{6}{\kappa_{\mathcal{L}}} \max\left\{\lambda\sqrt{s}, \mu\sqrt{r}\right\} + \sqrt{\frac{8r_\perp^*}{\kappa_{\mathcal{L}}}}, \tag{4.13}$$

where $s = |E|$*,* $r = rank(U)$*, and*

$$\kappa_{\mathcal{L}} := \frac{M-2}{2(M-1)}\kappa_{\min}^*, \tag{4.14}$$

$$r_\perp^* := \lambda \sum_{(j,k)\notin E} |\mathbf{S}_{jk}^*| + \mu \sum_{j=r+1}^{p} \sigma_j(\mathbf{L}^*). \tag{4.15}$$

Proof sketch. The proof is inspired by Yang & Ravikumar (2013), in which a parameter estimation error bound is proven for estimating a class of superposition-structured parameters, such as sparse plus low-rank, through M-estimation with decomposable regularizers. Critical to specializing this framework to our LVGGM estimation problem is to verify two conditions on the log-likelihood loss function (4.4): the restricted strong convexity (RSC) and structural incoherence (SI). The RSC condition (which originally proposed in Negahban et al. (2012)) specifies the loss function to be sufficiently curved (*i.e.* lower bounded by a quadratic function) along a restricted set of directions (defined by $\mathbb{C}(E)$ and $\mathbb{C}(U)$). On the other hand, the SI condition effectively limits certain interaction between elements from the above two structural error sets. In Yang & Ravikumar (2013), under certain *C-linear* assumptions, the RSC and SI conditions are verified for several problems with quadratic loss functions. For the LVGGM estimation problem, however, the technical difficulty lies in the non-quadratic log-likelihood loss (4.4), for which the previously established RSC and SI conditions do not hold.

To deal with this difficulty, we leverage the *almost strong convexity* properties (Kakade et al., 2010) to characterize the convergence behavior of the sum of higher-order terms in the Taylor series of the log-likelihood loss function. We show

that in the regime specified by condition (4.12), the loss function can be well-approximated by the sum of a quadratic function and a residual term. Under this condition, the RFE assumption (Assumption 4.1) guarantees the RSC condition (cf. Lemma 4.2), and the SFI assumption (Assumption 4.2) leads to SI condition to hold (cf. Lemma 4.4). Theorem 4.1 can then be proven by the general theorem in Yang & Ravikumar (2013). A detailed proof of Theorem 4.1 can be found in Section 4.8.1. □

We make the following remarks:

- The error bound (4.13) is a family of upper bounds defined by different sets of subspace pairs $(\mathcal{M}(E), \overline{\mathcal{M}}(E)^{\perp})$ and $(\mathcal{M}(U), \overline{\mathcal{M}}(U)^{\perp})$. The tightest bound can be achieved by appropriately choosing E and U. The first additive term in (4.13) captures effect of the estimation error, while the second term captures the approximation error. In many cases it is reasonable to assume the approximation error is zero, then the error bound reduces to the first additive term.

- We note that similar derivations also apply to ℓ_1-regularized estimation of sparse GGM. For the sparse GGM, only Assumption 4.1 is required, and the derivations largely simplify. The final error bound also contains estimation and approximation errors, depending only on the sparse matrix subspace pair. However, when the true precision matrix Θ^* cannot be well-approximated as a sparse matrix (such as the LVGGM case), the approximation error would be much worse, leading to an inefficient learning rate.

- We finally remark that the SFI assumption can be relaxed to an even milder incoherence condition, $\|\mathbf{L}\|_{\infty} \leq \alpha$, as considered in Agarwal et al. (2012). Following similar derivations as in the proof of Theorem 4.1, the corresponding

error bound can be obtained. However, as a result of this incoherence assumption, the error bound would contain an additional incoherence term which does not vanish to zero even with infinite samples. This disadvantage is overcome under the structural incoherence condition.

The statement of Theorem 4.1 is deterministic in nature and applies to any optimum of the convex program. However, the condition on the regularization parameters (4.11) and the error bound depend on the sampled data (in particular the sample covariance matrix $\widehat{\Sigma}$), which is random. Therefore the key to specifying the regularization parameters, and hence obtaining error bounds independent of data, is to derive tight deviation bounds of the sample covariance matrix in terms of the ℓ_∞ and ℓ_2 norms, such that condition (4.11) holds with high probability. These bounds can be obtained by using concentration inequalities for Gaussian distributions, which leads to the following corollary.

Corollary 4.1. *Let the same assumptions in Theorem 4.1 hold. Given constants $C_1 > 1$ and $C_2 \geq 1$, assume that the number of samples n satisfies $n \geq \max\{4C_1^2 \log p,\ C_2^2 p\}$, and that the regularization parameters satisfy*

$$\lambda = 160 C_1 \overline{\sigma}^* \sqrt{\frac{\log p}{n}} \quad \text{and} \quad \mu = 16 C_2 \rho^* \sqrt{\frac{p}{n}}, \tag{4.16}$$

where $\overline{\sigma}^* = \max_i \mathbf{\Sigma}^*_{i,i}$ and $\rho^* = \|\mathbf{\Sigma}^*\|_2$. Then with probability at least $1 - 4p^{-2(C_1-1)} - 2\exp(-\frac{C_2^2 p}{2})$, we have

$$\|\widehat{\Theta} - \Theta^*\|_F \leq c_1 \sqrt{\frac{s \log p}{n}} + c_2 \sqrt{\frac{rp}{n}}, \tag{4.17}$$

where $c_1 = \frac{960}{\kappa_{\mathcal{L}}} \overline{\sigma}^*$ and $c_2 = \frac{96}{\kappa_{\mathcal{L}}} \rho^*$.

Remark: The estimation error (4.17) consists of two terms corresponding to the sparse and low-rank components, respectively. Note its resemblance to the error

bounds of robust PCA (*e.g.,* Agarwal et al. (2012); Yang & Ravikumar (2013)) and the derived bound in Chandrasekaran et al. (2012). In particular, the first term in (4.17) was on the same order as the estimation error of a sparse GGM (Ravikumar et al., 2011). However, due to the presence of latent variables, both the sample requirement (*i.e.,* $n \gtrsim p$) and the combined error bound are worse than those for learning the sparse conditional GGM.

Next we consider a scenario under which this additional disadvantage is largely removed. Assume that the true marginal covariance matrix $\mathbf{\Sigma}^*$ has an effective rank $r_{\text{eff}} := r_{\text{eff}}(\mathbf{\Sigma}^*)$ (recall $r_{\text{eff}}(\mathbf{\Sigma}^*) := \text{tr}(\mathbf{\Sigma}^*)/\|\mathbf{\Sigma}^*\|_2$) that is much smaller than p. Then, by using recent advances on the asymptotic behavior of the sample covariance matrix (Lounici, 2012), we can obtain a much tighter bound which only depends on p logarithmically, as stated in the following theorem.

Theorem 4.2. *Let the same assumptions in Theorem 4.1 hold. Given a constant $C_1 > 1$, assume that the number of observations n satisfies $n \geq \max\left\{4C_1 \log p,\ C_3 r_{\mathit{eff}} \log^2(2p)\right\}$, and the regularization parameters satisfy*

$$\lambda = 160 C_1 \overline{\sigma}^* \sqrt{\frac{\log p}{n}} \quad \textit{and} \quad \mu = C_4 \rho^* \sqrt{\frac{r_{\mathit{eff}} \log p}{n}}, \tag{4.18}$$

where $\overline{\sigma}^* = \max_i \mathbf{\Sigma}^*_{i,i}$, $\rho^* = \|\mathbf{\Sigma}^*\|_2$, and $C_3, C_4 > 0$ are sufficiently large constants. Then with probability at least $1 - 2p^{-2(C_1-1)} - (2p)^{-1}$, we have

$$\|\widehat{\Theta} - \Theta^*\|_F \leq \tilde{c}_1 \sqrt{\frac{s \log p}{n}} + \tilde{c}_2 \sqrt{\frac{r_{\mathit{eff}} \cdot r \log(2p)}{n}}, \tag{4.19}$$

where $\tilde{c}_1 = \frac{960}{\kappa_{\mathcal{L}}} \overline{\sigma}^*$, $\tilde{c}_2 = \frac{8C_4}{3\kappa_{\mathcal{L}}} \rho^*$.

Proof sketch. Same as Corollary 4.1, we need to verify that the choices of regularization parameters (4.18) satisfy the condition (4.11) with high probability. Since the choice of λ has been verified in Corollary 4.1, it only remains to verify the condition

on μ. To this end, we make use of the following sharp bound on the spectral norm deviation of the sample covariance matrix:

Lemma 4.1 (Lounici (2012)). *Let $\widehat{\mathbf{\Sigma}}$ be a sample covariance matrix constructed from n* i.i.d. *samples from a p-dimensional Gaussian distribution $\mathcal{N}(0, \mathbf{\Sigma}^*)$. Then with probability at least $1 - (2p)^{-1}$,*

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_2 \leq C\|\mathbf{\Sigma}^*\|_2 \max\left\{\sqrt{\frac{2r_{\mathit{eff}}\log(2p)}{n}}, \frac{2r_{\mathit{eff}}\log(2p)(3/8+\log(2pn)}{n}\right\},$$

where $C > 0$ is an absolute constant.

Then as commented in Lounici (2012) (Prop. 3), when the sample size n is sufficiently large such that $n \geq C_3 r_{\text{eff}} \log^2 \max\{2p, n\}$, where $C_3 > 0$ is a large constant, the choice of regularization parameter μ as in (4.18) suffices for the condition (4.11) to hold with high probability. □

Notice that when $r_{\text{eff}} \ll p$, the error bound (4.19) is significantly tighter than the bound (4.17). Also the sample requirement $n \gtrsim r_{\text{eff}} \log(p)$ is much milder. This result implies the efficiency of LVGGM learning when the true covariance model has a low effective rank.

4.5 Experiments

We use a set of simulations on synthetic data to verify our reduced effective rank assumption on the covariance matrix of LVGGM, and the derived error bounds in Theorem 4.2.

4.5.1 Effective Rank of Covariance of LVGGM

To better understand the effective rank of the covariance matrix of LVGGM, it is convenient to consider a hierarchical generating process for the observed variables:

$x_O \sim \mathbf{A}x_L + z$, where $x_L \sim \mathcal{N}(\mathbf{0}, \mathbf{\Omega}_{L,L})$ are the latent variables, $\mathbf{A} := \mathbf{J}_{O,O}^{-1}\mathbf{J}_{O,L} \in \mathbb{R}^{p\times r}$, and $z \sim \mathcal{N}(\mathbf{0}, \mathbf{S}^{-1})$ captures the conditional effects. The marginal covariance matrix of the observed variables can be represented as

$$\mathbf{\Sigma} = \underbrace{\mathbf{A}\mathbf{\Omega}_{L,L}\mathbf{A}^T}_{\mathbf{G}} + \mathbf{S}^{-1}, \tag{4.20}$$

where $\mathbf{G}$ is a low-rank covariance matrix (global effects), and $\mathbf{S}^{-1}$ is a non-sparse covariance matrix (conditionally local effects) whose inverse is sparse. While the low-rank global effects naturally result in a concentrated spectrum, the sparse-inverse local effects generally contribute to a diffuse spectrum. The effective rank, which is the sum of all eigenvalues divided by the magnitude of the largest one, depends on the relative energy ratio between $\mathbf{G}$ and $\mathbf{S}^{-1}$.

Since an exact characterization of the effective rank in terms of $\mathbf{A}$, $\mathbf{\Omega}_{L,L}$, and $\mathbf{S}$ tends to be difficult, we use Monte Carlo simulations to investigate synthetic LVGGM that conform to our assumptions. We generate LVGGM with independent latent variables (*i.e.,* diagonal $\mathbf{J}_{L,L}$), dense latent-observed submatrix $\mathbf{J}_{L,O}$, and a sparse conditional GGM $\mathbf{J}_{O,O}$ for observed variable with a random sparsity pattern (sparsity level $\approx 5\%$). We fix the number of latent variables to be 10, and vary the number of observed variables $p = \{80, 120, 200, 500\}$. By scaling the magnitudes of the elements in the latent variable submatrix, we sweep through the relative energy ratio between the global and local factors, *i.e.,* $\mathrm{Tr}(\mathbf{G})/\mathrm{Tr}(\mathbf{S}^{-1})$ from 0.1 to 10. After 550 realizations for each value of p, we plot the empirical effective ranks of observed covariance matrices in Figure 4.2.

As seen in the figure, when the global factor dominates (*i.e.,* the ratio is large), the effective rank of the covariance matrix is very small, as expected. On the other hand, when the local effects become stronger (*e.g.,* when the number of observed variables p increases) the effective rank increases, but at a very mild rate. In particular, when p

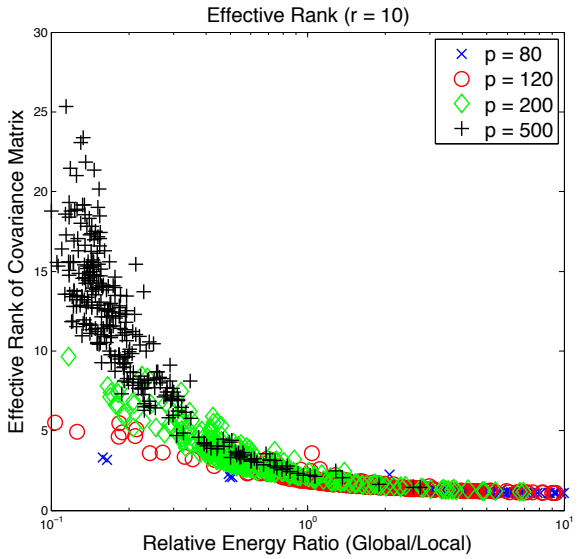


Figure 4.2: Effective ranks of covariance matrices of LVGGM with various global/local energy ratios.

increases from 80 to 500, the maximum empirical effective rank in our simulation only increases from 4 to 26. For all of our simulated LVGGM, the empirical effective ranks are observed as at least an order of magnitude smaller than p. This mild growing rate of the effective rank (compared to p) will lead to our improved error bound in Theorem 4.2 to hold.

4.5.2 Frobenius Norm Error of LVGGM Estimation

We simulate LVGGM data with number of observed variables $p = \{160, 200, 320, 400\}$ and number of latent variables in the set $r = \{0.1, 0.15, 0.2, 0.3\}p$. The sparse conditional GGM is a chain graph whose associated precision matrix is tridiagonal with off-diagonal elements $\mathbf{S}_{i,i-1} = \mathbf{S}_{i,i+1} = 0.4\mathbf{S}_{i,i}$ for $i = \{2, \ldots, p-1\}$. For each configuration of p and r, we draw n samples from the LVGGM, where n ranges from 200 to 1000. Using these samples, the precision matrix $\widehat{\Theta}$ is learned by solving the regularized ML estimation problem (4.5). As shown in Section 4.5.1, the effective rank of the covariance matrix grows mildly. Then Theorem 4.2 predicts that the Frobenius error of the estimated precision matrix of LVGGM should scale as

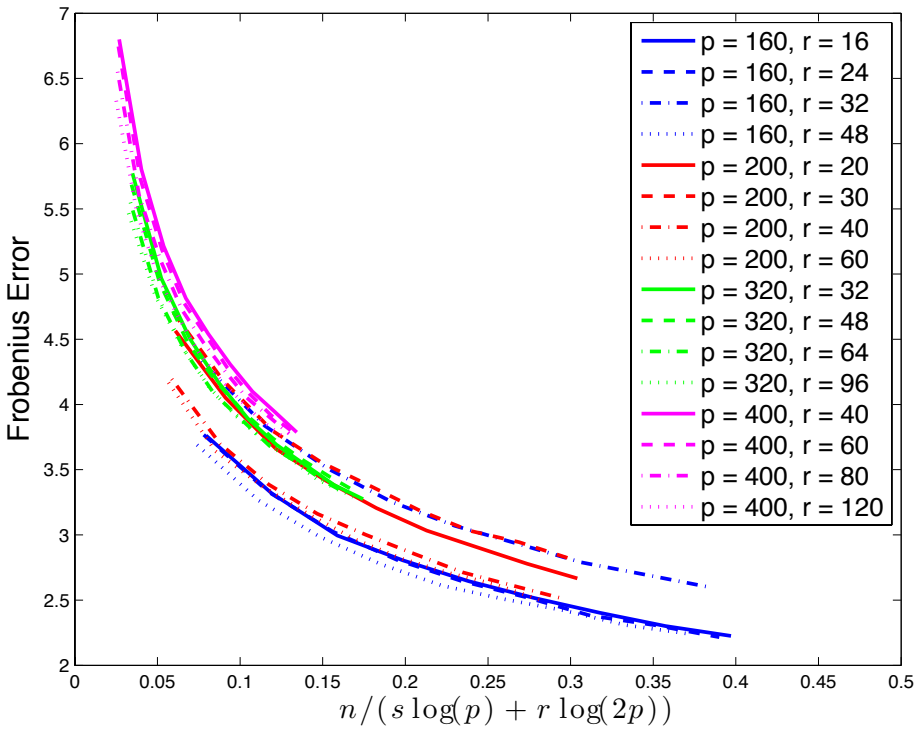


Figure 4.3: Simulations for chain graphical models with latent variables. Plots of Frobenius norm error $\|\widehat{\Theta} - \Theta^*\|_F$ versus the rescaled sample size $n/(s\log(p) + r\log(2p))$.

$\|\widehat{\Theta} - \Theta^*\|_F \asymp \sqrt{(s\log(p) + r\log(2p))/n}$, when the regularization parameters are chosen such that $\lambda \asymp \overline{\sigma}^* \sqrt{\frac{\log(p)}{n}}$ and $\mu \asymp \rho^* \sqrt{\frac{r_{\text{eff}}\log(p)}{n}}$. Guided by this theoretical result, we set the regularization parameters as $\lambda = C_a \overline{\sigma}^* \sqrt{\frac{\log(p)}{n}}$ and $\mu = C_b \rho^* \sqrt{\frac{r_{\text{eff}}\log(p)}{n}}$, where constants C_a and C_b are cross-validated and then fixed for all test data sets with different configurations. We plot the Frobenius estimation errors against the rescaled sample size $n/(s\log(p) + r\log(2p))$ in Figure 4.3. With a wide range of configurations, almost all the empirical error curves for models align and have the form of $f(t) \propto t^{-1/2}$ when the sample size is rescaled, as predicted by Theorem 4.2. In practice when the true model is unknown, one could set the regularization parameters according to the sample versions of the quantities $\overline{\sigma}^*$ and ρ^*, as discussed in Lounici (2012).

4.6 LVGGM for Recommender Systems

As mentioned in Section 1.2.2, the LVGGM can be motivated by recommender systems. In this section, we illustrate how the recommender system data can be modeled by LVGGM, and then present its superior performance in predicting missing values in recommender systems.

4.6.1 Background and Problem Formulation

In recommender systems, available is an incomplete data sample matrix $\mathbf{R} \in \mathbb{R}^{n \times p}$ where each element $R_{i,j}$ denotes the rating score that the i^{th} user gives to the j^{th} item (*e.g.,* movie). The goal is to predict the unobserved or missing ratings based on the observed ones. Fundamental to missing-value prediction is to choose and learn a model that governs the joint distribution of the movie rating variables.

Popularized by the famous *Netflix Prize* challenge[1], there has been a substantial body of literature on the recommender systems. Here we briefly review two most popular approaches which have been widely studied in the literature and implemented in practical recommender systems. Interested readers are referred to Su & Khoshgoftaar (2009) for a survey of various techniques, and Koren (2008); Bell & Koren (2007) for the approaches that won the competition (in which the two methods described below are both used).

One of the most successful approaches for recommender systems is the latent factor model, also known as the low rank model or matrix factorization technique Koren et al. (2009). As the names suggest, this model posits that there exist a small number (r) of global "latent factors" affecting the movie ratings given by the users. Accordingly, each user and item are represented as a low-dimensional feature vector, and the rating given by the i^{th} user to the j^{th} movie is modeled as the similarity, *i.e.,* inner

[1] http://www.netflixprize.com

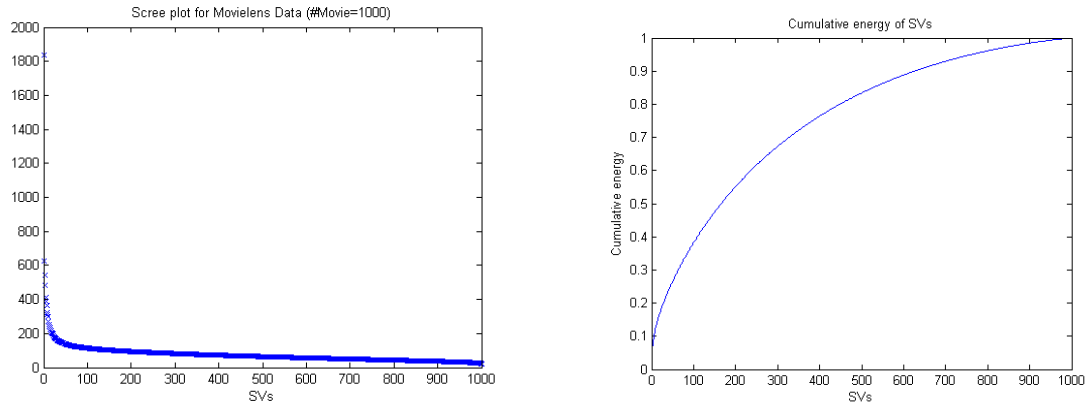


Figure 4.4: Illustration of the high-rank residual in the real-world *Movielens* data set. Left: Plot of magnitudes of singular values (SV) of top 1000 movies in *Movielens* data. Right: Plot of energy captured by the number of leading SVs. Only 50% of the total energy is captured by the largest 200 SVs, indicating the existence of a strong high-rank residual.

product of the two corresponding feature vectors:

$$R_{i,j} \approx \mathbf{u}_i^T \cdot \mathbf{v}_j, \tag{4.21}$$

where $\mathbf{u}_i, \mathbf{v}_j \in \mathbb{R}^r$ are r-dimensional user- and item-specific feature vectors, respectively. As a result, the rating matrix can be approximated by the product of two feature matrices:

$$\mathbf{R} \approx \mathbf{U}^T \mathbf{V}, \tag{4.22}$$

where $\mathbf{U} \in \mathbb{R}^{r \times n}$ and $\mathbf{V} \in \mathbb{R}^{r \times p}$ contains the user and item feature vectors, respectively. Note that r, the number of latent factors, is often much lower than the dimensions of the rating matrix. Therefore $\mathbf{R}$ is assumed to be low-rank, and it admits an approximate matrix factorization as shown in (4.22).

Despite the success of low-rank models on many real-world data sets, a complementary model is often required for capturing the residual information after removing

the low-rank dominating component, as the real-world data often exhibits a high-rank and heavy-tailed spectrum. In Figure 4.4, we plot the magnitudes of singular values (SV) of the rating matrix corresponding to the top 1000 movies in the *Movielens* data. As can be seen, only 50% of the energy is captured by the largest 200 SVs, suggesting the existence of a strong high-rank residual.

Another widely adopted method for recommender systems is the item-graph approach Koren (2008). In this framework, an item-graph is constructed where each node represents an item, and an edge connecting two nodes suggests significant correlation or similarity between the ratings given to the two associated items. Then the rating given by the i^{th} user to the j^{th} item is modeled as a weighted linear combination of the rating he/she gives to the neighboring movies:

$$R_{ij} \approx b_{ij} + \sum_{k \in N(j)} \theta^i_{jk}(R_{ik} - b_{ik}), \tag{4.23}$$

where $N(j)$ denotes the set of neighbors of node j in the item-graph, and θ^i_{jk} is a coefficient characterizing the rating similarity between item j and k for user i. The item graph captures the correlation structure between the ratings, and it utilizes this structure for missing-rating prediction. However, the main drawback of the item-graph approach is a lack of theoretical guidance on choosing appropriate graph structure and guarantees for the model estimation performance.

An alternative model that extends the item-graph approach is to use the Markov random field to model the rating variables. When the data is Gaussian distributed, this model is also known as the Gaussian graphical model (GGM). As described in Section 4.3.1, GGM provides an efficient characterization of the joint distribution using the conditional dependency graph, whose edge pattern coincides with the non-zero pattern of the inverse covariance matrix. For the recommender system data, we assume the rating vector for each user (which contains the ratings for all the movies)

is an *i.i.d.* sample from the joint distribution of the movie rating variables, which is modeled by a GGM. Note that the Gaussian distribution assumption (after removing the mean effects in the ratings) has been widely adopted in the recommender systems literature as a standard approximation technique (see, for example, Lawrence & Urtasun (2009)). More accurate distributional characterization has also been studied (see, for example, Salakhutdinov & Mnih (2008); Guo et al. (2014)), but at the cost of more computationally expensive estimation algorithms and more complicated theoretical analysis. The GGM approach is not constrained by the low-rank assumption imposed in the latent factor model, and by parameterizing the joint distribution by conditional dependencies, it is also statistically more efficient than the item-graph approach, which essentially parameterizes the distribution using marginal dependencies. Given the ratings from n users, we can learn the structure and parameters of the inverse covariance matrix using the ℓ_1-regularized maximum likelihood estimation (Ravikumar et al., 2011). Related literature on regularized learning of GGM was reviewed in Section 4.2.

Unfortunately, in the presence of global and latent variables, both theory and empirical results suggest the ineffectiveness of using GGM, since the marginal inverse covariance matrix corresponding to the observed variables (*i.e.,* the movie rating variables) is not sparse. Enforcing sparsity will lead to large bias and, as a result, suboptimal prediction performance.

As motivated in Section 4.3.2, we propose to use the LVGGM for modeling the recommender system data. Specifically, the latent variables are introduced to capture the global and latent effects that affecting a large portion of the movie ratings (such the movie genres). Conditioning on these latent factors, the conditional distribution of the observed movie rating variables is modeled by a sparse GGM. The resulting marginal inverse covariance matrix has a "low-rank plus sparse" structure, and can be learned through a regularized maximum likelihood estimation (4.5).

Note that only the sample covariance matrix of the movie rating variables are needed for the regularized ML estimation (4.5), therefore a "plug-in" estimator of the fully-observed covariance matrix constructed from the incomplete observations can be used instead. In particular, following Kolar & Xing (2012), we construct the sample covariance from the incomplete observations by weighting each sample covariances by its specific empirical missingness level. This construction yields a consistent estimator for the covariance matrix, and it alleviates the missing-value effect. We refer the interested readers to Kolar & Xing (2012) for analysis on this procedure and Loh & Wainwright (2012) for a recent related work.

4.6.2 Validation of Structural Assumptions of LVGGM

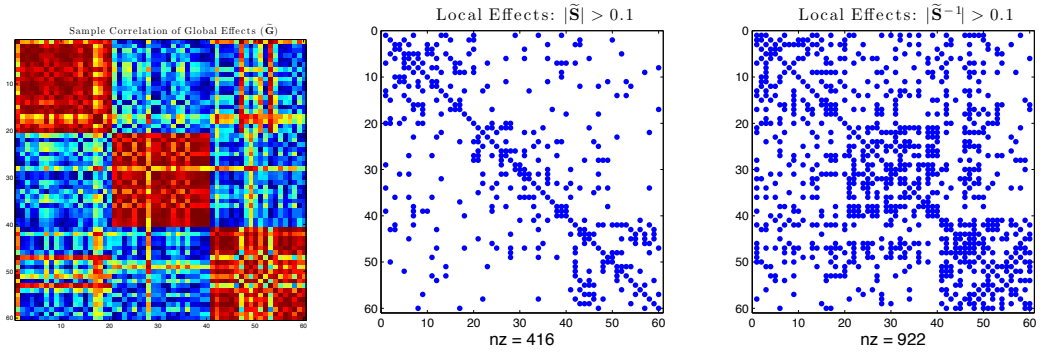
Next we use *Movielens*[2], a real-world movie rating data set, to validate the structural assumptions of the LVGGM. For this purpose, we manually impose the movie genre as a global effect by constructing a data set with 60 movies from three genres, where each genre contains 20 movies. To minimize the effect of the missingness in the data, we choose the ratings given by the most active 600 users and for the most rated 20 movies from each of the following three genres: *Horror*, *Children's*, and *Action*. This results in a 600×60 rating matrix with 56% completeness. We consider the joint distribution of 60 movie rating variables as a LVGGM with three latent variables. As mentioned above, each user's rating vector is treated as an *i.i.d.* sample from the LVGGM. Since the true covariance matrix is unknown, we use the sample covariance matrix as a reasonable proxy since $n \gg p$. Each covariance element is weighted by the actual number of observations to compensate for the missingness in the data.

According to the decomposition of covariance matrix of a LVGGM (see Eq. (4.20)), we expect that this genre effect can be extracted with a low-rank component $\mathbf{G}$ in the covariance matrix. Our goal is to examine whether the remaining residual component

[2]http://movielens.org

can be reasonably fitted by a sparse GGM (*i.e.,*, has a sparse precision matrix $\mathbf{S}$ but less sparse covariance matrix $\mathbf{S}^{-1}$. Note that, for illustration purpose, we will use eigen-decomposition to decompose the sample covariance matrix, which in general will results in different estimates as those obtained by the regularized maximum likelihood estimation.

We decompose the rating matrix into two matrices: a rank-3 matrix spanned by its top three leading singular vectors, and a residual matrix capturing the conditional effects. We denote the covariance matrix of the low-rank component as $\widetilde{\mathbf{G}}$, and the sparse precision matrix of the residual component as $\widetilde{\mathbf{S}}$. A heat map of the normalized $\widetilde{\mathbf{G}}$ is shown in Figure 4.5(a), and the sparsity patterns of the normalized $\widetilde{\mathbf{S}}$ and $\widetilde{\mathbf{S}}^{-1}$ (*i.e.,* the covariance of the residual) are shown in Figure 4.5(b), thresholded by 0.1. As expected, the low-rank $\widetilde{\mathbf{G}}$ captures the structure of the global effects (*i.e.,* movie genre), and the residual can be well-modeled by a sparse GGM – its precision matrix is much sparser than the covariance matrix. In addition, we find the effective rank of the covariance is equal to 7.4, much smaller than the number of variables, 60, which validates the low effective rank assumption.



(a) $\widetilde{\mathbf{G}}$ of *Movielens* (b) $|\widetilde{\mathbf{S}}|$ and $|\widetilde{\mathbf{S}}^{-1}|$ of *Movielens*

Figure 4.5: Illustration of LVGGM assumptions on *Movielens* data set. (a): Heat maps of the leading low-rank matrices capturing the global effects (movie genre). (b): Sparsity patterns of the precision and covariance matrices of the remaining conditional effects.

4.6.3 Missing-value prediction using LVGGM

Under the LVGGM, the observed movie rating variables are jointly Gaussian distributed as follows:

$$\mathbf{x}_O \sim \mathcal{N}(\mathbf{0}, \Theta^{-1}),$$

where the marginal inverse covariance matrix Θ is given in Eq. (4.2), and can be learned from the training data.

Given a user, we denote the set of observed (or *known*) movie ratings from this user by K, and denote the set of missing (or *unknown*) ratings by U. Then the (conditional) distribution of the variables $\mathbf{x}_U$ conditioning on $\mathbf{x}_K$ is

$$\mathbf{x}_U | \mathbf{x}_K \sim \mathcal{N}(-\Theta_{U,U}^{-1}\Theta_{U,K}\mathbf{x}_K, \Theta_{U,U}^{-1}).$$

For prediction of $\mathbf{x}_U$, we can use the estimated $\widehat{\Theta}$ to calculate the conditional distribution, and then use the corresponding conditional mean as the predicted values for $\mathbf{x}_U$, *i.e.,*

$$\text{(Prediction)} \quad \widehat{\mathbf{x}}_U | \mathbf{x}_K = -\Theta_{U,U}^{-1}\Theta_{U,K}\mathbf{x}_K. \tag{4.24}$$

This prediction rule notably bears much resemblance to the well-known item-graph approach for recommender systems (Koren, 2008).

4.6.4 Experiment results on *Movielens*

We implement and compare the above-mentioned three models for missing-value prediction on a subset of *Movielens* data set: the latent factor model (LFM), the GGM without latent variables, and the LVGGM. The LFM is learned for various number of latent factors and by the alternating least square (ALS) algorithm (Koren, 2008).

The GGM without latent variables is learned by the ℓ_1-regularized maximum likelihood (Ravikumar et al., 2011). The LVGGM is learned by Eq. (4.5). The number of latent factors in the LFM and regularization parameters in learning GGM and LVGGM are cross-validated on a training data set. The optimal predication performance (measured by the root-mean-square errors (RMSE)) on a testing data set and some descriptive statistics of the optimal models are reported in Table 4.1.

	LFM	GGM	LVGGM
Rank	20	\	55
Average graph degree	\	76.5	**16.4**
Total # parameters	30000	38727	36189
RMSE	0.8186	0.8166	**0.8115**

Table 4.1: Prediction performance and descriptive statistics for three models on *Movielens* data set with 500 movies and 1000 users. The data missing rate is 53%. LVGGM achieves the lowest RMSE with comparable total number of parameters. The conditional graph of LVGGM is much sparser than the GGM without latent variables (in terms of the average graph degree).

As can be seen, the proposed LVGGM achieves superior performance among three models with comparable total number of parameters. Notably, the conditional graph of the optimal LVGGM is much sparser than the optimal GGM without latent variables, which validates our intuitions.

4.7 Summary of Chapter 4

We have considered a family of latent variable Gaussian graphical model (LVGGM) whose precision matrix has a sparse plus low-rank structure. For this LVGGM, we derive the parameter estimation error bounds for regularized maximum likelihood estimation under mild conditions. The theory is validated by numerical simulations, and the proposed LVVGM achieves superior prediction performance on a recommender system data set.

4.8 Proofs for Chapter 4

4.8.1 Proof of Theorem 4.1

In Yang & Ravikumar (2013), the authors proved a general superpositioned parameter estimate error bound using the decomposable regularized framework. Theorem 4.1 can be proven similarly by specializing the result in Yang & Ravikumar (2013) to the LVGGM learning problem (4.5). Then it suffices to verify the two critical conditions (C3) and (C4) in Yang & Ravikumar (2013) (the other two conditions are trivial to verify for our problem), which we introduce and elaborate in this section.

Restricted strong convexity. Let $\delta\mathcal{L}(\Delta; \Theta^*)$ denote the remainder term in first-order Taylor series approximation of the loss function $\mathcal{L}(\cdot)$ at the true parameter Θ^* with respect to a perturbation $\Delta = \Theta^* - \widehat{\Theta}$:

$$\delta\mathcal{L}(\Delta; \Theta^*) := \mathcal{L}(\Theta^* + \Delta) - \mathcal{L}(\Theta^*) - \langle \nabla\mathcal{L}(\Theta^*), \Delta \rangle. \tag{4.25}$$

In Negahban et al. (2012), the authors introduce the *restricted strong convexity* (RSC) condition, which specifies that given some set $\mathbb{C} \subseteq \mathbb{R}^{p \times p}$, there exists some curvature parameter $\kappa_{\mathcal{L}} > 0$ and tolerance function $\tau_{\mathcal{L}}$, such that the following holds:

$$\delta\mathcal{L}(\Delta; \Theta^*) \geq \kappa_{\mathcal{L}} \|\Delta\|_F^2 - \tau_{\mathcal{L}}(\Theta^*), \ \forall \Delta \in \mathbb{C}. \tag{4.26}$$

The RSC condition guarantees sufficient curvature of the loss function at the true parameter along some directions specified by set $\mathbb{C}$. This condition is critical for consistent estimation in the high-dimensional regime, since standard strong convexity usually does not hold in the $p \gg n$ setting.

The following shows that the restricted Fisher eigenvalue conditions defined in Assumption 4.1 implies the RSC condition.

Lemma 4.2 (RSC condition)**.** *Suppose Assumption 4.1 holds for the true marginal precision matrix* Θ^* *and let* $M > 2$*. Then for all* $\Delta \in \mathbb{C}(E) \cup \mathbb{C}(U)$*, such that* $\|\Delta\|_{\mathcal{F}^*}^2 \leq \frac{1}{2M^2}$*, the RSC condition* (4.26) *is satisfied with the curvature parameter* $\kappa_{\mathcal{L}} = \frac{M-2}{2(M-1)}\kappa_{\min}^*$ *and the tolerance function* $\tau_{\mathcal{L}} = 0$*.*

The proof of Lemma 4.2 is largely inspired by Kakade et al. (2010), in which it is shown that exponential family distributions exhibit *almost strong convexity* in a neighborhood. The RFE assumption makes connection between this property and the RSC condition. A proof of Lemma 4.2 is given in the Section 4.8.2.

Note there is an important difference between the RSC condition considered here and the condition introduced in Agarwal et al. (2012). The RSC condition considered here is satisfied with respect to the error matrices of each simple structure separately, while the RSC condition in Agarwal et al. (2012) is required for the combined error matrices (defined in the product space of two sets), which could lie in a significantly larger set.

Structural incoherence. The second ingredient for consistent estimation of the sparse plus low-rank parameter Θ, is some type of incoherence condition between the sparse and low-rank components. In the present work, we consider the *structural incoherence* condition that was proposed more recently in Yang & Ravikumar (2013). This condition allows for a vanishing error bound when n goes to infinity, and is applicable to more general loss functions, such as the log-likelihood function in Eq. (4.4).

Define the following incoherence measure of the loss function $\mathcal{L}$ for two structural error matrices Δ_S and Δ_L:

$$c_{\mathcal{L}}(\Delta_S, \Delta_L; \Theta^*) := |\mathcal{L}(\Theta^* + \Delta_S + \Delta_L) + \mathcal{L}(\Theta^*)$$
$$- \mathcal{L}(\Theta^* + \Delta_S) - \mathcal{L}(\Theta^* + \Delta_L)|, \forall \Delta_S \in \mathbb{C}(E), \Delta_L \in \mathbb{C}(U).$$

Then the *structural incoherence* (SI) condition is satisfied if the following relation holds for all $\Delta_S \in \mathbb{C}(E)$ and $\Delta_L \in \mathbb{C}(U)$:

$$c_{\mathcal{L}}(\Delta_S, \Delta_L; \Theta^*) \leq \frac{\kappa_{\mathcal{L}}}{2} \|(\|\Delta_S\|_F^2 + \|\Delta_L\|_F^2), \tag{4.27}$$

where $\kappa_{\mathcal{L}}$ is the curvature parameter in the RSC condition (4.26).

The following lemma shows that, in addition to the restricted Fisher eigenvalue assumption (Assumption 4.1), if the true marginal model also satisfies the structural Fisher incoherence assumption (Assumption 4.2), then the above SI condition on the likelihood loss function is guaranteed.

Lemma 4.3 (SI condition)**.** *Suppose Assumption 4.1 and 4.2 hold for the true marginal precision matrix* Θ^* *and let* $M > 6$*. Then the SI condition* (4.27) *is satisfied for all* $\Delta_S \in \mathbb{C}(E)$ *and* $\Delta_L \in \mathbb{C}(U)$*, such that* $\max\{\|\Delta_S\|_{\mathcal{F}^*}^2, \|\Delta_L\|_{\mathcal{F}^*}^2\} \leq \frac{1}{6M^2}$*. The curvature parameter* $\kappa_{\mathcal{L}}$ *is the same as in Lemma 4.2, i.e.,* $\kappa_{\mathcal{L}} = \frac{M-2}{2(M-1)}\kappa_{\min}^*$.

The proof of Lemma 4.3 is in Section 4.8.3.

Finally, under Assumption 4.1 and Assumption 4.2, Lemma 4.2 and 4.3 imply the RSC and SI conditions hold for our LVGGM learning problem, respectively. Thus Theorem 4.1 can be proven by directly appealing to Theorem 1 in Yang & Ravikumar (2013).

4.8.2 Proof of Lemma 4.2

Proof. The remainder term in the first-order Taylor series of the negative log-likelihood (4.4) of GGM takes the following form:

$$\begin{aligned}\delta\mathcal{L}(\Delta; \Theta^*) &= \mathcal{L}(\Theta^* + \Delta) - \mathcal{L}(\Theta^*) - \langle \nabla\mathcal{L}(\Theta^*), \Delta\rangle \\ &= \langle \Theta^{*-1}, \Delta\rangle - \log\det(\Theta^* + \Delta) + \log\det(\Theta^*).\end{aligned}$$

For $s \in (0, 1]$, define the Taylor series of function $g(s; \Theta^*) := \log\det(\Theta^* + s\Delta)$ at Θ^*

$$g(s; \Theta^*) = \log\det(\Theta^* + s\Delta) = \sum_{k=0}^{\infty} \frac{c_k(\Delta)s^k}{k!}, \tag{4.28}$$

where $c_k(\Delta) := g^{(k)}(s; \Theta^*)$ is the k-th derivative of the log det function at Θ^*. Define $c_0(\Delta) := \log\det(\Theta^*)$, the remainder can be expressed as:

$$\delta\mathcal{L}(s\Delta; \Theta^*) = \sum_{k=2}^{\infty} \frac{c_k(\Delta)s^k}{k!} = \frac{c_2(\Delta)s^2}{2} + \sum_{k=3}^{\infty} \frac{c_k(\Delta)s^k}{k!} = \frac{c_2(\Delta)s^2}{2} + \delta g(s; \Delta, \Theta^*), \tag{4.29}$$

where the second term $\delta g(s)$ is defined as the second-order Taylor error of the log-determinant function. Next we show that this error term, which is the sum of all the higher-order terms, can be bounded by a quadratic term in a small neighborhood around Θ^*.

For exponential family distributions (Gaussian as an example), the log-partition function (*i.e.,* log det function for Gaussian) coincides with the *cumulant generating function*. This implies that the derivatives $c_k(\Delta)$ are the corresponding cumulants of the distribution, which can be shown to converge to zero quite rapidly. Indeed, in Kakade et al. (2010) the authors show that for a univariate random variable z under an exponential family distribution, its k-th order cumulant satisfies

$$\left|\frac{c_k(z)}{c_2(z)^{k/2}}\right| \leq \frac{1}{2}k!\alpha^{k-2}, \quad \forall k \geq 3, \tag{4.30}$$

where α is a finite constant, and the second-order cumulant coincides with the Fisher norm of the deviation $c_2(\Delta) = \|\Delta\|^2_{\mathcal{F}^*}$ due to the definition of the Fisher information. For multivariate Gaussian distributions, $\alpha = \sqrt{2}$ suffices for the above relation to hold (see Sec. 3.2.2 in Kakade et al. (2010)).

Therefore we bound the second-order Taylor error term in Eq. (4.29) as follows (similar to Kakade et al. (2010)):

$$|\delta g(s;\Delta,\Theta^*)| = \left|\sum_{k=3}^{\infty}\frac{c_k(\Delta)s^k}{k!}\right| \tag{4.31}$$

$$\leq \frac{1}{2}\sum_{k=3}^{\infty}2^{\frac{k}{2}-1}c_2(\Delta)^{k/2}s^k \tag{4.32}$$

$$\leq \frac{s^2c_2(\Delta)}{2}\sum_{k=1}^{\infty}(s\sqrt{2c_2(\Delta)})^k \tag{4.33}$$

$$\overset{(i)}{\leq} \frac{s^2c_2(\Delta)}{2}\sum_{k=1}^{\infty}\frac{1}{M^k} \tag{4.34}$$

$$= \frac{s^2c_2(\Delta)}{2(M-1)} \tag{4.35}$$

$$\leq \frac{c_2(\Delta)}{2(M-1)}\frac{1}{\max\{2M^2c_2(\Delta),1\}} \tag{4.36}$$

$$\overset{(ii)}{=} \frac{c_2(\Delta)}{2(M-1)} \tag{4.37}$$

where (i) and (ii) are due to our conditions on $c_2(\Delta)$ (*i.e.*, $\|\Delta\|_{\mathcal{F}^*}^2 \leq \frac{1}{2M^2}$) and $s \leq 1$. Then we obtain a lower bound for $\delta\mathcal{L}(\Delta;\Theta^*)$:

$$\delta\mathcal{L}(\Delta;\Theta^*) \geq \frac{c_2(\Delta)}{2} + \delta g(s;\Delta,\Theta^*) \geq \left(\frac{1}{2} - \frac{1}{2(M-1)}\right)c_2(\Delta) \overset{(ii)}{\geq} \frac{M-2}{2(M-1)}\kappa^*_{\min}\|\Delta\|_F^2, \tag{4.38}$$

where (ii) is due to the RFE condition. Therefore the RSC condition is satisfied with the curvature parameter $\kappa_{\mathcal{L}} := \frac{M-2}{2(M-1)}\kappa^*_{\min}$ and a zero tolerance parameter $\tau_{\mathcal{L}} = 0$.

□

4.8.3 Proof of Lemma 4.3

Proof. First we state the following lemma which gives a bound on the magnitude of *Fisher inner product* between elements from the two sets.

Lemma 4.4. *Suppose Assumption 4.1 and 4.2 hold for the true marginal precision matrix* Θ^*. *Then given a constant* $M \geq 6$, *the following inequality holds for all* $\Delta_S \in \mathbb{C}(E)$ *and* $\Delta_L \in \mathbb{C}(U)$ *such that* $\max\{\|\Delta_S\|_{\mathcal{F}^*}^2, \|\Delta_L\|_{\mathcal{F}^*}^2\} \leq \frac{1}{6M^2}$*:*

$$|\langle \Delta_S, \Delta_L \rangle_{\mathcal{F}^*}| \leq \psi \left(\|\Delta_S\|_{\mathcal{F}^*}^2 + \|\Delta_L\|_{\mathcal{F}^*}^2 \right), \tag{4.39}$$

where $\psi := \frac{1}{4} - \frac{3}{2M}$.

The proof of Lemma 4.4 follows similarly as that of the Proposition 2 in Yang & Ravikumar (2013), and hence is omitted.

Next we prove Lemma 4.3 using the above result. Following similar derivations as in the proof of Lemma 4.2, the incoherence measure in the SI condition can be simplified to

$$c_{\mathcal{L}}(\Delta_S, \Delta_L; \Theta^*) := |\delta\mathcal{L}(\Delta_S + \Delta_L; \Theta^*) - \delta\mathcal{L}(\Delta_S; \Theta^*) - \delta\mathcal{L}(\Delta_L; \Theta^*)| .$$

Using the remainder in the Taylor series of $\delta\mathcal{L}$ (4.29), the incoherence measure can

be expressed as:

$$
\begin{aligned}
&c_{\mathcal{L}}(\Delta_S, \Delta_L; \Theta^*) \\
=&\left| \frac{c_2(\Delta_S + \Delta_L)}{2} + \delta g(s; \Delta_S + \Delta_L) - \left(\frac{c_2(\Delta_S)}{2} + \delta g(s_1; \Delta_S) \right) - \left(\frac{c_2(\Delta_L)}{2} + \delta g(s_2; \Delta_L) \right) \right| \\
\leq& \left| \frac{c_2(\Delta_S + \Delta_L)}{2} - \frac{c_2(\Delta_S)}{2} - \frac{c_2(\Delta_L)}{2} \right| + |\delta g(s; \Delta_S + \Delta_L)| + |\delta g(s_1; \Delta_S)| + |\delta g(s_2; \Delta_L)| \\
\overset{(i)}{\leq}& |\langle \Delta_S, \Delta_L \rangle_{\mathcal{F}^*}| + \frac{c_2(\Delta_S + \Delta_L) + c_2(\Delta_S) + c_2(\Delta_L)}{2(M-1)} \\
=& |\langle \Delta_S, \Delta_L \rangle_{\mathcal{F}^*}| + \frac{\|\Delta_S\|_{\mathcal{F}^*}^2 + \|\Delta_L\|_{\mathcal{F}^*}^2 + \langle \Delta_S, \Delta_L \rangle_{\mathcal{F}^*}}{M-1} \\
\leq& \frac{M}{M-1} |\langle \Delta_S, \Delta_L \rangle_{\mathcal{F}^*}| + \frac{\|\Delta_S\|_{\mathcal{F}^*}^2 + \|\Delta_L\|_{\mathcal{F}^*}^2}{M-1} \\
\overset{(ii)}{\leq}& \frac{M\psi + 1}{M-1} (\|\Delta_S\|_{\mathcal{F}^*}^2 + \|\Delta_L\|_{\mathcal{F}^*}^2) \\
\overset{(iii)}{\leq}& \frac{M-2}{4(M-1)} \kappa_{\min}^* (\|\Delta_S\|_F^2 + \|\Delta_L\|_F^2) \\
\leq& \frac{\kappa_{\mathcal{L}}}{2} (\|\Delta_S\|_F^2 + \|\Delta_L\|_F^2),
\end{aligned}
$$

where in (i) we have apply (4.37) to bound the second-order Taylor error terms (note that the conditions on the error matrices also guarantees $\|\Delta_S + \Delta_L\|_{\mathcal{F}^*}^2 \leq \frac{1}{2M^2}$ due to Lemma 4.4). Inequality (ii) is due to Lemma 4.4. Inequality (iii) can be verified by the definitions of ψ and the RSC curvature parameter $\kappa_{\mathcal{L}}$. □

4.8.4 Proof of Corollary 4.1

Proof. Theorem 4.1 is a deterministic statement, however, the condition on the regularization parameters (4.11) and the error bound depend on the sample covariance matrix $\widehat{\Sigma}$ which is random. Note that the error bound directly follows from the deterministic error bound in Theorem 4.1 and the choices of regularization parameters as in Eq. (4.16). To prove Corollary 4.1, it only remains to verify that the condition (4.11) in Theorem 4.1 is guaranteed with high probability. More specifically, this requires bounding the deviation of the sample covariance matrix in terms of ℓ_∞ and

and spectral norms.

First we make use of the following lemma to characterize the element-wise deviation of the sample covariance matrix[3].

Lemma 4.5 (Ravikumar et al. (2011)). *For a p-dimensional Gaussian random vector with covariance matrix $\mathbf{\Sigma}^*$, the sample covariance matrix obtained from n samples $\widehat{\mathbf{\Sigma}}$ satisfies*

$$P\left\{|\widehat{\mathbf{\Sigma}}_{i,j} - \mathbf{\Sigma}^*_{i,j}| > \epsilon_1\right\} \leq 4\exp\left(-\frac{n\epsilon_1^2}{3200\overline{\sigma}^{*2}}\right), \tag{4.40}$$

for all $\epsilon_1 \in (0, 40\overline{\sigma})$, where $\overline{\sigma}^* := \max_{i=1,\ldots,p} \mathbf{\Sigma}^*_{i,i}$.

If the number of samples satisfies $n \geq 4\log p$, then by choosing $\frac{1}{2}\lambda \geq \epsilon_1 = 80C_1\overline{\sigma}^*\sqrt{\frac{\log p^2}{n}} \in (0, 40\overline{\sigma})$, where $C_1 > 1$ is an arbitrary constant, and applying the union bound we have

$$P\left\{\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_\infty \leq \frac{1}{2}\lambda\right\} \geq P\left\{\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_\infty \leq \epsilon_1\right\} \geq 1 - 4p^{-2(C_1-1)}.$$

Then the condition on λ is satisfied with high probability.

Next we consider the condition on the other regularization parameter μ, which requires bounding the deviation of the operation norm of the sample covariance matrix. The following lemma provides such a characterization.

Lemma 4.6 (Chandrasekaran et al. (2012), Lemma 3.9). *For a p-dimension Gaussian random vector with covariance matrix $\mathbf{\Sigma}^*$ and let $\rho^* = \|\mathbf{\Sigma}^*\|_2$. If the number of samples n be such that $n \geq \frac{64p\rho^{*2}}{\epsilon_2^2}$, then the sample covariance matrix $\widehat{\mathbf{\Sigma}}$ obtained*

[3]The original lemma applies to all sub-Gaussian variables, here we specialize to Gaussian random vectors.

from n samples satisfies

$$P\left\{\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_2 \geq \epsilon_2\right\} \leq 2\exp\left(-\frac{n\epsilon_2^2}{128{\rho^*}^2}\right), \tag{4.41}$$

for all $\epsilon_2 \in (0, 8\rho^*)$.

If $n \geq p$, then by choosing $\frac{1}{2}\mu \geq \epsilon_2 = 8C_2\rho^*\sqrt{\frac{p}{n}} \in (0, 8\rho^*)$, where $C_2 \geq 1$ is an arbitrary constant, we have

$$P\left\{\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_2 \leq \frac{1}{2}\mu\right\} \geq P\left\{\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_2 \leq \epsilon_2\right\} \geq 1 - 2\exp\left(-\frac{C_2^2 p}{2}\right).$$

Combining the above results we have verified the condition (4.11) in Theorem 4.1 holds with high probability, which concludes the proof. □

CHAPTER 5

Detecting Emerging Topics in Topic Models with Confidence

In this chapter, we focus on a specific Bayesian network, which is known in machine learning as a "topic model", namely the latent Dirichlet allocation (LDA) model. In particular, we consider the problem of detecting emerging topics in text document corpora based on the LDA model. A hypothesis testing framework for this detection problem is proposed, and as a surrogate for the standard generalized likelihood-ratio test, we consider and perform a simple test procedure called the surrogate Hausdorff test. We develop theory which shows that, in addition to significantly reduced computational cost, the proposed test procedure also has strong detection performance guarantees, such as the asymptotic consistency. Numerical experiments on both synthetic and real-world corpora validate and complement our theoretical results.

5.1 Introduction

Event detection is an important and challenging task which has many real-world applications, ranging from discovering breaking news from online articles, to identifying anomalous activities from social or behavioral data. For text document corpora, this task can be formulated as *topic detection*, where each topic captures an event, such as a piece of news or a trending meme on social media websites. In this paper, we

consider a hypothesis testing framework for this problem. This framework is built upon a specific type of topic model, namely the latent Dirichlet allocation (LDA) model.

The LDA model has been widely used in modeling collections of text documents since its first introduction to the computer science and applied statistics communities (Blei et al., 2003). Its empirical success has been observed in various areas and for diverse data sets (see, for example, Blei (2014, 2012) for recent surveys). The LDA model posits that each document (seen as a "bag of words") is sampled from a mixture of multinomial distributions over the vocabulary of words, where each mixture component is called a "topic". These topics are assumed to be shared by all documents in the corpus, while the mixing proportions are specific to each of the documents. A parallel and equally active line of work can be found in the genetics literature under the name *admixture model*, where the primary interest is in modeling the ancestral structure in the genotype data (see Pritchard et al. (2000) for an introduction). In this present paper we adopt the terminology used in the context of topic modeling for text documents. However, the theory and algorithms developed naturally apply to any field that uses LDA models.

The problem of topic detection in topic models is motivated by real-world applications that share the following general set-up. Suppose that a corpus of time stamped documents is collected from some source, *e.g.,* a publications database, over a period of time. At some later time another collection of documents is collected from the same source. We are interested in detection of novel topics in the newer collection that were not present in the older collection. These novel topics are called *emerging topics*. We note that the notion of topic can be generalized to any previously unseen *event*, in particular when the topic models are used for modeling non-text data, such as images.

Perhaps the most straightforward approach to incorporate these new topics into the

old topic model is to increase the number of topics when fitting the new document collection. However, without a clear characterization of the emerging event, it is difficult to specify how many additional topics are required. A standard practice in applying the LDA model is to over-fit the model with many more topics, and then remove the low-quality or duplicate ones. These redundant topics are often identified through either manual inspection of the top words, or some heuristic quality measures. Unfortunately, both empirical studies and recent theoretical analysis have shown that over-fitting degrades the learnability of the LDA topics (Tang et al., 2014; Nguyen, 2014). As a result, many topics obtained from such an over-fitting strategy will likely not capture the true topics.

An alternative approach to deal with the change of number of topics is to modify the standard LDA model. Some examples include Blei & Lafferty (2006); Wang & McCallum (2006); AlSumait et al. (2008); Wang et al. (2012). The shared idea is to couple multiple LDA models to capture the emergence or disappearance of topics. A more principled framework to handle the unknown and varying number of topics are nonparametric extensions of the LDA model, *e.g.,* the hierarchical Dirichlet process model (Teh et al., 2006) and its variants (Gao et al., 2011; Lin & Fisher, 2012; Chen et al., 2012). Unfortunately, most existing models and approaches lack theoretical guarantees, and are often computationally expensive due to the increased parameter space and model complexity.

In this paper, we consider a hypothesis testing framework for detecting emerging topics in the LDA model. In particular, we propose to estimate a LDA model from the new data set with only *one* additional topic on top of the existing ones. Then a test statistic is calculated by projecting the estimated new topic onto the convex hull of the old topics. This test statistic can be shown to be interpreted as a computationally simpler surrogate for the standard likelihood ratio. By leveraging the theory of empirical processes and some recently established results relating various forms of di-

vergence for LDA densities to the Hausdorff distance between the corresponding topic polytopes, we show that the proposed surrogate test guarantees the consistency of the original hypothesis testing problem, even when the true number of emerging topics is greater than one. In particular, the surrogate test statistic correctly converges to zero under the null hypothesis (*i.e.,* in the absence of new topics), and is bounded away from zero under the alternative hypothesis (*i.e.,* there occurs at least one new topic). We further provide the convergence rate of the surrogate test statistic and bounds on the hypothesis testing errors. The theory is validated through numerical experiments on both synthetic and real-world corpora.

The remainder of the paper is organized as follows. In Section 5.2 we describe the LDA model and the notion of topic polytope. A hypothesis testing framework for topic detection is introduced in Section 5.3 and our proposed surrogate test is described in Section 5.4. Our main results on the detection performance of the proposed surrogate test are stated in Section 5.5 and their proofs are included in Section 5.6. In Section 5.7 we provide numerical experimental results to validate our theory, and we conclude with future directions in Section 5.8.

Notation for Chapter 5. The Euclidean distance between a point θ and a set C is defined as $\operatorname{dist}(\theta, C) := \inf_{\mathbf{x}\in C} \|\theta - \mathbf{x}\|$. The Hausdorff distance between two convex bodies C and C' is defined as:

$$d_{\mathcal{H}}(C, C') = \max \left\{ \max_{\theta\in C} \operatorname{dist}(\theta, C'),\ \max_{\theta'\in C'} \operatorname{dist}(\theta', C) \right\}. \tag{5.1}$$

$B_d(\theta, R)$ denotes a closed d-dimensional Euclidean ball centered at point θ and has radius R. The notations $\operatorname{int} G$, $\operatorname{relint} G$, $\operatorname{extr} G$, $\operatorname{aff} G$, and $\operatorname{vol}_d G$ denote the interior, relative interior, the set of extreme points, the affine span, and the d-dimensional volume of set G. We define the dimension of a convex polytope to be the dimension of its affine hull. Δ^d denotes the d-dimensional probability simplex.

The following divergence measures for probability distributions are used: $K(p,q)$, $h(p,q)$, $V(p,q)$ denote, respectively, the Kullback-Leibler divergence, Hellinger distance and total variation distances between two densities p and q defined with respect to a measure on a common space: $K(p,q) = \int p \log(p/q)$, $h^2(p,q) = \frac{1}{2}\int(\sqrt{p} - \sqrt{q})^2$ and $V(p,q) = \frac{1}{2}\int |p - q|$.

5.2 The LDA model and topic polytope

Latent Dirichlet Allocation (LDA) models were first introduced in Blei et al. (2003) and Pritchard et al. (2000). We briefly review the model assumptions, and give an equivalent representation that will be used to derive the surrogate test statistic. Interested readers are referred to the above papers for more details.

Available are a collection of m documents. The i^{th} document is represented as a "bag of words": $\mathcal{S}^i_{[n]} := (X_{ij})_{j=1}^n$, where $X_{ij} \in \{0, 1, \ldots, d\}$ is a word in the vocabulary, and n is the number of words in the document. Without loss of generality, we have assumed all documents have the same number of words. The LDA model assumes that the words in each document are independent and identically distributed (*i.i.d.*) samples from a document-specific mixture of K multinomial distributions over the vocabulary, where each multinomial distribution is called a "topic" and is parameterized by a vector $\phi_k \in \Delta^d, k = 1, \ldots, K$. Here Δ^d is the d-dimensional probability simplex. In practice, it is reasonable to assume $K \ll d$ (for example, Griffiths & Steyvers (2004) found that around 300 topics are optimal for modeling a corpus of $28,154$ abstract documents with a vocabulary of $20,551$ words). The collection of m documents $\mathcal{S}^{[m]}_{[n]} := (\mathcal{S}^i_{[n]})_{i=1}^m$ is generated independently in the same manner, conditioning on the same set of K topic vectors $\Phi = (\phi_1, \ldots, \phi_K)$. One of the primary goals in LDA modeling is inference of the topic parameters Φ on the basis of the sampled $m \times n$ words $\mathcal{S}^{[m]}_{[n]}$.

As first introduced in Nguyen (2014), instead of studying the individual topic vectors of the LDA model, we propose to consider their convex hull:

$$G = \text{conv}(\phi_1, \ldots, \phi_K), \tag{5.2}$$

which we shall refer to as the *topic polytope*. By studying this topic polytope, we effectively remove the "label-switching" ambiguity between the true and estimated topics and also handle the difficulty of estimating topics that are inside the polytope (*i.e.,* topics that can be expressed as convex combination of others). The critical role of the topic polytope G can be seen from an equivalent representation of the LDA model described as follows. For each document $i = 1, \ldots, m$, we associate it with a random vector in the topic polytope $\eta_i \in G$, parameterized by $\eta_i = \theta_{i,1}\phi_1 + \ldots + \theta_{i,K}\phi_K$, where the random vector $\theta_i = (\theta_{i,1}, \ldots, \theta_{i,K}) \in \Delta^{K-1}$ contains the topic mixing proportions. Conditioning on η_i, the words in the document $\mathcal{S}^i_{[n]} = (X_{ij})_{j=1}^n$ are i.i.d. draws from the multinomial distribution Multi(η_i) specified by η_i, *i.e.,* $P(X_{ij} = l \mid \eta_i) = \eta_{il}$ for $l = 0, \ldots, d$. As in the standard LDA model, we assume the topic proportion vector θ follows a Dirichlet distribution Dir(α), where $\alpha = (\alpha_1, \ldots, \alpha_K)$ are the concentration parameters. Then the random vector η follows an induced distribution over topic polytope G, denoted as $P_{\eta|G}$.

The distribution of a document $\mathcal{S}^i_{[n]}$, denoted by $P_{\mathcal{S}_{[n]}|G}$, is obtained by integrating out the random vector η over the topic polytope G, yielding the following density with respect to counting measure:

$$p_{\mathcal{S}_{[n]}|G}(\mathcal{S}^i_{[n]}) = \int_G \prod_{j=1}^{n} \prod_{l=0}^{d} \eta_{il}^{\mathbb{I}(X_{ij}=l)} dP_{\eta|G}(\eta_i). \tag{5.3}$$

Accordingly, the joint distribution of the full data set $\mathcal{S}^{[m]}_{[n]}$, denoted as $P^m_{\mathcal{S}_{[n]}|G}$, is the

product distribution of all single-document distributions:

$$P^m_{\mathcal{S}_{[n]}|G}(\mathcal{S}^{[m]}_{[n]}) := \prod_{i=1}^{m} P_{\mathcal{S}_{[n]}|G}(\mathcal{S}^i_{[n]}). \tag{5.4}$$

Note that our formulation of the LDA model focuses on the topic-level and document-level characteristics, since only the marginal distribution (5.4) is needed for the proposed hypothesis testing with respect to the topic polytope. Therefore we do not need to introduce the latent word-topic assignment variables as in Blei et al. (2003). Instead, they have been marginalized out in our representation.

The topic parameters Φ can be estimated by either empirical Bayes approaches or in a full Bayesian framework. In the full Bayesian framework, the topic parameters are assumed random and endowed with a prior distribution. Then their posterior distributions given the observed documents can be obtained through approximate inference techniques. In this paper, we take the empirical Bayes approach (as in Blei et al. (2003)), where a point estimate of the topic parameters Φ (and hence their induced topic polytope G) is obtained by maximizing the marginal likelihood of the samples (5.4). This perspective allows us to use some established empirical process theory to derive the concentration behavior of the topic parameter estimates, which complements recent theory in the Bayesian framework (Nguyen, 2014).

5.3 Hypothesis testing framework for detecting emerging topics

Suppose there is an available LDA model learned from historical data. Our goal is to detect the occurrence of emerging topics. Alternatively, one can also consider this task as an anomaly detection problem, where the existing LDA model corresponds to the nominal model, and the emerging topics capture certain anomalous activities.

We formulate the topic detection problem as a composite hypothesis test. The null hypothesis assumes that the new samples are generated from the existing LDA model, while the alternative hypothesis assumes that they are generated from a new LDA model which contains emerging topics in addition to the existing ones. Let the LDA model in the null hypothesis has K topics, denoted as $\Phi = \{\phi_1, \ldots, \phi_K\}$, and let their convex hull be $G_0 = \text{conv}(\Phi)$ (called the *null topic polytope*). Given a text document corpus $\mathcal{S}_{[n]}^{[m]}$, the emerging topic detection problem can be formulated as the following composite hypothesis test (HT-q):

$$\textbf{(HT-}q\textbf{)} \begin{cases} H_0 : \mathcal{S}_{[n]}^{[m]} \sim P_{\mathcal{S}_{[n]}|G_0}^m \\ H_1 : \mathcal{S}_{[n]}^{[m]} \sim P_{\mathcal{S}_{[n]}|G_q}^m, \end{cases} \tag{5.5}$$

where $G_q = \text{conv}(\Phi, \phi_{K+1}, \ldots, \phi_{K+q})$ is the topic polytope of the LDA model under the alternative hypothesis. The set of the extreme points of G_q includes the existing topics Φ and q new topics ($q > 0$), each of which captures an event.

A standard test procedure for composite hypothesis testing is the generalized likelihood ratio test (GLRT), in which the maximum likelihood (ML) estimation of the alternative model is followed by a log-likelihood ratio test (LRT). For our specific test HT-q, the ML estimation step involves the inference of the unknown topic parameters $\phi_{K+1}, \ldots, \phi_{K+q}$ (note that q is also unknown) or, equivalently, the topic polytope G_q.

Define $\overline{\mathcal{G}}^q$ as the set of all polytopes with *at most* $(K + q)$ extreme points that include Φ:

$$\overline{\mathcal{G}}^q = \{G \mid \Phi \subset \text{extr } G, |\text{extr } G| \leq K + q\}. \tag{5.6}$$

Note that $G_0 \in \overline{\mathcal{G}}^q$ by definition. The ML polytope under the alternative hypothesis

is

$$\widehat{G}_q = \arg\max_{G \in \overline{\mathcal{G}}^q} \; P^m_{\mathcal{S}_{[n]}|G}(\mathcal{S}^{[m]}_{[n]}). \tag{5.7}$$

Equivalently, this ML estimation is also with respect to the topic parameters, *i.e.,*

$$\widehat{G}_q = \mathrm{conv}(\Phi, \widehat{\phi}_{K+1}, \ldots, \widehat{\phi}_{K+q}), \tag{5.8}$$

where $\{\widehat{\phi}_{K+j}, j = 1, \ldots, q\}$ denote the ML estimates for the topic vectors.

The standard log-likelihood ratio test statistic can be formulated as

$$\sum_{i=1}^{m} \log \frac{P_{\mathcal{S}_{[n]}|\widehat{G}_q}}{P_{\mathcal{S}_{[n]}|G_0}} \underset{H_0}{\overset{H_1}{\gtrless}} \tau_{m,n}, \tag{5.9}$$

where the decision threshold $\tau_{m,n}$ is often a function of the sample sizes m and n chosen by the user.

Although conceptually straightforward, performing the above GLRT procedure involves several challenges. The first difficulty is that the ML estimation under the alternative hypothesis involves LDA model selection for all possible $q > 0$, which is computationally expensive and statistically vulnerable to overfitting. Moreover, evaluating the exact likelihood $P_{\mathcal{S}_{[n]}|G}$ for a given topic polytope G, or the log-likelihood ratio in Eq. (5.9), is in general computationally intractable due to the integral (5.3). Sampling-based algorithms or variational approximation are often used in practice, but the results are mixed (see, *e.g.,* Wallach et al. (2009); Taddy (2012); Scott & Baldridge (2013); Foulds & Smyth (2014) for some recent discussions). Finally, choosing an appropriate threshold function $\tau_{m,n}$ that guarantees a certain level of detection performance (*e.g.,* a given false positive rate) is a largely unexplored problem in the LDA model selection literature.

5.4 Surrogate Hausdorff test

We propose a surrogate testing procedure to address the above challenges. First we introduce a new hypothesis testing problem (HT-1) in which the generating LDA model in the alternative hypothesis has only $(K+1)$ topics:

$$\textbf{(HT-1)} \begin{cases} H_0 : \mathcal{S}_{[n]}^{[m]} \sim P^m_{\mathcal{S}_{[n]}|G_0} \\ H_1 : \mathcal{S}_{[n]}^{[m]} \sim P^m_{\mathcal{S}_{[n]}|G_1}, \end{cases} \tag{5.10}$$

where $G_1 = \text{conv}(\Phi, \phi_{K+1})$, and $\phi_{K+1} \in \Delta^d$ is a new topic vector. Under this setting, the space of alternative hypotheses is substantially reduced, and hence ML estimation of the single-topic parameters is computationally more tractable than it is in the HT-q problem (5.5).

Our proposed test described below is motivated by the idea of using this HT-1 problem as a surrogate for the HT-q problem. Let $\overline{\mathcal{G}}^1$ denote the set of all polytopes with *at most* $(K+1)$ extreme points that include Φ:

$$\overline{\mathcal{G}}^1 = \{G \mid \Phi \subset \text{extr } G, |\text{extr } G| \leq K+1\}. \tag{5.11}$$

Then the ML estimate of the $(K+1)$-polytope, denoted as $\widehat{G}_1$ and referred to as the *surrogate ML polytope*, is defined as:

$$\widehat{G}_1 = \underset{G \in \overline{\mathcal{G}}^1}{\arg\max}\ P^m_{\mathcal{S}_{[n]}|G}(\mathcal{S}_{[n]}^{[m]}). \tag{5.12}$$

Its associated ML $(K+1)^{th}$ topic vector can be represented as:

$$\widehat{\phi}_{K+1} = \underset{\phi_{K+1} \in \Delta^d}{\arg\max}\ P^m_{\mathcal{S}_{[n]}|G}(\mathcal{S}_{[n]}^{[m]}),\ G = \text{conv}(\Phi, \phi_{K+1}). \tag{5.13}$$

As a surrogate for the likelihood ratio (5.9), we propose to use $d_{\mathcal{H}}(\widehat{G}_1, G_0)$, the

Hausdorff distance between the surrogate ML polytope $\widehat{G}_1$ and the null polytope G_0 as our test statistic. The decision rule is:

$$d_{\mathcal{H}}(\widehat{G}_1, G_0) \underset{H_0}{\overset{H_1}{\gtrless}} \rho, \tag{5.14}$$

where ρ is a decision threshold. We shall refer to this proposed estimation-detection procedure as the *surrogate Hausdorff test*.

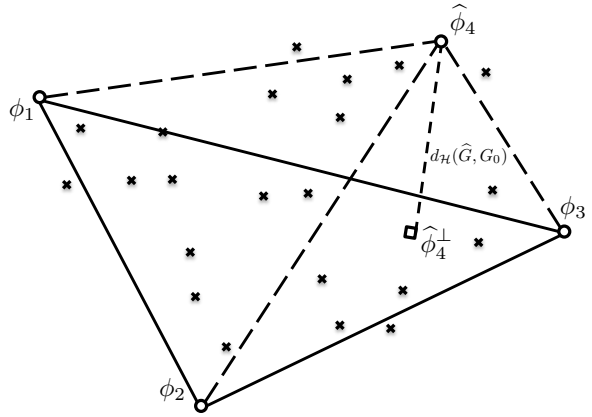


Figure 5.1: Illustration of the Hausdorff distance between null and alternative polytopes. All points denote vectors in Δ^d, the d-dimensional word probability simplex (not shown). "$\times$" denotes the empirical word frequency vector corresponding to a document. $\widehat{\phi}_4^{\perp}$ is the projection of $\widehat{\phi}_4$ onto G_0. When $\widehat{\phi}_4 \notin \text{aff } G_0$, the Hausdorff distance between $\widehat{G} = \text{conv}(\phi_1, \phi_2, \phi_3, \widehat{\phi}_4)$ and $G_0 = \text{conv}(\phi_1, \phi_2, \phi_3)$ is the Euclidean distance between $\widehat{\phi}_4$ and $\widehat{\phi}_4^{\perp}$.

Computation of the Hausdorff distance. When the extreme points of $\widehat{G}_1$ are composed of the extreme points of G_0 and $\widehat{\phi}_{K+1}$ (this holds almost surely under our assumed setting $K \ll d$), the Hausdorff distance between G_0 and $\widehat{G}_1$ coincides with the following projected Euclidean distance (see Figure 5.1 for an illustration)

$$d_{\mathcal{H}}(\widehat{G}_1, G_0) = d(\widehat{\phi}_{K+1}, G_0) = \min_{\mathbf{1}^T\mathbf{w}=1, \mathbf{w}\succeq\mathbf{0}} \|\widehat{\phi}_{K+1} - \sum_{k=1}^{K} w_k \phi_k\|, \tag{5.15}$$

which can be formulated as a quadratic programming optimization (by equivalently

optimizing the squared Euclidean distance) with respect to a low-dimensional Δ^{K-1}-constrained vector $\mathbf{w}$. Therefore the Hausdorff distance can be calculated efficiently by generic convex optimization techniques (Boyd & Vandenberghe, 2009).

5.5 Performance guarantees of surrogate Hausdorff test

In this section we state the theoretical results on performance guarantees for the proposed surrogate Hausdorff test. The proofs of the main theorems are given in Section 5.6.

Asymptotic setting. The data samples available are a collection of m documents, $\mathcal{S}_{[n]}^{[m]} = (\mathcal{S}_{[n]}^{i})_{i=1}^{m}$. Our main focus is to establish performance guarantees for the proposed surrogate Hausdorff test as the total number of words (or tokens) $m \times n$ goes to infinity. In particular, we consider the asymptotic setting where both m, the number of documents, and n, the number of words in each document, are finite and allowed to increase to infinity while the constraint $\log n = o(m)$ is satisfied. Note that our asymptotic setting requires a mild increasing rate of n, as contrasted to the setting of Nguyen (2014), in which both m and n increase to infinity with the constraint $\log \log m \leq \log n = o(m)$.

Assumptions. Recall from the model description that each document is distributed according to the LDA density, *i.e.,* $\mathcal{S}_{[n]}^{i} \sim P_{\mathcal{S}_{[n]}|G}$. The (latent) topic mixing proportions vector θ is endowed with a Dirichlet prior distribution $\mathrm{Dir}(\alpha)$. We make the following assumptions.

(A0) The number of topics is much smaller than the number of words in the vocabulary, *i.e.,* $K \ll d$.

(A1) The true LDA models under the null and alternative hypotheses have identical Dirichlet concentration parameters $\alpha_1, \ldots, \alpha_K$. The alternative hypothesis has additional concentration parameters $\alpha_{K+1}, \ldots, \alpha_{K+q}$.

(A2) For the LDA models under both hypotheses, $\alpha_j \in (0, 1]$ for all $j = 1, \ldots, K+q$.

(A3) The topic polytope G_0 is fixed and not degenerate, *i.e.,* for some $r, R > 0$, $\mathbf{v} \in \Delta^d$, G_0 contains the spherical ball $B_K(\mathbf{v}, r)$ and is contained in $B_K(\mathbf{v}, R)$.

(A4) Topic vectors $\phi_j = (\phi_{j,0}, \ldots, \phi_{j,d})$, $j = 1, \ldots, K+q$, are bounded away from the boundary of Δ^d, *i.e.,* $\min_{l=0,\ldots,d} \phi_{j,l} > c_0$ for some $c_0 > 0$.

(A5) The surrogate ML polytope $\widehat{G}^1$ under the alternative hypothesis of HT-q converges to a $(K+1)$-polytope $G_1^* \in \overline{\mathcal{G}}^1$ almost surely. In addition, the Hausdorff distance between G_1^* and G_0 is bounded away from zero almost surely[1].

Under these assumptions, the following theorem states the consistency of the proposed surrogate Hausdorff test.

Theorem 5.1. *Let φ be the test function associated with the proposed surrogate Hausdorff test $\varphi := \mathbb{I}\{d_{\mathcal{H}}(\widehat{G}_1, G_0) > \rho\}$, where $\widehat{G}_1$ is the surrogate ML polytope and ρ is a decision threshold. Under Assumptions (A0 - A5), there exists some $\rho > 0$, such that as $m \to \infty$ and $n \to \infty$ such that $\log n = o(m)$, the following holds:*

$$P^m_{\mathcal{S}_{[n]}|G_0}\ [\varphi] \to 0, \tag{5.16}$$

$$P^m_{\mathcal{S}_{[n]}|G'}\ [1 - \varphi] \to 0, \tag{5.17}$$

where $P^m_{\mathcal{S}_{[n]}|G}\ [\varphi]$ denotes the expectation of test φ under distribution $P^m_{\mathcal{S}_{[n]}|G}$, and $G' \in \overline{\mathcal{G}}^q$, the set of all $(K+q)$-polytopes with any finite integer $q > 0$.

[1]We conjecture that this statement can be proven to be true. Due to technical reason, we present it as an assumption. See Section 5.6.2 for more details.

Theorem 5.1 shows that the the proposed test based on Hausdorff distance between the surrogate ML polytope $\widehat{G}^1$ and the null polytope G_0 is asymptotically consistent for the hypothesis testing problem HT-q, in particular both the false alarm (Type-I) and false negative (Type-II) rates approach zero with sufficient samples. The proof of Theorem 5.1 is given in Section 5.6.2.

The convergence of false alarm rate with respect to zero relies on the convergence behavior of the surrogate ML polytope $\widehat{G}_1$, which is interesting in its own right. In particular, when the true polytope G^* lies in $\overline{\mathcal{G}}^1$, by leveraging the theory of empirical processes we have the following theorem on the convergence of $\widehat{G}_1$ with respect to G^*.

Theorem 5.2. *Assume Assumptions (A0 - A4) hold and the true topic polytope $G^* \in \overline{\mathcal{G}}^1$. Let $\widehat{G}_1$ be the surrogate ML polytope estimated from samples $\mathcal{S}_{[n]}^{[m]}$. As $m \to \infty$ and $n \to \infty$ such that $\log n = o(m)$, for some sufficiently large constant C independent of m and n, a universal constant c, and for all $\delta \geq \delta_{m,n}$, where $\delta_{m,n}$ is the vanishing sequence defined as*

$$\delta_{m,n} = \left[\frac{\log m}{m} + \frac{\log n}{n} + \frac{\log n}{m} \right]^{\frac{1}{2}}, \tag{5.18}$$

we have

$$P^m_{\mathcal{S}_{[n]}|G^*} \left(d_{\mathcal{H}}(\widehat{G}_1, G^*) > C\delta^{\frac{1}{\gamma}} \right) \leq c \exp\left(-m\delta^2/c^2 \right), \tag{5.19}$$

where the exponent $\gamma = 1$ when $G^* = G_0$, and $\gamma = K = |extr\, G_0|$ when $G^* \in \overline{\mathcal{G}}^1 \backslash \{G_0\}$.

We make the following remarks:

1. It is interesting to note that the convergence rate of the Hausdorff distance $\delta_{m,n}$ parallels the *posterior contraction rate* of the topic polytope recently established by Nguyen (2014) in a Bayesian setting (cf. Theorem 2 therein), which is validated

by empirical results (Tang et al., 2014). However, note that the asymptotic setting of our results is different (see the remark in the beginning of Section 5.5).

2. It can be shown that the quantity $d_{\mathcal{H}}(\widehat{G}_1, G^*)^\gamma$ serves as a lower bound for $h(P_{\mathcal{S}_{[n]}|\widehat{G}_1}, P_{\mathcal{S}_{[n]}|G^*})$, the Hellinger divergence between densities $P_{\mathcal{S}_{[n]}|\widehat{G}_1}$ and $P_{\mathcal{S}_{[n]}|G^*}$ (cf. Lemma 5.1). Therefore Eq. (5.19) implies that, when $G^* \in \overline{\mathcal{G}}^1$, the detection error exponent for the proposed surrogate Hausdorff test is asymptotically on the order of the squared Hellinger divergence $h^2(P_{\mathcal{S}_{[n]}|\widehat{G}_1}, P_{\mathcal{S}_{[n]}|G^*})$ or, equivalently, the Kullback-Leibler divergence $K(P_{\mathcal{S}_{[n]}|\widehat{G}_1}, P_{\mathcal{S}_{[n]}|G^*})$.

3. When the true polytope $G^* = G_0$, *i.e.,* under the null hypothesis of the HT-q problem, Eq. (5.19) ensures the convergence of $\widehat{G}^1$ with respect to G_0. Moreover, it provides guidance for choosing the decision threshold in the surrogate Hausdorff test (Eq. (5.14)) and an upper bound on the false alarm rate (Type-I error) for the HT-q problem.

4. Under the alternative hypothesis of the HT-q problem, *i.e.,* when $G^* \in \overline{\mathcal{G}}^q \backslash \overline{\mathcal{G}}^1$, using our surrogate test, the ML polytope $\widehat{G}^1 \in \overline{\mathcal{G}}^1$ will not converge to the true model G^*. Therefore Eq. (5.19) does not generally apply due to the model mis-specification. However, under a slightly more restrictive, but still realistic, setting, the results of Theorem 5.2 apply to both the null and alternative hypotheses, as shown below.

Performance guarantees for HT-1. We revisit the HT-1 problem (5.10), in which the LDA models under the alternative hypothesis are assumed to have exactly $(K+1)$ topics (*i.e.,* the alternative topic polytope has $(K+1)$ extreme points). In this scenario, we have the following theorem.

Theorem 5.3. *Assume Assumptions (A0 - A4) hold and the true topic polytope* $G^* \in \overline{\mathcal{G}}^1$. *Let* $\widehat{G}_1$ *be the surrogate ML polytope estimated from samples* $\mathcal{S}_{[n]}^{[m]}$. *As*

$m \to \infty$ and $n \to \infty$ such that $\log n = o(m)$, for some sufficiently large constants C and C' independent of m and n, some universal constants c and c', and for all $\delta \geq \delta_{m,n}$, where $\delta_{m,n}$ is a vanishing sequence defined in Eq. (5.18)*, we have the following relations:*

$$P^m_{\mathcal{S}_{[n]}|G_0}\left(d_{\mathcal{H}}(\widehat{G}_1, G_0) > C\delta\right) \leq c\exp\left(-m\delta^2/c^2\right) \tag{5.20}$$

$$\sup_{G' \in \overline{\mathcal{G}}^1} P^m_{\mathcal{S}_{[n]}|G'}\left(d_{\mathcal{H}}(\widehat{G}_1, G') < C'\delta^{\frac{1}{K}}\right) \leq c'\exp\left(-m\delta^2/c'\right). \tag{5.21}$$

Furthermore, there exists some constant $\rho > 0$, such that the following guarantees hold for the Hausdorff test $\varphi := \mathbb{I}\{d_{\mathcal{H}}(\widehat{G}_1, G_0) > \rho\}$:

$$P^m_{\mathcal{S}_{[n]}|G_0}\left[\varphi\right] \leq c\exp\left(-m\delta^2/c^2\right), \tag{5.22}$$

$$\sup_{G' \in \overline{\mathcal{G}}^1} P^m_{\mathcal{S}_{[n]}|G'}\left[1-\varphi\right] \leq c'\exp\left(-m\delta^2/c'\right), \tag{5.23}$$

where $P^m_{\mathcal{S}_{[n]}|G}\left[\varphi\right]$ denotes the expectation of test φ under distribution $P^m_{\mathcal{S}_{[n]}|G}$.

Theorem 5.3 sharpens the consistency results in Theorem 5.1 by providing exponential bounds for both the Type-I and Type-II errors of the HT-1 problem. These results suggest that, if the emerging event(s) can be well captured by a single topic, which might be reasonable in practice, our theory provides strong guarantees for the proposed surrogate Hausdorff test.

5.6 Proofs of main theorems

In this section, we provide proofs of the main theorems. First we introduce the following lemma, which is derived from some results established in Nguyen (2014). This key lemma links the Hellinger divergence h between LDA densities and the Hausdorff distance $d_{\mathcal{H}}$ between the corresponding topic polytopes. For completeness, its proof is given in Section 5.9.

Lemma 5.1. *Let G, G' be polytopes in $\overline{\mathcal{G}}^1$. Assume that $d_{\mathcal{H}}(G, G') > c\sqrt{\log n/n}$ for some sufficiently large constant c, then there is a constant $C_1 > 0$ independent of G, G' such that*

$$h(p_G, p_{G'}) \geq C_1 d_{\mathcal{H}}(G, G')^{\gamma}, \tag{5.24}$$

where the exponent $\gamma = 1$ when $G = G_0, G' \in \overline{\mathcal{G}}^1 \backslash \{G_0\}$, and $\gamma = K = |extr\ G_0|$ when $G, G' \in \overline{\mathcal{G}}^1 \backslash \{G_0\}$.

Next we prove Theorem 5.2, which serves as a key component for the derivations of the other main theorems.

5.6.1 Proof of Theorem 5.2 (convergence of surrogate ML polytope)

Our proof leverages a result in van de Geer (2000) on the convergence rate of the ML estimator (in terms of Hellinger metric) for a general class of densities using the concentration behavior of empirical processes. We specialize this general result to the proposed hypothesis testing problem and derive a convergence rate of the ML topic polytope, using the relation in Lemma 5.1.

Notation. We first introduce some necessary definitions and notation. The LDA densities under consideration are indexed by topic polytopes in $\overline{\mathcal{G}}^1$. Let p_G be a shorthand notation for the density of $P_{\mathcal{S}_{[n]}|G}$ for some topic polytope G. Denote the "average density"[2] between a topic polytope G and the null polytope G_0 as $\overline{p}_G := (p_G + p_{G_0})/2$. Define

$$\overline{\mathcal{P}}^{1/2}(\delta) = \left\{ \overline{p}_G^{1/2} : h(\overline{p}_G, p_{G_0}) \leq \delta \right\} \tag{5.25}$$

[2]The use of the averaged densities ensures the log-likelihood $\log \overline{p}_G / p_{G_0}$ is bounded from below, which allows for the convenience of using Hellinger metric (van de Geer, 2000).

as the local Hellinger ball around p_{G_0} intersected with the space of square-root averaged densities $\overline{\mathcal{P}}^{1/2} := \left\{\overline{p}_G^{1/2}, \forall G\right\}$. It will be shown that the convergence rate of the maximum likelihood LDA density depends crucially on the richness of this set, measured by its *entropy with bracketing*. We briefly introduce this entropy notion, more detailed can be found in van der Vaart & Wellner (1996); van de Geer (2000). For any $u > 0$ and a function class $\mathcal{F}$, let $N_B(u, \mathcal{F})$ be the smallest value of N for which there exist pairs of functions (called *brackets*) $\{[f_j^L, f_j^U]\}_{j=1}^N$ such that $\|f_j^U - f_j^L\| \leq u$, and for each $f \in \mathcal{F}$ there exists a j such that $f_j^L \leq f \leq f_j^U$. Then $H_B(u, \mathcal{F}) = \log N_B(u, \mathcal{F})$ is defined as the entropy with bracketing of $\mathcal{F}$ (with respect to a certain metric).

For small δ, $H_B(u, \overline{\mathcal{P}}^{1/2}(\delta))$ denotes the entropy with bracketing of the defined set $\overline{\mathcal{P}}^{1/2}(\delta)$. Define an entropy integral as

$$J_B(\delta, \overline{\mathcal{P}}^{1/2}(\delta)) = \int_{\delta^2/c'}^{\delta} H_B^{1/2}(u, \overline{\mathcal{P}}^{1/2}(\delta))du \vee \delta, \tag{5.26}$$

where c' is an absolute constant (*e.g.,* 2^{13} is used in van de Geer (2000)). Define $\Psi(\delta) \geq J_B(\delta, \overline{\mathcal{P}}^{1/2}(\delta))$ as any function such that $\Psi(\delta)/\delta^2$ is non-increasing in δ.

Proof of Theorem 5.2. Let G^* be the true topic polytope and its associated distribution which generates the data samples be P_{G^*}. Recall the surrogate ML polytope estimate $\widehat{G}_1$ and its associated density $P_{\widehat{G}_1}$. The following proposition, which is a version of Theorem 7.4 in (van de Geer, 2000), provides a convergence rate of $p_{\widehat{G}_1}$ with respect to p_{G^*} in terms of Hellinger divergence (note that we slightly abuse the notation $\delta_{m,n}$ which is different to that defined before).

Proposition 5.1. *Assume the following holds for a vanishing sequence $\delta_{m,n}$ (as a*

function of m and n) and a universal constant c:

$$\sqrt{m}\delta_{m,n}^2 \geq c\Psi(\delta_{m,n}). \tag{5.27}$$

Then we have for all $\delta \geq \delta_{m,n}$,

$$P_{\mathcal{S}_{[n]}|G^*}\left(h(p_{\widehat{G}_1}, p_{G^*}) > \delta\right) \leq c\exp\left(-m\delta^2/c^2\right). \tag{5.28}$$

Under the null hypothesis, the convergence limit of $\widehat{G}^1$, denoted as G_0^*, is the null polytope G_0 (see Figure 5.2). Proposition 5.1 implies that the Hellinger divergence between the density parameterized by the surrogate ML polytope and the true density converges to zero at the rate of $O_P(\delta_{m,n})$. Due to Lemma 5.1 (for $G = G_0$), the Hausdorff distance between the ML and the true polytopes converges to zero at the same rate as the Hellinger divergence, also $O_P(\delta_{m,n})$. Therefore it remains to find a sequence $\delta_{m,n}$ that satisfies the entropy integral condition (5.27) to complete the proof of Theorem 5.2.

Define $\Phi(\delta)$ as the δ-parallel of G_0 (which is a d-dimensional body in Δ^d), *i.e.,*

$$\Phi(\delta) = \left\{\phi \in \Delta^d \backslash G_0 \mid \text{dist}(\phi, G_0) \leq \delta\right\}. \tag{5.29}$$

By definition, the set $\Phi(\delta)$ includes all $(K+1)^{th}$ topic vectors of polytopes in $\overline{\mathcal{G}}^1$ which are at most δ away from G_0 in terms of the Hausdorff distance. The following lemma establishes an upper bound of the entropy number with bracketing of the set $\overline{\mathcal{P}}^{1/2}(\delta)$ (defined in Eq. (5.25)) in terms of the (standard) entropy of the set $\Phi(\delta)$ with the Euclidean metric. The proof of Lemma 5.2 is given in Section 5.10.

Lemma 5.2. *For small $\delta, \epsilon > 0$, we have the following relation between the entropy*

with bracketing of $\overline{\mathcal{P}}^{1/2}(\delta)$ *and the entropy of* $\Phi(\delta)$ *(with Euclidean metric).*

$$H_B\left(\sqrt{\frac{n\overline{\alpha}_{K+1}\epsilon}{2c_0}}, \overline{\mathcal{P}}^{1/2}(\delta)\right) \leq H\left(\epsilon, \Phi(\delta)\right) = O\left(\frac{d}{K}\log\left(\frac{\delta}{\epsilon}\right)\right), \tag{5.30}$$

where $\overline{\alpha}_{K+1} := \alpha_{K+1}/\sum_{j=1}^{K+1}\alpha_j$ *is a scaled Dirichlet concentration parameter for the topic proportions, and* $c_0 > 0$ *is a constant.*

With this result, we could further upper bound the entropy integral defined in (5.26) as follows:

$$\begin{aligned}
J_B(\delta, \overline{\mathcal{P}}^{1/2}(\delta)) &\leq \int_{\delta^2/c'}^{\delta} H^{1/2}\left(2c_0\overline{\alpha}_{K+1}^{-1}n^{-1}u^2, \Phi(\delta)\right) du \vee \delta \\
&\lesssim \left(\int_{\delta^2/c'}^{\delta} \sqrt{\log\left(\delta n u^{-2}\right)} du\right) \vee \delta \\
&\leq \left(\int_{\delta^2/c'}^{\delta} \sqrt{\log\left(\delta n (\delta^2/c')^{-2}\right)} du\right) \vee \delta \\
&\lesssim (\delta - \delta^2/c')\sqrt{\log\left(\delta^{-3}n\right)} \vee \delta \\
&\lesssim \delta\sqrt{\log\left(\delta^{-3}n\right)} := \Psi(\delta),
\end{aligned}$$

where we have defined a proper $\Psi(\delta)$ (up to a constant). One can easily verify that Eq. (5.27) is satisfied if δ is bounded below by a sufficiently large multiple of $\left[(\log n/m)^{1/2} + (\log m/m)^{1/2}\right]$. Therefore the conditions in Proposition 5.1 hold. Combining this result with the condition in Lemma 5.1 yields the choice of $\delta_{m,n}$ (Eq. (5.18)). Theorem 5.2 follows immediately from Proposition 5.1 and Lemma 5.1 (for $G = G_0$). □

5.6.2 Proof of Theorem 5.1

We show that, under our specified asymptotic setting, a proper decision threshold ρ exists which correctly distinguishes the null and alternative hypotheses. The two key components of our proof are: (1) Under the null hypothesis, the Hausdorff distance

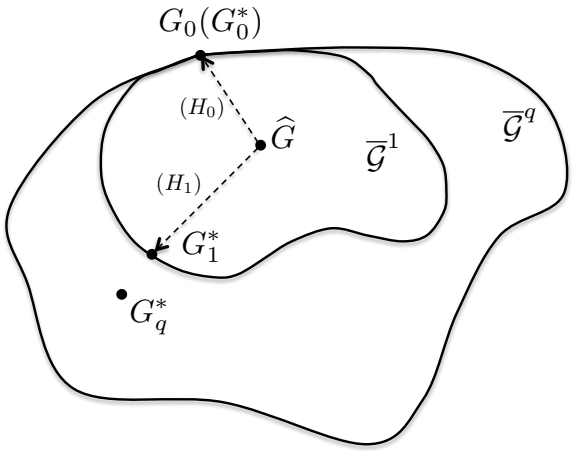


Figure 5.2: Illustration of the convergence of surrogate ML polytope estimate.

between the limiting surrogate ML polytope (denoted as G_0^*) and the null polytope G_0 converges to zero; (2) Under the alternative hypothesis, the Hausdorff distance between the limiting surrogate ML polytope (denoted as G_1^*) and the null polytope G_0 is bounded from zero. Then a threshold ρ can be chosen to ensure the consistency of the surrogate Hausdorff test. The convergence behavior of the surrogate ML estimate is illustrated in Figure 5.2.

Proof of Theorem 5.1. In Theorem 5.2, we have shown that under the null hypothesis, the surrogate ML estimate converges to G_0 at a rate described in Eq. (5.18)

$$d_{\mathcal{H}}(\widehat{G}^1, G_0) = d_{\mathcal{H}}(\widehat{G}^1, G^*) = O_P(\delta_{m,n}) \to 0, \quad \text{as } m, n \to \infty. \tag{5.31}$$

Therefore there exists some constant $C_1 > 0$, such that for all threshold $\rho > C_1\delta_{m,n}$, we have

$$P^m_{\mathcal{S}_{[n]}|G_0}\left(d_{\mathcal{H}}(\widehat{G}^1, G_0) > \rho\right) \to 0, \quad \text{as } m, n \to \infty, \tag{5.32}$$

which guarantees the convergence of false alarm rate with respect to zero under the null hypothesis. It only remains to prove that this chosen threshold ρ also leads to

the convergence of false negative rate under the alternative hypothesis.

Let the true polytope under the alternative hypothesis be G_q^*. Generally, G_q^* has more than $(K+1)$ extreme points, *i.e.,* $G_q^* \in \overline{\mathcal{G}}^q \backslash \overline{\mathcal{G}}^1$. Therefore, the surrogate ML polytope $\widehat{G}^1$ and its limit G_1^* will not converge to G_q^*. This results in a model misspecification in the proposed surrogate estimation procedure (see Figure 5.2). It is well known that, maximum likelihood estimator under model misspecification (also known as the quasi-maximum likelihood estimator, or QMLE) converges to a well-defined limit in the misspecified space under mild conditions (see, for example, White (1982)). However, our specific asymptotic setting, in particular both m and n are allowed to increase to infinity, requires an extension of the standard theory.

Under the alternative hypothesis, the true generating LDA density is $p_{\mathcal{S}_{[n]}|G_q^*}$, *i.e.,* $\mathcal{S}_{[n]}^i \sim P_{\mathcal{S}_{[n]}|G_q^*}$, $i = 1, \ldots, m$, and the quasi-likelihood for the surrogate estimation is the LDA density $p_{\mathcal{S}_{[n]}|G'}$, $G' \in \overline{\mathcal{G}}^1$. The proposed surrogate ML estimate $\widehat{G}^1$ is equivalent to the following QMLE:

$$\widehat{G}^1 = \underset{G' \in \overline{\mathcal{G}}^1}{\arg\max} \; L^m(\mathcal{S}_{[n]}^{[m]}; G'), \tag{5.33}$$

where the quasi-likelihood is defined as

$$L^m(\mathcal{S}_{[n]}^{[m]}; G') = \frac{1}{n} \log P^m_{\mathcal{S}_{[n]}|G'}(\mathcal{S}_{[n]}^{[m]}) = \sum_{i=1}^{m} \frac{1}{n} \log P_{\mathcal{S}_{[n]}|G'}(\mathcal{S}_{[n]}^i). \tag{5.34}$$

Note that the factor $1/n$ properly scales the log-likelihood of each document.

We make the following identifiability assumption on the surrogate ML polytope under the alternative hypothesis:

Assumption (A5) Under the alternative hypothesis, as $m \to \infty$, $n \to \infty$ such that $\log(n) = o(m)$, the surrogate ML polytope $\widehat{G}^1$ converges to a $(K+1)$-polytope $G_1^* \in \overline{\mathcal{G}}^1$ almost surely. In addition, the Hausdorff distance between G_1^* and G_0 is

bounded away from zero, *i.e.,* $d_{\mathcal{H}}(G_1^*, G_0) > 0$ almost surely.

Under Assumption (A5), the surrogate ML polytope G_1^* is bounded away from the null polytope G_0 in Hausdorff distance as $m, n \to \infty$ under our specified setting. This implies that by choosing the threshold $\rho < d_{\mathcal{H}}(G_1^*, G_0)$, the proposed surrogate Hausdorff test is consistent under the alternative hypothesis, *i.e.,*

$$P^m_{\mathcal{S}_{[n]}|G_q^*}\left(d_{\mathcal{H}}(G_1^*, G_0) < \rho\right) \to 0, \qquad \text{as } m, n \to \infty. \tag{5.35}$$

Due to the vanishing behavior of the sequence $\delta_{m,n}$, for some sufficiently large m and n we have $C_1\delta_{m,n} < d_{\mathcal{H}}(G_1^*, G_0)$. Therefore there exists some decision threshold ρ that satisfies $C_1\delta_{m,n} < \rho < d_{\mathcal{H}}(G_1^*, G_0)$. Such threshold leads to the correct convergences of the detection errors under both the null and alternative hypotheses. Therefore the consistency of the hypothesis testing problem HT-q is proven.

□

Remarks on Assumption (A5): The first part of Assumption (A5) specifies the convergence of the surrogate ML polytope, which is an extension of the results established in White (1982). Specifically, Assumption (A5) assumes the convergence of a specifically structured quasi-maximum likelihood polytope estimator, under an asymptotic setting where both m and n are allowed to increase to infinity. The second part of Assumption (A5) further characterizes the limiting polytope G_1^* and assumes its distinguishability from the null polytope G_0. Here we provide some intuitions for these assumptions.

For any fixed n, using similar arguments as in White (1982), the empirical quasi-likelihood (5.34) can be shown to converge to its expectation, which is a well-defined function of G' (dependent on n). Then the quasi-maximum likelihood polytope converges to the (unique) optimum of the expectation function by the continuous map-

ping theorem for M-estimation (van der Vaart & Wellner, 1996). As $n \to \infty$, we have increasingly sufficient observations for each document, the empirical word frequency vector of each document converges to its population mean, which lies in the true topic polytope G_q^* under the alternative hypothesis. Correspondingly, the $(K+1)^{th}$ extreme point of the quasi-maximum likelihood polytope is expected to converge to some (relative) interior point of G_q^*, which is bounded away from G_0 (a boundary facet of G_q^*) almost surely. See Section 5.11 for a more elaborated presentation of these arguments.

We conjecture that the statement of Assumption (A5) can be shown to be true using a rigorous proof (and possibly along the line of arguments as described above). However, due to technical reason, we present it as an assumption in the current paper, and leave its rigorous proof for future work (Meng et al., 2014b).

5.6.3 Proof of Theorem 5.3

Proof. The first set of inequalities (Eq. (5.20) and (5.21)) are direct implications of Theorem 5.2. The second set of relations (Eq. (5.22) and (5.23)) can be proven similarly as Theorem 5.1, by noting that Theorem 5.2 now provides convergence rates (Eq. (5.19)) of the ML polytope $\widehat{G}^1$ with respect to G_0 (and G_1^*) under the null hypothesis (alternative hypothesis, respectively). Therefore, there exists some decision threshold ρ such that

$$C_1\delta_{m,n} < \rho < d_{\mathcal{H}}(G_1^*, G_0) - C_1'\delta_{m,n}^{\frac{1}{K}}, \tag{5.36}$$

where G_1^* is the true polytope under the alternative hypothesis, C_1 and C_1' are constants. Therefore correct convergences of the detection errors under both hypotheses are guaranteed, along with the exponential upper bounds (Eq. (5.19)). □

5.7 Algorithmic implementation and experiments

In this section, we first describe the algorithmic implementation of the proposed test, and then present results of simulations on synthetic data sets. These experimental results validate our theoretical development.

5.7.1 Algorithms for the surrogate Hausdorff test

The main computation of the proposed surrogate Hausdorff test is the ML estimation of the $(K+1)^{th}$ topic vector $\widehat{\phi}_{K+1}$ (Eq. (5.13)). This problem can be solved using variants of standard LDA learning algorithms, including the sampling-based techniques (*e.g.,* the collapsed Gibbs sampling (Griffiths & Steyvers, 2004; Porteous et al., 2008)) or variational inference (Blei et al., 2003; Foulds et al., 2013). Under the setting of HT-1, the collection of LDA topics under the alternative hypothesis is assumed to include all the (known) topics under the null hypothesis. Therefore an *incremental* version of the original LDA estimation algorithms is needed. Specifically, at the end of each iteration of the estimation algorithm (either sampling-based or variational inference), only the parameters of the $(K+1)^{th}$ topic are updated, while the remaining K topic vectors are kept fixed throughout the estimation process.

In our numerical experiments, we implement an incremental version of the variational inference algorithm (Blei et al., 2003), which is proven to be sufficiently accurate and efficient for our purpose.

5.7.2 Simulations on synthetic data sets

We demonstrate the detection performance of the proposed surrogate Hausdorff test through numerical simulations on synthetic data generated from LDA models.

The LDA models are parameterized as follows. The vocabulary size $d = 100$, the number of topics under the null hypothesis $K = 10$, the true topic vectors are

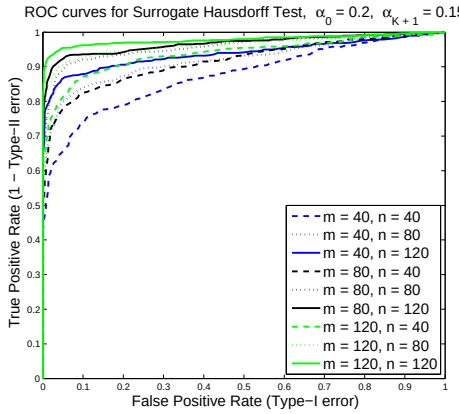
randomly sampled from a d-dimensional Dirichlet distribution with symmetric concentration parameter $\beta = 0.2$. The Dirichlet concentration parameter vector for the mixing proportions of the null topics is set to be symmetric and each component equals α_0, *i.e.,* $\alpha_j = \alpha_0$, for all $j = 1, \ldots, K$. The value of α is selected from one of the following three values $\{0.2, 0.5, 0.8\}$. The concentration parameter for the emerging topic(s) is set to be $\alpha_j = 0.15, j = K+1, \ldots, K+q$. We vary the value of m, the number of documents, and n, the number of words per documents. For each setting of (m, n), 80 pairs of document corpuses, one under the null hypothesis and the other under the alternative hypothesis, are generated according to the LDA model. Using these sampled document corpus, the proposed surrogate Hausdorff test statistics are calculated.

In Figure 5.3(a) - 5.3(c), we consider the HT-1 problem (Eq. (5.10)), in which the true LDA model under the alternative hypothesis contains one new topic. We plot the receiver operating characteristic (ROC) curves for the three settings $\alpha_0 = \{0.2, 0.5, 0.8\}$, as described above. In Figure 5.3(d), we plot the ROC curves for a HT-q problem ($q = 2$), *i.e.,* the true LDA model contains $K + 2 = 12$ topics under the alternative hypothesis. All the reported curves are averaged over 10 Monte Carlo simulations.

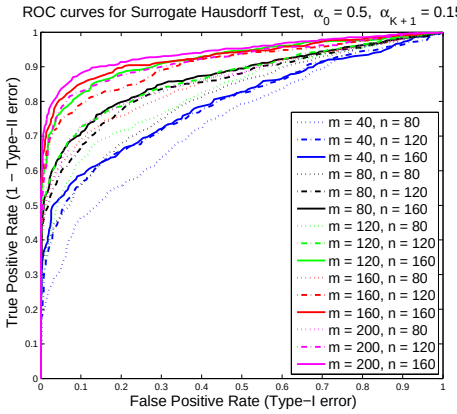
From these simulation results, several observations can be made.

When m and n both increase, the detection performance of the proposed test improves consistently under all settings. In particular, both the false positive and false negative errors decrease and approach zero, as predicted by Theorem 5.3. Note that increasing either m or n while the other quantity kept fixed does not guarantee the improved performance. This observation agrees with the derived convergence rate of the Hausdorff distance (cf. Eq. (5.18) in Theorem 5.2) in which both $m^{-\frac{1}{2}}$ and $n^{-\frac{1}{2}}$ appear.

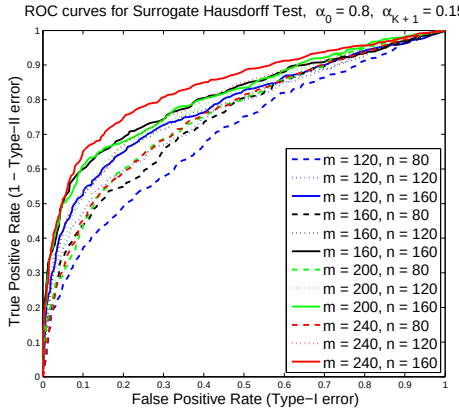
It is interesting to note that the detection performance degrades when α_0, the



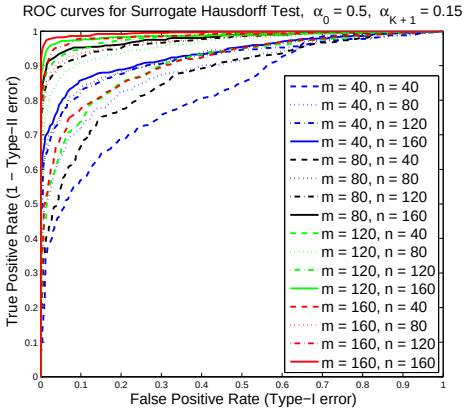
(a) ROC curves for HT-1 ($\alpha_0 = 0.2$, $\alpha_{K+1} = 0.15$).



(b) ROC curves for HT-1 ($\alpha_0 = 0.5$, $\alpha_{K+1} = 0.15$).



(c) ROC curves for HT-1 ($\alpha_0 = 0.8$, $\alpha_{K+1} = 0.15$).



(d) ROC curves for HT-q ($q = 2$, $\alpha_0 = 0.5$, $\alpha_{K+1} = \alpha_{K+2} = 0.15$).

Figure 5.3: ROC curves for the HT-1 and HT-q problems using the proposed surrogate Hausdorff test. When m and n both increase, the detection performance of the proposed test improves consistently under various LDA settings (*i.e.*, both the false positive and false negative errors decrease). The detection is more challenging when α_0, the Dirichlet concentration parameters for the null topics, is larger.

Dirichlet concentration parameters for the null topics, increases. For example, when $\alpha_0 = 0.2$ (Figure 5.3(a)), the detection performance of the proposed test using only $m = 40$ documents with length $n = 40$, almost matches that under $\alpha_0 = 0.8$ (Figure 5.3(c)) using $m = 240$ documents with length $n = 160$. Intuitively, large α_0 implies strong mixing of the topics among the documents under the null hypothesis (*i.e.,* documents are "topic dense"), which, not surprisingly, leads to difficulty in detecting emerging topics which has weakly mixing (*i.e.,* relatively small α_{K+1}). In practice, this suggests that topic detection is easier in thematically simple text corpus (such as news article), and vice versa.

While our asymptotic theory does not provide a direct explanation for such non-asymptotic phenomenon, some insights can be drawn from our derivation. In particular, the existence of a proper decision threshold that guarantees consistency relies on relation (5.36), which hold for sufficiently large m and n. Therefore a larger $d_{\mathcal{H}}(G_1^*, G_0)$ results in lower sample complexity. In Section 5.11, we show that $d_{\mathcal{H}}(G_1^*, G_0)$ depends on the magnitude of the $(K+1)^{th}$ mixing proportion θ_{K+1} (cf. (5.56)), which, in turn, depends on the relative weight of the Dirichlet parameter $\frac{\alpha_{K+1}}{\sum_{j=1}^{K+1}\alpha_j} = \frac{\alpha_{K+1}}{K\alpha_0+\alpha_{K+1}}$ (due to the property of Dirichlet distribution). Therefore increasing α_0 and/or decreasing α_{K+1} results in a large $d_{\mathcal{H}}(G_1^*, G_0)$, which leads to better non-asymptotic detection performance.

Finally, for the HT-q problem ($q = 2$), the detection performance is improved over the similarly parameterized HT-1 problem ($\alpha_0 = 0.5$ for Fig. 5.3(b) and Fig. 5.3(d)). This could possibly be explained by the fact that the increase in the number of new topics leads to an increase in the relative magnitudes of the mixing proportion for the new topics (see discussion above), thereby resulting in larger $d_{\mathcal{H}}(G_1^*, G_0)$ and better performance. In practice, this result implies that multiple emerging topics are easier to identify (although not necessarily easier to *estimate*) than a single emerging topic.

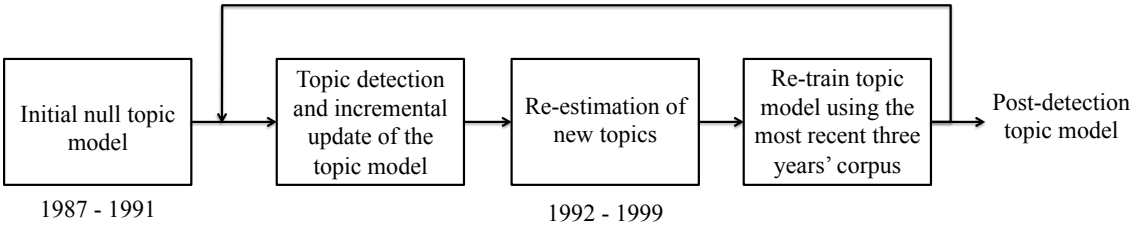


Figure 5.4: Set-up of the sequential emerging topic detection in the NIPS corpus.

5.7.3 Experiments on the NIPS corpus

We consider the problem of detecting emerging topics in a real-world document corpus, the NIPS corpus, which consists of the published papers on Neural Information Processing Systems (NIPS) conferences from 1987 to 1999 [3]. This corpus contains 2,484 documents, each of which is represented as a word count vector over a vocabulary of 14,036 words. We model the corpus using the LDA model, and apply the proposed emerging topic detection algorithm in a sequential manner.

The set-up of the sequential detection is as follows (also see Figure 5.4 for an illustration). We train an initial LDA model using the collection of the first five years' NIPS papers from 1987 to 1991. The topic polytope associated with this initial model is used as the null polytope for the first detection. For each year from 1992 to 1999, we perform the proposed surrogate Hausdorff test to discover new topics and update the current topic model incrementally. Specifically, the test is performed multiple times until either the null hypothesis is accepted or the maximum number of new topics per year (set as three) is achieved. In our experiments, we choose the decision threshold of the proposed test as half of the Hausdorff distance between the first detected new topic in that year and the previous topic polytope. After the detection terminates, we perform a re-estimation of the new topics while holding the null topics fixed. This re-estimation step is important to extract the "clean" new

[3]The data set was processed by Sam Roweis and Gal Chechik. It can be retrieved at http://ai.stanford.edu/~gal/data.html.

topics, since our proposed surrogate estimation, which only adds one topic at a time, is not guaranteed to be consistent (in terms of estimation) when there exist more than emerging topics (*e.g.,* the surrogate ML topic estimate could be a convex combination of the new emerging topics, see Figure 5.5 for an example). The re-estimation of the new topics will refine the previous estimation and alleviate the model misspecification effect. After the re-estimation, we re-train and update the entire topic model using the most recent three years' corpus. The re-trained model then is used as the null model for the subsequent year. This post-detection re-training captures the changes, or even disappearance, of the pre-existing topics in the more recent corpus, which is important for an accurate detection in a time-varying setting. For all the (re-)training of null models, we discard the redundant and junk topics, which are identified by near-zero Hausdorff distance with respect to the topic polytope of the other topics, and negligible Dirichlet concentration parameter estimate (estimated from the data as described in Blei et al. (2003)), respectively. In practice, we find these two criteria yield plausible results.

The discovered new topics (after the re-estimation) for each year are listed in Section 5.13, and the pre- and post-detection topics are included in Section 5.14. For each topic, we show the top 30 words that have the largest probabilities. It can be observed that the detected topics clearly demonstrate certain evolving trends in the NIPS papers over the considered decade. In particular, many earlier discovered topics emphasize on neural-related subjects (such as *rat, hippocampal*, and *visual*), while some of the more recent topics present word combinations about emerging machine learning techniques (such as *kernel, support, vector, machines* and *independent, component, analysis*).

Due to the absence of ground truth, the quality of topic modeling on real-world corpora is often more difficult to evaluate quantitatively than those on the synthetic data. Some empirical measures, such as the pointwise mutual information (PMI) (Newman

et al., 2010), often utilize external information. For our NIPS corpus, we consider to investigate the coherence between the detected topics and the key words in the annual Call For Papers (CFP) of the NIPS conferences, which reflect the emerging trends and subjects specified by the organizers. Using this evidence, many interesting results can be found. For example, our proposed method identifies the emergence of the *support vector machines* (SVM) on the year of 1998. This is coherent to the CFP of 1998, where the key words of SVM first appeared in the CFP of NIPS. Another SVM-related topic is again detected in the year of 1999 by our algorithm, where *svm* co-occurs with *theorem, proof, conditions, bound*, etc., indicating that it is more focused on the theory of SVM. Another notable finding is the topics related to *independent component analysis* (ICA), which are discovered on the year of 1996 and 1997, respectively. This detection is interestingly ahead of the first appearance of ICA in the CFP of NIPS on the year of 2000. Also note that from 1996 to 1999, ICA-related topics evolve from emerging topics to a persistent component of the post-detection topic collections (*i.e.,* Topic 5 in the post-detection model).

Another interesting phenomenon is the effect of model misspecification under our surrogate detection procedure, and its correction using the re-estimation. For instance, the first detected topic in the year of 1994 *before the re-estimation* is the follows:

Detected Topic 1 in 1994: field em tasks teacher map mlp skills user saliency variance memory student mixtures noise subjects recall active context pyramid hme mixture wta **jordan** item **michael** experts chess eq long loss...

The co-occurrence of *michael jordan* may not be surprising considering the fact that Dr. Michael I. Jordan co-authored seven papers in that year of NIPS. However, after the re-estimation of the new topics, this co-occurrence becomes less notable

(disappeared from the top 30 words). One possible explanation is that, since our proposed estimation only uses one topic as a surrogate for multiple emerging topics, the "Michael Jordan" topic is identified as the first surrogate topic due to its high volume of of occurrences. However, it might be the case that this "Michael Jordan" topic lies in the convex hull of some other emerging topics, and hence it is replaced by the "vertex topics" during the re-estimation (again, see Figure 5.5 for an illustration). This observation shows consistency to our geometric intuitions for the proposed detection procedure under model misspecification.

5.8 Summary of Chapter 5

We propose the surrogate Hausdorff test for topic detection in the latent Dirichlet allocation model. Our theory shows that this computationally efficient test procedure has strong performance guarantees for detecting emerging topics. The theory is supported by numerical simulation on synthetic data sets and experiments on a real-world corpus.

5.9 Proof of Lemma 5.1

Proof. The proof of Lemma 5.1 follows similar steps of Theorem 5 and Lemma 2 in Nguyen (2014). The main difference is that when $G = G_0$, a new geometric lemma is needed to characterize the volume of $G' \backslash G_0$ in terms of the Hausdorff distance $d_{\mathcal{H}}(G', G_0)$, where $G' \in \overline{\mathcal{G}}^1$. The proof of Lemma 5.3 is included in Section 5.12.

Lemma 5.3. *Under Assumption (A0), for all $G' \in \overline{\mathcal{G}}^1$, there exists a positive constant $C_2 = C_2(G_0)$ such that the following holds*

$$\mathrm{vol}_K(G' \backslash G_0) \geq C_2 d_{\mathcal{H}}(G_0, G'). \tag{5.37}$$

When $G, G' \in \overline{\mathcal{G}}^1 \backslash \{G_0\}$, a similar relation with exponent equals to K can be shown to hold following similar arguments to Lemma 2 in Nguyen (2014). Note that the thick body assumption required for this lemma is trivially satisfied, since we have assumed the null polytope is not degenerate and fixed.

For arbitrary polytopes $G = G_0$ or $G \in \overline{\mathcal{G}}^1$ and $G' \in \overline{\mathcal{G}}^1$, we lower bound the Hellinger divergence between the LDA densities $p_{\mathcal{S}_{[n]}|G}$ and $p_{\mathcal{S}_{[n]}|G'}$ by their total variation distance:

$$h(p_{\mathcal{S}_{[n]}|G}, p_{\mathcal{S}_{[n]}|G'}) \geq V(p_{\mathcal{S}_{[n]}|G}, p_{\mathcal{S}_{[n]}|G'}) = \sup_A \left| P_{\mathcal{S}_{[n]}|G}(\widehat{\eta} \in A) - P_{\mathcal{S}_{[n]}|G'}(\widehat{\eta} \in A) \right|,$$

where $\widehat{\eta}$ is the empirical word frequency vector corresponding to document $\mathcal{S}^i_{[n]}$, $\widehat{\eta}_l := \frac{1}{n}\sum_{t=1}^n \mathbb{I}(X_t = l)$ for $l = 0, \ldots, d$, and the supremum is taken over all measurable subsets of Δ^d. In the proof of Theorem 5 in Nguyen (2014), a suitable test set is constructed to further lower bound the above variational distance in terms of the Hausdorff distance $d_{\mathcal{H}}(G, G')$. Let ϵ be some constant such that $0 < \epsilon \leq d_{\mathcal{H}}(G, G')/4$, it can be shown that

$$\left| P_{\mathcal{S}_{[n]}|G}(\widehat{\eta} \in A^*) - P_{\mathcal{S}_{[n]}|G'}(\widehat{\eta} \in A^*) \right| \geq C_4 \epsilon - 2(d+1)\exp(-2n\epsilon^2/(d+1)), \quad (5.38)$$

where we have used the 0-regularity of the Dirichlet-induced density $P_{\eta|G'}$ (or $P_{\eta|G}$) proven in Lemma 4 of Nguyen (2014), and $C_4 = C_4(\alpha, K, d)$ is a constant (independent of m and n) defined therein.

When $\epsilon = O(\log n/n)$ as assumed in Lemma 5.1, the second term in (5.38) is dominated by the first term, which concludes the proof. □

5.10 Proof of Lemma 5.2

Proof. We first state the following lemma which provides an upper bound on the Kullback-Leibler divergence of two alternative LDA densities in terms of the Euclidean distance between their $(K+1)^{th}$ topic vectors. The proof of Lemma 5.4 is included in Section 5.12.

Lemma 5.4. *Let $G, G' \in \overline{\mathcal{G}}^1$ be two polytopes with $(K+1)$ extreme points such that any point $\eta \in G \cup G'$ satisfies $\min_{l=0,\ldots,d} \eta_l > c_0$ for some constant $c_0 > 0$. Then the Kullback-Leibler divergence between their associated densities satisfy*

$$K(p_G, p_{G'}) \leq \frac{n\overline{\alpha}_{K+1}}{c_0} \|\phi_{K+1} - \phi'_{K+1}\|, \tag{5.39}$$

where $\overline{\alpha}_{K+1} := \alpha_{K+1} / \sum_{j=1}^{K+1} \alpha_j$ is a scaled Dirichlet concentration parameter for the mixing proportions, ϕ_{K+1} and ϕ'_{K+1} are the corresponding $(K+1)^{th}$ extreme points of G and G', respectively.

For any averaged LDA density $\overline{p}_G \in \overline{\mathcal{P}}(\delta)$, by definition we have $h(\bar{p}_G, p_{G_0}) \leq \delta$. By Lemma 4.2 in van de Geer (2000), it can be shown that $h(p_G, p_{G_0}) \leq 4h(\bar{p}_G, p_{G_0}) \leq 4\delta$. Then using Lemma 5.1 (for $G \in \overline{\mathcal{G}}^1 \backslash \{G_0\}$), $d_{\mathcal{H}}(G, G_0) \leq (4\delta/C_1)^{1/K} = O(\delta^{1/K})$, which implies that the volume of $\Phi(\delta)$ (defined in Eq. (5.29)) is $O(\delta^{d/K})$, *i.e.*, $\mathrm{vol}_d\ \Phi(\delta) = O(\delta^{d/K})$.

Let $\{\phi_{t_1}, \phi_{t_2}, \ldots, \phi_{t_N}\}$ denote an ϵ-net of $\Phi(\delta)$, i.e. $\epsilon \leq \|\phi_{t_i} - \phi_{t_j}\| \leq 2\epsilon$, $\forall i, j = 1, \ldots, N$. For each ϕ_{t_i}, an associated polytopes $G_{t_i} = \mathrm{conv}(\Phi, \phi_{t_i})$ can be defined. Accordingly, we obtain a set of alternative models associated with the N points,

$\{G_{t_1}, \ldots, G_{t_N}\}$. For any $i, j = 1 \ldots, N$, we have

$$\begin{aligned} h^2(\bar{p}_{G_i}, \bar{p}_{G_j}) &\overset{(i)}{\leq} \frac{1}{2} K(\bar{p}_{G_i}, \bar{p}_{G_j}) \\ &\leq \frac{1}{2} \int \log \frac{\frac{1}{2}p_{G_i} + \frac{1}{2}p_{G_0}}{\frac{1}{2}p_{G_j} + \frac{1}{2}p_{G_0}} (\frac{1}{2}p_{G_i} + \frac{1}{2}p_{G_0}) d\mu \\ &\overset{(ii)}{\leq} \frac{1}{2} \left(\frac{1}{2} K(p_{G_i}, p_{G_j}) + \frac{1}{2} K(p_{G_0}, p_{G_0}) \right) \\ &= \frac{1}{4} K(p_{G_i}, p_{G_j}) \\ &\overset{\text{(Lem. 5.4)}}{\leq} \frac{n\overline{\alpha}_{K+1}}{4c_0} \|\phi_{t_i} - \phi_{t_j}\| \\ &\leq \frac{n\overline{\alpha}_{K+1}\epsilon}{2c_0}, \end{aligned}$$

where (i) is a basic inequality, (ii) is due to the convexity of KL divergence.

Then the following brackets cover the set $\overline{\mathcal{P}}^{1/2}(\delta)$:

$$\left[\sqrt{\bar{p}}^{L_i}, \sqrt{\bar{p}}^{U_i}\right] := \left[\bar{p}_{G_{t_i}}^{1/2} - (\frac{n\overline{\alpha}_{K+1}\epsilon}{2c_0})^{1/2},\ \bar{p}_{G_{t_i}}^{1/2} + (\frac{n\overline{\alpha}_{K+1}\epsilon}{2c_0})^{1/2}\right],\ i = 1, 2, \ldots, N.$$

To see this, note that for any point $\tilde{\phi} \in \Phi(\delta)$, there exists some j such that $\|\phi_{t_j} - \tilde{\phi}\| \leq 2\epsilon$. By similar arguments as above, we have $h^2(\bar{p}_{\tilde{G}}, \bar{p}_{G_j}) \leq \frac{n\overline{\alpha}_{K+1}\epsilon}{2c_0}$, which implies that $\sqrt{\bar{p}_{\tilde{G}}}$ lies in the j^{th} bracket.

Therefore, the entropy with bracketing of $\overline{\mathcal{P}}(\delta)$ endowed with Hellinger metric, i.e. $H_B\left(\sqrt{\frac{n\overline{\alpha}_{K+1}\epsilon}{2c_0}}, \overline{\mathcal{P}}^{1/2}(\delta)\right)$, is upper bounded by the entropy number of $\Phi(\delta)$, $H\left(\epsilon, \Phi(\delta)\right) = \log N$, which is $O\left(\frac{d}{K} \log\left(\frac{\delta}{\epsilon}\right)\right)$ since $\mathrm{vol}_d\ \Phi(\delta) = O(\delta^{d/K})$. □

5.11 Intuitions for Assumption (A5) and some results

In this section, we provide some intuitions for Assumption (A5) and present some relevant results.

Step 1. For any $n > 0$, we first establish the convergence behavior of the empirical quasi-likelihood. Let the true generating density be $p_{\mathcal{S}_{[n]}|G_q^*}$. The quasi-likelihood for the surrogate estimation under the alternative hypothesis is the LDA density with respect to $p_{\mathcal{S}_{[n]}|G'}$ for $G' \in \overline{\mathcal{G}}^1$. Due to the unique mapping between the $(K+1)$-polytope in $\overline{\mathcal{G}}^1$ and its $(K+1)^{th}$ extreme point, the LDA density can be equivalently parameterized by ϕ_{K+1}:

$$p_{\mathcal{S}_{[n]}}(\mathcal{S}_{[n]}^i; \phi_{K+1}) = \int_{\Delta^K} \exp\left\{ n \sum_{l=0}^{d} \widehat{\eta}_{i,l} \log\left(\sum_{j=1}^{K} \phi_{jl}\theta_j + \phi_{K+1,l}\theta_{K+1} \right) \right\} dP_{\theta;\alpha}.$$

The quasi-log-likelihood is (also parameterized by ϕ_{K+1}):

$$L^m(\mathcal{S}_{[n]}^{[m]}; \phi_{K+1}) = \sum_{i=1}^{m} n^{-1} \log P_{\mathcal{S}_{[n]}}(\mathcal{S}_{[n]}^i; \phi_{K+1}). \tag{5.40}$$

First we note that, for any fixed $n > 0$, $\left| n^{-1} \log p_{\mathcal{S}_{[n]}(\mathcal{S}_{[n]}^i)} \right|$ is bounded. To see this, apply the mean value theorem (Olmsted, 1961) to the integration with respect to θ, we have

$$\begin{aligned}
\left| n^{-1} \log p_{\mathcal{S}_{[n]}}(\mathcal{S}_{[n]}^i) \right| &= \left| n^{-1} \log \exp\left\{ n \sum_{l=0}^{d} \widehat{\eta}_{i,l} \log\left(\sum_{j=1}^{K} \phi_{jl}\overline{\theta}_j + \phi_{K+1,l}\overline{\theta}_{K+1} \right) \right\} \right| \\
&= \left| \sum_{l=0}^{d} \widehat{\eta}_{i,l} \log\left(\sum_{j=1}^{K} \phi_{jl}\overline{\theta}_j + \phi_{K+1,l}\overline{\theta}_{K+1} \right) \right|,
\end{aligned}$$

for some $\overline{\theta} \in \Delta^K$ (note that $\overline{\theta}$ is dependent on n). Due to Assumption (A4) that the components of topic vectors ϕ_{jl} are bounded away from zero, and the constraint that $\sum_{j=1}^{K+1} \overline{\theta}_j = 1$, the right-hand side of the last display is bounded.

Given the boundedness and continuity of the quasi-log-likelihood, and that the parameter ϕ_{K+1} lies in a compact subset of a Euclidean space (*i.e.*, Δ^d), Mickey's

Theorem (given as Theorem 2 in Jennrich (1969)) states that

$$m^{-1}\sum_{i=1}^{m} n^{-1}\log P_{\mathcal{S}_{[n]}}(\mathcal{S}^i_{[n]};\phi_{K+1}) \xrightarrow{m\to\infty} \mathbb{E}\left[n^{-1}\log P_{\mathcal{S}_{[n]}}(\mathcal{S}^i_{[n]};\phi_{K+1})\right] \tag{5.41}$$

uniformly for all $\phi_{K+1} \in \Delta^d$ and almost every sequence $(\mathcal{S}^i_{[n]})_{i=1}^m$. Note that the expectation is taken with respect to the true generating distribution $P_{\mathcal{S}_{[n]}|G^*_q}$.

Step 2. Next, we consider the limit of the expected quasi-log-likelihood (*i.e.,* the right-hand side quantity in (5.41)) as $n \to \infty$.

Note that the expected quasi-log-likelihood can be expressed as follows

$$\mathbb{E}\left[n^{-1}\log P_{\mathcal{S}_{[n]}|G'}(\mathcal{S}^i_{[n]})\right] = n^{-1}\int_{\Delta^d} \log P_{\mathcal{S}_{[n]}|G'}(\mathcal{S}^i_{[n]}) dP_{\mathcal{S}_{[n]}|G^*_q}. \tag{5.42}$$

Recall the generative process of LDA model, the marginal density can be expressed as an integration over the random vector η in the topic polytope G. Therefore, we have $P_{\mathcal{S}_{[n]}|G'} = \int_{G'} P_{\mathcal{S}_{[n]}|\eta'} dP_{\eta'|G'}$ and $P_{\mathcal{S}_{[n]}|G^*_q} = \int_{G^*_q} P_{\mathcal{S}_{[n]}|\eta} dP_{\eta|G^*_q}$ for the quasi- and true likelihoods, respectively. These relations lead to

$$\begin{aligned}
&\mathbb{E}\left[n^{-1}\log P_{\mathcal{S}_{[n]}|G'}(\mathcal{S}^i_{[n]})\right] \\
&= n^{-1}\int_{\Delta^d} \log \int_{G'} \exp\left[n\sum_{l=0}^{d} \widehat{\eta}_{il}\log \eta'_{il}\right] dP_{\eta'|G'}(\eta'_i) dP_{\mathcal{S}_{[n]}|G^*_q}(\mathcal{S}^i_{[n]}) \qquad (5.43)\\
&= n^{-1}\int_{G^*_q}\int_{\Delta^d} \log\left\{\int_{G'} \exp\left[n\sum_{l=0}^{d} \widehat{\eta}_{il}\log \eta'_{il}\right] dP_{\eta'|G'}(\eta'_i)\right\} dP_{\mathcal{S}_{[n]}|\eta_i} dP_{\eta|G^*_q}(\eta_i), \qquad (5.44)
\end{aligned}$$

where $\widehat{\eta}_i$ (the empirical word frequency vector) lies in Δ^d, η_i (the population word frequency vector) lies in the true generating polytope G^*_q, and η'_i (the surrogate population word frequency vector) lies in the surrogate polytope G'.

Conditioning on η_i (the population word frequency vector in G^*_q), as $n \to \infty$, the empirical word frequency vector $\widehat{\eta}_i$ of the words $(X_{ij})_{j=1}^n$, which are i.i.d. draws from

the multinomial distribution Multi(η_i), converges to the population word frequency vector η_i almost surely. Therefore, loosely speaking, $P_{\mathcal{S}_{[n]}|\eta_i} \to \mathbb{I}(\widehat{\eta}_i = \eta_i)$, which leads to the following conjectured relation:

$$\begin{aligned}
&\mathbb{E}\left[n^{-1}\log P_{\mathcal{S}_{[n]}|G'}(\mathcal{S}^i_{[n]})\right] \\
&\xrightarrow{n\to\infty} \int_{G_q^*} n^{-1}\log\int_{G'}\exp\left[n\sum_{l=0}^{d}\eta_{il}\log\eta'_{il}\right]dP_{\eta'|G'}(\eta'_i)dP_{\eta|G_q^*}(\eta_i). \qquad (5.45)
\end{aligned}$$

Note that the right-hand side quantity in (5.45) is still dependent on n. Therefore Eq. (5.45) is *not* a well-defined convergence argument.

Although the convergence with respect to the right-hand side quantity in (5.45) is not established, it is interesting to investigate its properties, in particular its concavity in ϕ_{K+1}. Apply the mean value theorem (Olmsted, 1961) to the outer and inner integrals, respectively, we have

$$\begin{aligned}
&\int_{G_q^*} n^{-1}\log\int_{G'}\exp\left[n\sum_{l=0}^{d}\eta_{il}\log\eta'_{il}\right]dP_{\eta'|G'}(\eta'_i)dP_{\eta|G_q^*}(\eta_i) \\
&= n^{-1}\log\int_{G'}\exp\left[n\sum_{l=0}^{d}\overline{\eta}_l\log\eta'_{il}\right]dP_{\eta'|G'}(\eta'_i) \\
&= n^{-1}\log\exp\left[n\sum_{l=0}^{d}\overline{\eta}_l\log\overline{\eta'}_l\right] \\
&= \sum_{l=0}^{d}\overline{\eta}_l\log\overline{\eta'}_l, \qquad (5.46)
\end{aligned}$$

for some $\overline{\eta} \in$ relint G_q^* and $\overline{\eta'} \in$ relint G' (the existence of these two vectors is guaranteed by the continuity of the integrated function, see, for example, Olmsted (1961, Page 346)). Note that both $\overline{\eta}$ and $\overline{\eta'}$ are potentially dependent on n.

Since $\overline{\eta'}$ can be parameterized by $\overline{\eta'} = \sum_{j=1}^{K+1}\phi_j\overline{\theta}_j$, where $\overline{\theta} \in$ relint Δ^K is the corresponding topic proportion vector (again, $\overline{\theta}$ depends on n). Then the quantity

in (5.46) can be expressed as a function of ϕ_{K+1}:

$$\sum_{l=0}^{d} \overline{\eta}_l \log \overline{\eta'}_l = \sum_{l=0}^{d} \overline{\eta}_l \log \left(\sum_{j=1}^{K} \phi_{j,l} \overline{\theta}_j + \phi_{K+1,l} \overline{\theta}_{K+1} \right). \tag{5.47}$$

Combine (5.41), (5.45), (5.46) and (5.47), as $m \to \infty$ and $n \to \infty$, we have the following *conjecture*:

$$m^{-1} \sum_{i=1}^{m} n^{-1} \log P_{\mathcal{S}_{[n]}}(\mathcal{S}^i_{[n]}; \phi_{K+1}) \to \sum_{l=0}^{d} \overline{\eta}_l \log \left(\sum_{j=1}^{K} \phi_{j,l} \overline{\theta}_j + \phi_{K+1,l} \overline{\theta}_{K+1} \right) \tag{5.48}$$

uniformly in ϕ_{K+1}.

Step 3. *Assume* the convergence argument in (5.48) holds for some limiting function that takes the following form

$$L^*(\phi_{K+1}) = \sum_{l=0}^{d} \overline{\eta}_l \log \left((1 - \overline{\theta}_{K+1}) \mathbf{a}_l + \phi_{K+1,l} \overline{\theta}_{K+1} \right), \tag{5.49}$$

where $\overline{\eta}$ and $\overline{\theta}$ are *independent* of n, and the vector $\mathbf{a}$ is defined as

$$\mathbf{a} = \frac{1}{1 - \overline{\theta}_{K+1}} \sum_{j=1}^{K} \phi_{jl} \overline{\theta}_j. \tag{5.50}$$

It can be easily verified through the calculation of second-order derivative that $L^*(\phi_{K+1})$ is strictly concave. Therefore it has a unique maximum, denoted as ϕ^*_{K+1}:

$$\widehat{\phi}^*_{K+1} = \underset{\phi_{K+1} \in \Delta^d}{\arg\max} \sum_{l=0}^{d} \overline{\eta}_l \log \left((1 - \overline{\theta}_{K+1}) \mathbf{a}_l + \phi_{K+1,l} \overline{\theta}_{K+1} \right). \tag{5.51}$$

Recall our surrogate ML estimate of ϕ_{K+1} can be equivalently defined as the fol-

lowing maximizer of the empirical quasi-likelihood

$$\widehat{\phi}_{K+1} \in \underset{\phi_{K+1}\in\Delta^d}{\arg\max}\ m^{-1}\sum_{i=1}^{m} n^{-1}\log P_{\mathcal{S}_{[n]}}(\mathcal{S}^i_{[n]};\phi_{K+1}). \tag{5.52}$$

Due to the *assumed* convergence (5.48) and the uniqueness of the optimum of $L^*(\phi_{K+1})$, the continuous mapping theorem for the M-estimation (also known as the *Argmax continuous mapping theorem* (van der Vaart & Wellner, 1996, Theorem 3.2.2)) implies:

$$\widehat{\phi}_{K+1} \xrightarrow{a.s.} \widehat{\phi}^*_{K+1}. \tag{5.53}$$

By the unique mapping between the $(K+1)^{th}$ extreme point and the $(K+1)$-polytope in $\overline{\mathcal{G}}^1$, we have

$$\widehat{G}^1 \xrightarrow{a.s.} G^*_1, \tag{5.54}$$

which is the first part of Assumption (A5).

Step 4. Recall $\overline{\theta} \in \text{relint } \Delta^K$, it can be verified that the vector $\mathbf{a}$ defined in (5.50) satisfies $\mathbf{a} \in \text{relint } G_0$, where we recall $G_0 = \text{conv}(\phi_1, \ldots, \phi_K)$ (see Figure 5.5).

It can be shown through the Karush-Kuhn-Tucker conditions of the optimization problem (5.51) that its unique optimum ϕ^*_{K+1} satisfies

$$\phi^*_{K+1,l} = 0, \quad \text{or} \quad \phi^*_{K+1,l} = \lambda\overline{\eta}_l - \frac{1-\overline{\theta}_{K+1}}{\overline{\theta}_{K+1}}\mathbf{a}_l, \quad l = 0, \ldots, d \tag{5.55}$$

for some constant λ. Due to our assumption that the topic vector components are bounded away from zero (Assumption (A4)) and the normalization constraint, it can

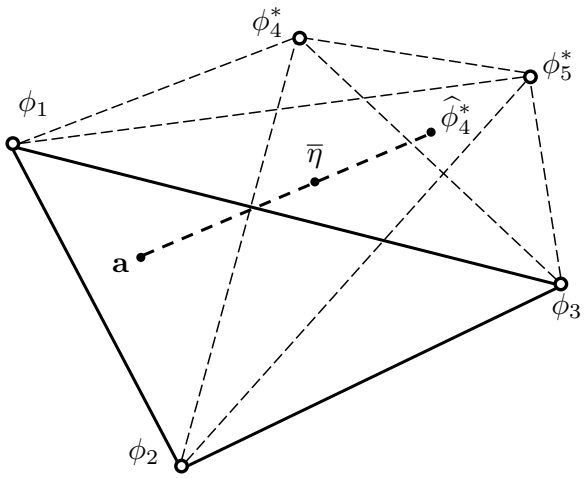


Figure 5.5: Illustration of Eq. (5.56). The null topic polytope $G_0 = \text{conv}(\phi_1, \phi_2, \phi_3)$ ($K = 3$). The true topic polytope (under the alternative hypothesis) $G_q^* = \text{conv}(\phi_1, \phi_2, \phi_3, \phi_4^*, \phi_5^*)$ has five extreme points ($q = 2$). Vectors $\mathbf{a} \in \text{relint } G_0$ and $\overline{\eta} \in \text{relint } G_q^*$. As a result of Eq. (5.56), the limiting surrogate ML estimate of the extreme point $\widehat{\phi}_{K+1}^*$ is bounded away from G_0. All points denote vectors in Δ^d, the d-dimensional word probability simplex (not shown).

be verified that $\lambda = 1/\overline{\theta}_{K+1}$, which leads to the relation

$$\overline{\eta} = \overline{\theta}_{K+1}\widehat{\phi}^* + (1 - \overline{\theta}_{K+1})\mathbf{a}. \tag{5.56}$$

The above display implies that, the vector $\overline{\eta}$ lies on the line segment of $(\mathbf{a}, \widehat{\phi}^*)$ (excluding the endpoints since $\overline{\theta}_{K+1} > 0$). Under Assumption (A0), the null polytope G_0 is a boundary facet of the higher dimensional polytope G_q^* almost surely (for $q \geq 1$ under the alternative hypothesis). This implies vector $\mathbf{a}$, which is a relative interior point of G_0, is also on the boundary of G_q^*. Since vector $\overline{\eta}$ is a relative interior point of G_q^*, (5.56) implies that $\widehat{\phi}^*$ is bounded away from G_0 (see Figure 5.5 for an illustration). Therefore, $d_{\mathcal{H}}(\text{conv}(G_0, \widehat{\phi}^*), G_0) > 0$, which is the second part of Assumption (A5).

Remarks: We conjecture that the statements of Assumption (A5) can be shown to

be true using a rigorous proof (which is possibly similar to the above arguments). However, the main technical difficulty is to establish a relation that is similar to (5.48), in which the right-hand side quantity is independent of n. Another possibly minor technical issue is the equivalence between our desirable asymptotic setting (*i.e.,* both m and n increase to infinity simultaneously) and the setting presented in the above arguments (*i.e.,* for any fixed n, first let $m \to \infty$, then let $n \to \infty$). We conjecture that the limits under two asymptotic settings are equal for our problem, which requires certain additional uniform convergence argument. For these reasons, we introduce Assumption (A5) and leave its rigorous proof for future work.

5.12 Proofs of auxiliary lemmas

Proof of Lemma 5.3. Let $\phi_{K+1}^{\perp}$ be the projection of ϕ_{K+1} onto G_0 (see Figure 5.1). When $d \gg K$, the volume of $G' \backslash G_0$ is almost surely a K-pyramid which has apex ϕ_{K+1} and base G_0. Therefore $C_2 = \mathrm{vol}_{K-1}(G_0) \sin\varphi$ suffices for Eq. (5.37) to hold, where φ is the angle between aff G_0 and the line segment $[\phi_{K+1}, \phi_{K+1}^{\perp}]$. □

Proof of Lemma 5.4. Our proof utilizes an established upper bound on the Kullback-Leibler divergence $K(p_{\mathcal{S}_{[n]}|G}, p_{\mathcal{S}_{[n]}|G'})$ in terms of the Wasserstein distance between $p_{\eta|G}$ and $p_{\eta|G'}$, defined as

$$W_1(p_{\eta|G}, p_{\eta|G'}) := \inf_{Q \in \mathcal{Q}} \int \|\eta_1 - \eta_2\| dQ(\eta_1, \eta_2), \tag{5.57}$$

where $Q(\eta_1, \eta_2)$ denotes a coupling of $P_{\eta|G}$ and $P_{\eta|G'}$, *i.e.,* a joint distribution over $G \times G'$ whose induced marginal distributions of η_1 and η_2 are equal to $P_{\eta|G}$ and $P_{\eta|G'}$, respectively. $\mathcal{Q}$ is the set of all such couplings.

The following lemma (Lemma 6 in Nguyen (2014)) provides an upper bound on the Kullback-Leibler divergence $K(p_{\mathcal{S}_{[n]}|G}, p_{\mathcal{S}_{[n]}|G'})$:

Lemma 5.5. *Let $G, G' \subset \Delta^d$ be closed convex subsets such that any $\eta = (\eta_0, \ldots, \eta_d) \in G \cup G'$ satisfies $\min_{l=0,\ldots,d} \eta_l > c_0$ for some constant $c_0 > 0$. Then*

$$K(p_{\mathcal{S}_{[n]}|G}, p_{\mathcal{S}_{[n]}|G'}) \leq \frac{n}{c_0} W_1(p_{\eta|G}, p_{\eta|G'}). \tag{5.58}$$

For distributions $P_{\eta|G}$ and $P_{\eta|G'}$, where $G, G' \in \overline{\mathcal{G}}^1$, we construct a coupling $Q^*(\eta, \eta')$ such that under Q^*, $\eta = \sum_{j=1}^{K+1} \theta_j \phi_j$ and $\eta' = \sum_{j=1}^{K+1} \theta_j \phi'_j$, *i.e.,* η and η' share the same Dirichlet-distributed topic mixing proportion vector θ. This construction is a valid coupling, since the Dirichlet distributions are identically parameterized by Assumption (A1), and the two polytopes share the same K extreme points $\phi_1, \ldots, \phi_K$. We have

$$\begin{aligned}
K(p_{\mathcal{S}_{[n]}|G}, p_{\mathcal{S}_{[n]}|G'}) &\leq \frac{n}{c_0} W_1(p_{\eta|G}, p_{\eta'|G'}) \\
&\leq \frac{n}{c_0} \int \|\eta - \eta'\| dQ^*(\eta, \eta') \\
&\leq \frac{n}{c_0} \mathbb{E}_{Q^*} \sum_{j=1}^{K+1} \theta_j \|\phi_j - \phi'_j\| \\
&= \frac{n}{c_0} \mathbb{E}_{Q^*} \theta_{K+1} \|\phi_{K+1} - \phi'_{K+1}\| \\
&= \frac{n \overline{\alpha}_{K+1}}{c_0} \|\phi_{K+1} - \phi'_{K+1}\|,
\end{aligned}$$

where the last step is due to the property of Dirichlet distribution, and $\overline{\alpha}_{K+1} := \alpha_{K+1} / \sum_{j=1}^{K+1} \alpha_j$. □

5.13 Detected topics in NIPS corpus

Year 1992:

Detected Topic 1 :

query uncertainty gender insertion hole mill epsilon pendulum lqr assembly filter

rom hjb realization slopes queries committee sensed infomax modes realizations interconnection gullapalli interconnections restart filters inductive load theoretic monkey

Detected Topic 2 :

rat distance character place hippocampal image goal tangent representations channel environment regions hand pre location representation scenes visual window manager inverse fault spiral manifolds circular dimensionality workspace waves food codebook

Detected Topic 3 :

rules approximation probability evidence model stochastic perturbation bootstrap distribution building bayesian pruning rule examples eq hints posterior children blocks selection prior similarity criterion moody lemma descent asymptotic membership theorem missing

Year 1993:

Detected Topic 1 :

filter leech entropy silicon face odor binding dopamine retina eeg exemplar contour nucleotides fault primary smyth krogh friesen shapes dna subjects site detection bee splice mach substructures becker genes sites

Detected Topic 2 :

theorem routing asynchronous convergence algorithm finite obs markov algorithms dynamics state obd transition momentum discrete stochastic pruning monte carlo lms communication behavior programming states pca observable methods tree proof board

Detected Topic 3 :

distance cross data clustering feature features cluster tangent classification images code validation sound address owl missing block surface signature auditory regression analysis similarity implicit localization reconstruction vectors nearest activity

representations

Year 1994:

Detected Topic 1 :

transformation regime actor plasma means spectral online estimators talkers aggregation bishop batch subspace attenuation electrotonic linsker series decorrelation separation moment transformed periodic andreas min blind magnetic limiting angle pan predictor

Detected Topic 2 :

field hand subjects force hint operator motor optical manifold position facial subject auditory tau hints interpolation interference song mst day location learned shift blocking video group reinforcement learning delta tracking

Detected Topic 3 :

em teacher tasks noise skills saliency mlp active user pyramid student recall queries hme wta chess mixtures item loss map context memory module capacity eq query experts coding compression long

Year 1995:

Detected Topic 1 :

loss spike character characters field entropy train transfer bounds lvq rate template window minima rates bound logistic isi reference receptive poisson warmuth pixels detections constraints ocr particle ill dimensionality msec

Detected Topic 2 :

robot student period symmetric sensory system dynamical filter frequency channel dynamics overlaps stage bifurcation motion cochlear integrator onset teacher environment gains motor dependencies cochlea filters path place stability simd navigation

Detected Topic 3 :

tree trees ensemble motor eeg risk stopping perceptrons ica depth decision features family algorithm energy table visible pruning variables boosting split terminal early perceptron representations clusters spin shavlik validation yang

Year 1996:

Detected Topic 1 :

experts hme expert batch utility entropic gating stationary growing spherical mem unlabelled grbf accelerator collective tdnns delays updates utilities welch controller hebb baum diagnosis tdnn wan questions disease earth saddle

Detected Topic 2 :

disparity carlo monte biases interpolation contour spectrum demonstration player rivalry mse kalman contours sampling pole elements power gammon trial binocular natural penalty edge spatiotemporal neal players hyperparameters curves electric segments

Detected Topic 3 :

separation graph som source sources ica call algorithms independent channel blind feature component style decision graphs classifier matrix samples price filters query energy tree algorithm content weak bagging classification extra

Year 1997:

Detected Topic 1 :

release robot module cia monotonic hybrid subject conditioning feedback observer adaptation ideal reflectance monotonicity winnow shading diffusion words refractory path subjects observers compensation synapse generative rating free control production refractoriness

Detected Topic 2 :

tree decision trees distribution graph hierarchical edge ham nodes blood ltm com-

posite divergence edges level adaboost leaves rectified kullback covariance margin weiss kalman ekf hierarchy matching leibler opinion measurement conditional

Detected Topic 3 :

words analysis eeg wavelet rules retrieval target time coding system frequency component preference independent memory ranking ica source word similarity rule sources artifacts low human user displays gradient sec attractor

Year 1998:

Detected Topic 1 :

latent sensor pca documents rotation sparse options rotated shrinkage fusion digit words image labeled unlabeled lie visualisation query vertex singular option upright minimax images retrieval plan lookahead tipping integral transformation

Detected Topic 2 :

rule state message synergy mode states transition dynamics hmms series stationary exact fisher memoryless replica dynamical messages tap hebbian modes ltp inference merge detection mutual equations coin coolen recall behaviors

Detected Topic 3 :

kernel margin support svm clustering inverse arm forward adaboost human subjects objects pca segments positive familiarity smo metric hand distance position threshold fig speed convolution wind euclidean machines motor curve

Year 1999:

Detected Topic 1 :

channel nonlinear stochastic capacity channels voltage ekf mode predictive densities ion states branch steady resonance pairing modes density current entropy wavelet dual dynamic statistic law gain fluctuations past nonparametric missing

Detected Topic 2 :

kernel svm support margin boosting solution belief tree adaboost machines nodes node theorem svms propagation decision variational hypothesis potential convex weak class vapnik algorithms decoding vector unique working proof conditions

Detected Topic 3 :

information localization sound iiii cue attractor mutual speech language spectral task cues context location som features light similarity length human wire acquisition arbors subject languages matlab spatial documents subjects mobile

5.14 Pre- and post-detection topics

Initial topics in the null model (trained on the corpus from 1987 to 1991) :

Topic 1:

network neural input learning networks time output set function units figure model training number system data weights layer hidden problem error algorithm unit results information pattern state performance systems net

Topic 2:

model cells cell neurons input firing response neuron synaptic activity time stimulus figure spike frequency cortex fig phase potential inhibitory synapses membrane excitatory threshold network single system patterns inhibition neural

Topic 3:

speech recognition training word classifiers performance classifier classification speaker mlp phoneme trained hmm words test set language phonetic experiments segmentation system frame signal mixture features multi characters rbf acoustic waibel

Topic 4:

image images pixel feature vision color pixels object surface maps field optical visual edges line segmentation depth points recognition discontinuities features illumination ring digit operator poggio edge competitive data map

Topic 5:

learning reinforcement action task state controller expert control robot actions sutton barto environment world current connectionist forward tasks learn adaptive exploration planning goal architecture policy evaluation jordan steps category play

Topic 6:

chip circuit analog neuron figure voltage vlsi neural current synapse weight digital neurons circuits synaptic charge chips implementation mead design input synapses cmos silicon device bit pulse output transistor gate

Topic 7:

error generalization probability distribution convergence dimension examples sample complexity class log classifier bound bounds capacity algorithm theorem minimum risk polynomial vapnik decay distributions haussler empirical bayesian uniform baum bayes converge

Topic 8:

memory capacity hopfield matrix associative vectors neurons memories stored code vector recall neuron states equilibrium codes attraction theorem probability storage stable state winner outer optical retrieval inverse bit convergence operation

Topic 9:

functions basis regression tree prediction data cart approximation rbf radial error moody polynomial trees validation mars spline splines cross variables friedman centers kernel receptive nearest mlp nonlinear classification gaussians density

Topic 10:

motion visual direction velocity eye moving image field spatial receptive murphy

wind stimulus joint cells position arm figure motor stimuli rotation speed object sensory target location activity processing velocities directions

Topic 11:

visual cells cortical eye cell neurons dominance cortex model activity map ocular field center brain receptive development maps synaptic synapses wave spatial spin orientation monocular lgn correlation owl stimuli rod

Topic 12:

functions threshold function polynomial theorem approximation depth optimization circuit boolean ary proof circuits degree bounded constrained gate points smooth wavelet differential lagrangian annealing boltzmann rational shallow variables spectral complexity constraint

Topic 13:

control motor arm trajectory model movements movement muscle feedback controller cerebellar brain cerebellum command fig cmac joint position kawato cortex commands change climbing fiber minimum velocity parallel hand organization torque

Topic 14:

object objects features recognition views feature image view visual images aspect gating instantiation frame edelman module attention traffic viewpoint reference rotation visit recognize intrator orientation manipulated rigid invariant operator extracted

Topic 15:

orientation feature map maps ocular dominance space principle cells neighborhood cortex infomax cortical visual models centers lateral topographic extraction features loss projection dimensional spatial bifurcation regions kohonen organization hubel cell

Topic 16:

eq star adjoint barhen trajectory dipole toomarian equations path trajectories paths

dynamical processors processor slots gulati operator message sensor sensitivity attitude green settling eqn pen scales zipser simd operators reverse

Topic 17:

attractors symbols content hit role symbol man baird projection normal filters periodic chaotic filter eigenvectors dog associative recursive eigenvalues ghost bifurcation basic verb semantic coordinates attractor trees stability blocks coarse

Topic 18:

delay oscillation stability bifurcation dynamics oscillatory bat symmetric westervelt echoes sonar feedback sustained gain echo marcus lattice delayed basin inhibition eq simmons fixed attractors lateral phys fig attractor moss snr

Topic 19:

vor velocity head eye gain vestibular storage reflex movements vestibulo ocular robinson compensation anastasio responses normal pathway canal slip compensatory solid dotted lisberger oculomotor muscles dashed monkey compensated motoneurons pair

Topic 20:

eeg kernel diagnosis fault vigilance antenna diagnostic faults diagnostics erp engine plant inference vehicle smyth theoretic classifiers controllers radiology pointing funds bull head operating monitoring lda event fund magnetic potentials

Topic 21:

strings stack state grammar recurrent automata finite string giles language chen grammars pollack length grammatical tomita extracted inference languages song sun fsa lee induction seed symbol regular action automaton watrous

Topic 22:

ann impulse rap receiver ripple receivers user optimum depth ring connect video spectral access fire impulses regional anns kohn shot matched software users channel median anomalous detector conventional library communications

Topic 23:

concept mistake tracker adversary concepts hypothesis tracking examples conservative sampling selective load instance pac label change rate theorem benign half rivest day greedy chooses planes atlas incremental kuh drawn varying

Topic 24:

fish tin electric integrator echoes echo harmonic winter column assembly source chorales moore bach melody confidence sensory clicks body music skeleton transitions potential objects assemblies animal skin chord harmony ryan

Topic 25:

role representation tensor grammar product binding tree regular structures roles language filler parse markov connectionist vectors source cons distributed representations represented entry representing bindings depth bayesian decomposition bayes smolensky inference

Topic 26:

leg insect controller nervous walking stance cochlear locomotion legs dispersion foot swing body design behaviors speeds patients stepping artificial temperatures natural insects controllers movements sensors angle cell currents membrane heterogeneous

Topic 27:

activation pairing saliency grouping signature pot role binding bindings spreading elements inside john groups object word spread concept localist salient node language conceptual nodes structured actor paths maximal contours plan

Topic 28:

road vehicle autonomous lane driving navigation steering video retina drive roads direction image finder sensor ans cmu intensity pomerleau camera situations exemplars ahead images travel curvature robot thorpe obstacle land

Topic 29:

jutten template signals separation waveforms waveform matching sources cohen pure

templates rault blind mixed music delayed overlapping auto delays herault interference pouliquen vittoz andreou cancellation cij superimposed filtering comparisons faggin

Topic 30:

fuzzy cell anna dsp board expert chip extraction rules linguistic crisp cells group automated antecedent lisp step instruction groups importance intermediate gallant proposition select bus neurocomputer boser billion chips disorders

Topic 31:

genetic memory vowel harmony weather location round phonology sparse kanerva addresses fitness holland predicting rogers rain clustering month population buffer counters hybrid hours members segments trigger wheeler davis phonological locations

Topic 32:

gamma routing message focused principe tdnn interconnection routes memory utilized vries route greedy kernels depth convolution exhaustive ports eeg multistage adaline additive goudreau legal tapped delays kuo mackey giles port

Topic 33:

font relevance letter skeleton letters skeletonization mozer production exception strokes trimming style smolensky stork trim car fonts hunt median east west assessment metric diamond music ron judged pratt composition project

Topic 34:

robot path planning inverse workspace kinematics configuration sta redundant manipulator oscillatory terminal dof branches norm demers neighborhood kinematic positioning bellman branch obstacles disturbance classifier neighborhoods robotics globally labels generically balls

Topic 35:

fault tolerance faults cmac cluster saturated exemplar judge operational synergy sensitivity generator lincoln placement activated discontinuous impact mistake toler-

ant adjustment adjustable drift skrzypek clustered rms stuck addresses manipulator retraining robotic

Topic 36:

classifier message match post heart strength classifiers links messages environmental jabri genetic environment list chamber ventricular cycle node link leong matched parents davis flower chi reproduction xie dual matches population

Topic 37:

faces face sex subjects humans human facial emotion compression female male emotions holons gender cottrell russell eyes discrimination expressions responses emotional identity mouth munro gray monkey discriminations relaxed served cortex

Topic 38:

linsker miller eigenvectors eigenvalues mackay eigenvalue centre tit eigenvector hebbian covariance surround eigenfunctions principal eigenfunction structures perturbed synaptic modes correlations regimes goodhill populations eigen dayan correlational oct synapses operator angular

Topic 39:

food insect feeding arousal animal reticular facial lobe search energy behavior beer behaviors sensor spinal goal fish mouth environment motivated hungry fin tracking patch weiss artificial insects intelligent chemical oral

Topic 40:

map retinal eye tectal maps markers brain axons fibers half cowan normal cij schmidt experiment meyer sij regeneration fiber development formation tectum retina yoon expanded experiments stick friedman contact details

Topic 41:

energy tank boltzmann ensemble hopfield evidence tour tours hop landscapes attractor valid wta machines clique winner spaces cities clause traveling hypothesis thresholds probe salesman escape minima gap shapes probes uphill

Topic 42:

clause proof protein clauses constraints instance backbone distance resolution query row bound matrix proteins proofs amino bohr prediction base literal instances structures copy rows logic original atomic secondary exists propositional

Topic 43:

sleep rem sequencing wake dream sutton cholinergic brainstem wanted whitaker loops cognitive synthesis wet mind permission influences loop disorder college norepinephrine cued widespread populations bifurcations mckenna rhythms cognition neuromodulation excitability

Topic 44:

rules dna splice stop kbann links knowledge extracted towell shavlik minus junction base junctions extraction ann acc don promoter sequences domain superior genes anns knn noordewier refinement symbolic pages protein

Topic 45:

stress syllable heavy vowel primary syllables linguistic nucleus phonology language languages word edge learnability wheeler touretzky gupta string metrical alt penultimate voicing vowels long deletion onset secondary perceptron alternation french

Topic 46:

signatures signature true verification false acceptance rejection handwritten roc goodman writer handwriting people stanford rosenberg writing centralized info cards individuals checks transactions cumulative radar automated tures nettalk eng plot jackel

Topic 47:

mfa spin annealing temperature averages equilibrium spins graph raleigh ece bout bins den hamiltonjan van miller relaxation ising snyder lowered markov chain bin experimentally temperatures begins hopfield attractive wesley iterations

Topic 48:

chain region mouse human cell receptor beta proteins domains heavy alpha protein class domain rat false bengio conserved detected amino positives acids score programs sequences fruit epsilon fly search database

Topic 49:

cues target cue access targets intersection strengths humphreys wiles product dennis auto arc cross retrieved weakly tensor processes strongly orthogonal associator temporary semantic cued multiplication cancel blend direct multiplying combination

Post-detection topics (re-trained on the corpus from 1995 to 1999):

Topic 1:

data model learning algorithm set function distribution training number neural models figure error results gaussian parameters networks probability case network space problem linear approach time vector method matrix noise input

Topic 2:

model neurons time input figure network visual synaptic neuron response stimulus cell neural information spike cells learning activity firing rate system cortical stimuli noise frequency units cortex function patterns shown

Topic 3:

image images recognition object face features feature figure visual network objects texture based set target information detection human performance pattern representation faces vision system task local scene search pixel examples

Topic 4:

state learning policy time reinforcement action function states optimal algorithm control reward actions problem figure model step system rl agent markov algorithms set number decision robot goal mdp environment based

Topic 5:

ica independent source basis sources separation blind component signal components

signals eeg natural sparse image analysis coefficients matrix wavelet images noise sejnowski mixing artifacts functions data gradient processing coding code

Topic 6:

kernel margin svm support training set adaboost vector algorithm function class decision boosting machines functions error examples solution algorithms classification theorem classifiers cost svms tree generalization classifier bound working convex

Topic 7:

circuit chip system figure analog output time current control vlsi motion signal input circuits field flow image cell shown systems frequency processing voltage visual implementation response design neural direction template

Topic 8:

iiii sound clustering information task language similarity localization attractor human concept som cluster light speech subjects cues cue mutual documents sequences spectral acquisition arbors wire length languages matlab source location

Topic 9:

loss bounds functions function theorem bound networks neural threshold dimension proof linear network class polynomial computational number upper input lower analog units sigmoidal unit result experts gate bounded ranking ltm

Topic 10:

words instruction system instructions wavelet rules user word preference gradient coding network retrieval blocks schedule time semantic melody table dec syllable scheduling video pact errors knowledge concentration guidance bits instance

Topic 11:

speech hmm state recognition hmms word words acoustic system states segmentation time transition continuous character segmented articulatory emission fusion markov pdfs articulator audio acoustics arc entropic continuity speaker topology constrained

Topic 12:

graph tree nodes trees node level edge matching hierarchical edges graphs clique parent clustering match matches structures structural structure opinion cut markov hierarchy correspondence randomized similarity pairwise maximal logarithmic transducers

Topic 13:

control forward inverse sv hand motor arm feedback position kernels movement controller wind adaptive vapnik convolution support desired kernel front sch familiarity ridge human regularization kopf smola impulse force movements

Topic 14:

tap solutions codes stationary message sequential replica phys boltzmann tanh spin recall weiss code shortcut equations messages glass decoding ferromagnetic nadal saddle gallager correcting connections equilibrium behaviour coolen gee stationarity

Topic 15:

teacher student dynamics xly hebbian phys macroscopic perceptron equations generalization coolen decay saad replica observables restricted saddle batch lyapunov sgn evolution asymptotic activations simulations theory activation dynamical formalism regime cavity

Topic 16:

adaptation cia monotonic words feedback subject speech conditioning winnow monotonicity path vowel compensation production generalization diffusion control altered experiment sound rating formants subjects blocking attributes bond hyperplane word testing abu

Topic 17:

steady prop gain jacobian industrial plant historical momentum pearlmutter leapfrog optimizer tensor chemical setpoint austin hamiltonian technion parsimonious technologies refining johnson exploration neuron auxiliary keeler riemannian martin discretization scheduling implementing

Topic 18:

documents sensor fusion words retrieval lsi document queries axes sensors word topic webb collection band topics magnification relevant indexing query semantic complementary polarity fused ml text projected collections kurtosis spie

Topic 19:

storage capacity committee gardner connected phys fully internal replica kwon volumes elementary rev overlaps parity volume breaking symmetry lett conventional korea barkai mitchison durbin mechanics machine violates europhys representations engel

Topic 20:

cleaning zimmermann neuneier dilemma observer cleaned eta forecasting var unified awt neuro covers penalty return obd flat damage siemens bond reviewed financial finnoff increments forecast trading markets psychological asset hit

Topic 21:

environment learnt characters distortion ocr japanese baxter canonical neighbour bartlett jonathan environments character misclassification pac cedar neighbours possessing err learner triples printed recogni buffalo maintained quantization classifi july english clas

CHAPTER 6

Conclusions and Future Work

Probabilistic graphical models provide a principled and flexible framework for compactly characterizing structural high-dimensional distributions. This dissertation has focused on three related problems on graphical modeling: (i) Distributed estimation of parameters and principal components of Gaussian graphical models; (ii) Learning of high-dimensional Gaussian graphical models with latent variables; and (iii) Detection of emerging topics in topic models. For each of these three problems, we have proposed novel and efficient algorithms, derived theoretical guarantees for the performance, and provided numerical experiments on both synthetic and real-world data sets to validate and complement the theory.

In Chapter 2, we have proposed a distributed framework for estimating the inverse covariance matrix in Gaussian graphical models based on convex relaxations of the marginal likelihood maximization problem in local neighborhoods. A global estimate is obtained by combining the local estimates without the need of iterative global message passing. We derived asymptotic properties for the proposed distributed estimator, as well as bounds on the high-dimensional error rates, which are comparable to those of the centralized maximum likelihood estimator. We validated the statistical and computational efficiencies of the proposed estimator through numerical experiments on synthetic graphical models with various structures and a real-world sensor network data sets.

In Chapter 3, we continued our exploration of distributed estimation framework for Gaussian graphical models, but focused on directed acyclic graphs and the estimation of principal components. We proposed a distributed algorithm, named DDPCA, for performing principal component analysis (PCA) in directed Gaussian graphical models. DDPCA was designed for exploiting structured sparsity arising from the Cholesky factors of the concentration matrix, and was shown to be closely related to the sum-product algorithm. We illustrated DDPCA's computational and communication advantages in a synthetic online subspace tracking problem and the anomaly detection in a real-world computer network.

In Chapter 4, we have presented a family of latent variable Gaussian graphical model (LVGGM) whose inverse covariance matrix has a sparse plus low-rank structure, extending the sparsity assumption used in the previous two chapters. This LVGGM was motivated by the missing-value prediction problem in recommender systems, and its structural assumptions were validated on real data. For LVGGM, we derived parameter estimation error bounds for regularized maximum likelihood estimation in the high-dimensional setting. Results of numerical simulations agreed with our theory, and the proposed LVGGM-based method achieved superior missing-value prediction performance on a real-world movie rating data set.

In Chapter 5, we have proposed the surrogate Hausdorff test for detecting emerging topics in a specific topic model, the latent Dirichlet allocation (LDA) model. The test statistic was shown to be a computationally tractable surrogate for the standard likelihood ratio, and is proven to guarantee the asymptotic consistency of the topic detection problem. We further provided the convergence rate of the test statistic and bounds on the detection errors when there is at most a single emerging topic. The theory is validated through numerical experiments on both synthetic and real-world document corpora.

6.1 Future Work

There are many interesting directions that are worthy of future study:

First, it would be worthwhile to investigate the non-Gaussian extensions of the proposed learning frameworks in Chapter 2 through Chapter 4. More specifically, extending the distributed estimation framework in Chapter 2 to Markov random fields with non-Gaussian distributions will lead to significant improvement in the computational complexity with respect to the centralized learning algorithms. There have been some recent work along this direction, *e.g.,* Mizrahi et al. (2013); Massam & Wang (2013)), but the understanding is still limited. The non-Gaussian scenario of the distributed algorithm proposed in Chapter 3 is also worth exploring. The learning algorithm and analysis framework for LVGGM considered in Chapter 4 both are potentially generalizable to graphical models parameterized by broader class of distributions.

With the recent advances of parallel computing machineries, such as Hadoop [1] and Spark [2], the parallel learning framework proposed in Chapter 2 can be easily implemented and applied to very large-scale graphs. A real-world example is the social networks, which typically contain millions of nodes. As illustrated in Section 2.6, the proposed algorithm provides a reasonable tradeoff between the statistical approximation precision and the computational complexity on synthetic small-world graphs, one of the standard models for social networks. An interesting future direction would be to pursue this direction using real data and more realistic (and challenging) graphs that relax the Markovian and/or Gaussian assumptions, such as the Erd

Hos-R

'enyi graphs (Erdős & Rényi, 1976), scale-free or power law networks (Liu & Ihler, 2011) and Kronecker graphs (Leskovec et al., 2010). Another challenge for learning

[1] http://hadoop.apache.org
[2] https://spark.apache.org

many real networks is the possible existence of "hubs" – nodes that have very large number of neighbors. Extending the current learning algorithm to networks with hubs is also a future direction worth studying.

In Chapter 2 and 3, the structure of the graphical model is assumed known. An interesting direction is to estimate the structure of graphical models in a similar distributed fashion. This appears to be a harder problem, but some insights regarding distributed algorithms and convex relaxation in particular can perhaps be shared. Furthermore, approximating loopy undirected models by proper directed models is one possible way to leverage DDPCA's advantageous computation complexity and convergence properties.

The current LVGGM learning and prediction framework is constrained by the computational complexities of (i) regularized maximum likelihood estimation of the marginal inverse covariance matrix; (ii) the conditional mean prediction rule which involves matrix inversion. Computationally more efficient optimization techniques for LVGGM learning and approximation algorithm for the matrix inversion that take advantage of the "low-rank plus sparse" structure are both interesting topics for future efforts. Another direction worth investigating is the effect of missing observations on LVGGM learning, and the robustness of the learning algorithm against such missingness. The proposed learning and prediction framework can also be utilized for other related applications besides recommender systems, such as the stock market data (see, for example, Luo (2011); Meng et al. (2014a) for empirical studies).

Lastly, for the topic detection framework proposed in Chapter 5, future directions include theoretical investigation of Assumption (A5) and the convergence rate of the maximum likelihood polytope under model misspecification (*e.g.,* when there exist more than a single emerging topics) or parameter misspecification (*e.g.,* unknown Dirichlet concentration parameters). More in-depth empirical studies, in particular on real data, are also worth exploring.

BIBLIOGRAPHY

Agarwal, Alekh, Negahban, Sahand, and Wainwright, Martin J. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics*, 40(2):1171–1197, 2012.

AlSumait, Loulwah, Barbará, Daniel, and Domeniconi, Carlotta. On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *Proceedings of the Eighth IEEE International Conference on Data Mining*, pp. 3–12. IEEE, 2008.

Anderson, T.W. *An introduction to multivariate statistical analysis*, volume 2. Wiley New York, 1958.

Banerjee, O., El Ghaoui, L., d'Aspremont, A., and Natsoulis, G. Convex optimization techniques for fitting sparse Gaussian graphical models. In *ACM International Conference Proceeding Series*, volume 148, pp. 89–96. Citeseer, 2006.

Bell, Robert M and Koren, Yehuda. Lessons from the netflix prize challenge. *ACM SIGKDD Explorations Newsletter*, 9(2):75–79, 2007.

Bickel, Peter J, Ritov, Yaacov, and Tsybakov, Alexandre B. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Blei, David M. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

Blei, David M. Build, compute, critique, repeat: data analysis with latent variable models. *Annual Review of Statistics and Its Application*, 1:203–232, 2014.

Blei, David M and Lafferty, John D. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 113–120. ACM, 2006.

Blei, David M, Ng, Andrew Y, and Jordan, Michael I. Latent Dirichlet allocation. *the Journal of Machine Learning Research*, 3:993–1022, 2003.

Boyd, Stephen and Vandenberghe, Lieven. *Convex optimization*. Cambridge university press, 2009.

Candès, Emmanuel J, Li, Xiaodong, Ma, Yi, and Wright, John. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

Chandrasekaran, Venkat, Parrilo, Pablo A, and Willsky, Alan S. Latent variable graphical model selection via convex optimization. *Annals of Statistics*, 40(4): 1935–1967, 2012.

Chen, Changyou, Ding, Nan, and Buntine, Wray. Dependent hierarchical normalized random measures for dynamic topic modeling. *arXiv preprint arXiv:1206.4671*, 2012.

Chen, Yilun, Wiesel, Ami, and Hero, Alfred O. Robust shrinkage estimation of high-dimensional covariance matrices. *Signal Processing, IEEE Transactions on*, 59(9): 4097–4107, 2011.

Choi, Myung Jin, Chandrasekaran, Venkat, and Willsky, Alan S. Gaussian multiresolution models: Exploiting sparse Markov and covariance structure. *Signal Processing, IEEE Transactions on*, 58(3):1012–1024, 2010.

Choi, Myung Jin, Tan, Vincent YF, Anandkumar, Animashree, and Willsky, Alan S. Learning latent tree graphical models. *Journal of Machine Learning Research*, 12: 1729–1770, 2011.

Dahl, J., Vandenberghe, L., and Roychowdhury, V. Covariance selection for non-chordal graphs via chordal embedding. *Optimization Methods and Software*, 23(4): 501–520, 2008.

Erdős, P and Rényi, A. On the evolution of random graphs. *Selected Papers of Alfréd Rényi, vol*, 2:482–525, 1976.

Foulds, J. R. and Smyth, P. Annealing paths for the evaluation of topic models. In *Proceedings of the Thirtieth Conference Conference on Uncertainty in Artificial Intelligence*, 2014.

Foulds, James, Boyles, Levi, DuBois, Christopher, Smyth, Padhraic, and Welling, Max. Stochastic collapsed variational Bayesian inference for latent Dirichlet allocation. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 446–454. ACM, 2013.

Friedman, J., Hastie, T., and Tibshirani, R. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag New York, 2009.

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

Gao, Zekai, Song, Yangqiu, Liu, Shixia, Wang, Haixun, Wei, Hao, Chen, Yang, and Cui, Weiwei. Tracking and connecting topics via incremental hierarchical dirichlet processes. In *Proceedings of the 11th IEEE International Conference on Data Mining*, pp. 1056–1061. IEEE, 2011.

Gastpar, M., Dragotti, P.L., and Vetterli, M. The distributed karhunen–loève transform. *Information Theory, IEEE Transactions on*, 52(12):5177–5196, 2006.

Golub, G.H. and Van Loan, C.F. *Matrix computations*. Johns Hopkins Univ Pr, 1996.

Griffiths, Thomas L and Steyvers, Mark. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004.

Guestrin, C., Bodik, P., Thibaux, R., Paskin, M., and Madden, S. Distributed regression: an efficient framework for modeling sensor network data. In *Information Processing in Sensor Networks, 2004. IPSN 2004. Third International Symposium on*, pp. 1–10. IEEE, 2004.

Guo, Jian, Levina, Elizaveta, Michailidis, George, and Zhu, Ji. Graphical models for ordinal data. *Journal of Computational and Graphical Statistics*, (just-accepted): 00–00, 2014.

Heinemann, U. and Globerson, A. What cannot be learned with bethe approximations. *arXiv preprint arXiv:1202.3731*, 2012.

Hsieh, Cho-Jui, Sustik, Mátyás A, Dhillon, Inderjit, Ravikumar, Pradeep, and Poldrack, Russell. BIG & QUIC: Sparse inverse covariance estimation for a million variables. In *Advances in Neural Information Processing Systems*, pp. 3165–3173, 2013.

Huang, J.Z., Liu, N., Pourahmadi, M., and Liu, L. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98, 2006a.

Huang, Ling, Nguyen, XuanLong, Garofalakis, Minos, Jordan, Michael I, Joseph, Anthony, and Taft, Nina. In-network pca and anomaly detection. In *Advances in Neural Information Processing Systems*, pp. 617–624, 2006b.

Jennrich, Robert I. Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics*, pp. 633–643, 1969.

Johnson, C.C., Jalali, A., and Ravikumar, P. High-dimensional sparse inverse covariance estimation using greedy methods. *arXiv preprint arXiv:1112.6411*, 2011.

Johnson, J. Fisher information in Gaussian graphical models. *unpublished technical note*, 2006.

Jordan, M.I. and Bishop, C.M. An introduction to graphical models. *unpublished book*, 2001.

Kakade, Sham, Shamir, Ohad, Sindharan, Karthik, and Tewari, Ambuj. Learning exponential families in high-dimensions: Strong convexity and sparsity. In *International Conference on Artificial Intelligence and Statistics*, pp. 381–388, 2010.

Kalaitzis, Alfredo and Lawrence, Neil D. Residual component analysis: Generalising PCA for more flexible inference in linear-Gaussian models. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pp. 209–216, 2012.

Kemmerer, R.A. and Vigna, G. Intrusion detection: a brief history and overview. *Computer*, 35(4):27–30, 2002.

Kolar, Mladen and Xing, Eric P. Consistent covariance selection from data with missing values. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pp. 551–558, 2012.

Koller, D. and Friedman, N. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Koren, Yehuda. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 426–434. ACM, 2008.

Koren, Yehuda, Bell, Robert, and Volinsky, Chris. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

Kschischang, F.R., Frey, B.J., and Loeliger, H.A. Factor graphs and the sum-product algorithm. *Information Theory, IEEE Transactions on*, 47(2):498–519, 2001.

Lakhina, A., Crovella, M., and Diot, C. Diagnosing network-wide traffic anomalies. In *Proceedings of ACM SIGCOMM 2004*, pp. 219–230, August 2004.

Lauritzen, S.L. *Graphical models*, volume 17. Oxford University Press, USA, 1996.

Lawrence, Neil D and Urtasun, Raquel. Non-linear matrix factorization with gaussian processes. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 601–608. ACM, 2009.

Ledoit, Olivier and Wolf, Michael. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5):603–621, 2003.

Lee, Jason, Sun, Yuekai, and Taylor, Jonathan E. On model selection consistency of penalized M-estimators: a geometric theory. In *Advances in Neural Information Processing Systems*, pp. 342–350, 2013.

Leskovec, Jure, Chakrabarti, Deepayan, Kleinberg, Jon, Faloutsos, Christos, and Ghahramani, Zoubin. Kronecker graphs: An approach to modeling networks. *The Journal of Machine Learning Research*, 11:985–1042, 2010.

Li, Lin, Scaglione, Anna, and Manton, Jonathan H. Distributed principal subspace estimation in wireless sensor networks. *Selected Topics in Signal Processing, IEEE Journal of*, 5(4):725–738, 2011.

Liang, Percy and Jordan, Michael I. An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators. In *Proceedings of the 25th international conference on Machine learning*, pp. 584–591. ACM, 2008.

Lin, Dahua and Fisher, John W. Coupling nonparametric mixtures via latent dirichlet processes. In *Advances in Neural Information Processing Systems*, pp. 55–63, 2012.

Liu, Q. and Ihler, A. Learning scale free networks by reweighted ℓ_1 regularization. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

Liu, Qiang and Ihler, Alexander. Distributed parameter estimation via pseudo-likelihood. *International Conference on Machine Learning (ICML)*, June 2012.

Liu, Ying and Willsky, Alan. Learning Gaussian graphical models with observed or latent FVSs. In *Advances in Neural Information Processing Systems*, pp. 1833–1841, 2013.

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

Lounici, Karim. High-dimensional covariance matrix estimation with missing observations. *arXiv preprint arXiv:1201.2577*, 2012.

Luo, Xi. Recovering model structures from large low rank and sparse covariance matrix estimation. *arXiv preprint arXiv:1111.1133*, 2011.

Ma, Shiqian, Xue, Lingzhou, and Zou, Hui. Alternating direction methods for latent variable Gaussian graphical model selection. *Neural computation*, 25(8):2172–2198, 2013.

Malioutov, Dmitry M, Johnson, Jason K, and Willsky, Alan S. Walk-sums and belief propagation in Gaussian graphical models. *The Journal of Machine Learning Research*, 7:2031–2064, 2006.

Massam, Helene and Wang, Nanwei. Distributed parameter estimation of discrete hierarchical models via marginal likelihoods. *arXiv preprint arXiv:1310.5666*, 2013.

Meinshausen, Nicolai and Bühlmann, Peter. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

Meng, Zhaoshi, Wiesel, Ami, and Hero, Alfred O. Distributed principal component analysis on networks via directed graphical models. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 2877–2880. IEEE, 2012.

Meng, Zhaoshi, Wei, D., Hero, A.O., and Wiesel, A. Marginal likelihoods for distributed estimation of graphical model parameters. In *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2013 IEEE 5th International Workshop on*, pp. 73–76, Dec 2013a.

Meng, Zhaoshi, Wei, Dennis, Wiesel, Ami, and Hero, Alfred O. Distributed learning of Gaussian graphical models via marginal likelihoods. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pp. 39–47, 2013b.

Meng, Zhaoshi, Eriksson, Brian, and Hero, Alfred O. Learning latent variable gaussian graphical models. In *Proceedings of The 31st International Conference on Machine Learning*, pp. 1269–1277, 2014a.

Meng, Zhaoshi, Nguyen, XuanLong, Hero, Alfred O., and Mei, Qiaozhu. Detecting emerging topics in topic models with confidence. *Technical Report*, 2014b.

Meng, Zhaoshi, Wei, Dennis, Wiesel, Ami, and Hero, Alfred O. Marginal likelihoods for distributed parameter estimation of gaussian graphical models. *Signal Processing, IEEE Transactions on*, 62(20):5425–5438, Oct 2014c.

Mizrahi, Yariv Dror, Denil, Misha, and de Freitas, Nando. Linear and parallel learning of Markov random fields. *arXiv preprint arXiv:1308.6342*, 2013.

Murphy, Kevin P, Weiss, Yair, and Jordan, Michael I. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pp. 467–475. Morgan Kaufmann Publishers Inc., 1999.

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

Newman, David, Lau, Jey Han, Grieser, Karl, and Baldwin, Timothy. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 100–108. Association for Computational Linguistics, 2010.

Nguyen, XuanLong. Posterior contraction of the population polytope in finite admixture models. *Bernoulli*, 2014.

O'Leary, D.P. and Whitman, P. Parallel qr factorization by householder and modified gram-schmidt algorithms. *Parallel computing*, 16(1):99–112, 1990.

Olmsted, John Meigs Hubbell. *Advanced calculus*. Prentice Hall, 1961.

Pacheco, Jason and Sudderth, Erik B. Minimization of continuous Bethe approximations: A positive variation. In *Advances in Neural Information Processing Systems*, pp. 2573–2581, 2012.

Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.

Petersen, Kaare Brandt, Pedersen, Michael Syskind, Larsen, Jan, Strimmer, Korbinian, Christiansen, Lars, Hansen, Kai, He, Liguo, Thibaut, Loic, Baro, Miguel, Hattinger, Stephan, Sima, Vasile, and The, We. The matrix cookbook. Technical report, 2006.

Porteous, Ian, Newman, David, Ihler, Alexander, Asuncion, Arthur, Smyth, Padhraic, and Welling, Max. Fast collapsed Gibbs sampling for latent Dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 569–577. ACM, 2008.

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

Ravikumar, P., Wainwright, M.J., Raskutti, G., and Yu, B. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Rothman, A.J., Bickel, P.J., Levina, E., and Zhu, J. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

Rütimann, P. and Bühlmann, P. High dimensional sparse covariance estimation via directed acyclic graphs. *Electronic Journal of Statistics*, 3:1133–1160, 2009.

Salakhutdinov, Ruslan and Mnih, Andriy. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *Proceedings of the 25th international conference on Machine learning*, pp. 880–887. ACM, 2008.

Scott, James and Baldridge, Jason. A recursive estimate for the predictive likelihood in a topic model. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pp. 527–535, 2013.

Shental, O., Siegel, P.H., Wolf, J.K., Bickson, D., and Dolev, D. Gaussian belief propagation solver for systems of linear equations. In *Information Theory, 2008. ISIT 2008. IEEE International Symposium on*, pp. 1863–1867. IEEE, 2008.

Shojaie, A. and Michailidis, G. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, 97(3):519–538, 2010.

Smith, M. and Kohn, R. Parsimonious covariance matrix estimation for longitudinal data. *Journal of the American Statistical Association*, 97(460):1141–1153, 2002.

Stoev, S., Michailidis, G., and Vaughan, J. On global modeling of backbone network traffic. In *INFOCOM, 2010 Proceedings IEEE*, pp. 1–5. IEEE, 2010.

Su, Xiaoyuan and Khoshgoftaar, Taghi M. A survey of collaborative filtering techniques. *Advances in artificial intelligence*, 2009:4, 2009.

Taddy, Matt. On estimation and selection for topic models. In *International Conference on Artificial Intelligence and Statistics*, pp. 1184–1193, 2012.

Tang, Jian, Meng, Zhaoshi, Nguyen, Xuanlong, Mei, Qiaozhu, and Zhang, Ming. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proceedings of The 31st International Conference on Machine Learning*, pp. 190–198, 2014.

Teh, Yee Whye, Jordan, Michael I, Beal, Matthew J, and Blei, David M. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476), 2006.

Tron, R. and Vidal, R. Distributed computer vision algorithms through distributed averaging. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 57–63. IEEE, 2011.

van de Geer, Sara. *Empirical Processes in M-estimation*, volume 45. Cambridge University Press Cambridge, 2000.

Van der Vaart, A.W. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

van der Vaart, AW and Wellner, Jon. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 1996.

Vershynin, Roman. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Wainwright, Martin J. Estimating the wrong graphical model: Benefits in the computation-limited setting. *The Journal of Machine Learning Research*, 7:1829–1859, 2006.

Wainwright, MJ. Sharp thresholds for noisy and high-dimensional recovery of sparsity using $\ell 1$-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.

Wainwright, M.J. and Jordan, M.I. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

Wallach, Hanna M, Murray, Iain, Salakhutdinov, Ruslan, and Mimno, David. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1105–1112. ACM, 2009.

Wang, Chengjing, Sun, Defeng, and Toh, Kim-Chuan. Solving log-determinant optimization problems by a newton-cg primal proximal point algorithm. *SIAM Journal on Optimization*, 20(6):2994–3013, 2010.

Wang, Chong, Blei, David, and Heckerman, David. Continuous time dynamic topic models. *arXiv preprint arXiv:1206.3298*, 2012.

Wang, Xuerui and McCallum, Andrew. 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*, pp. 424–433. ACM, 2006.

Watts, D. and Strogatz, S. The small world problem. *Collective Dynamics of Small-World Networks*, 393:440–442, 1998.

Weiss, Y. and Freeman, W.T. Correctness of belief propagation in gaussian graphical models of arbitrary topology. *Neural computation*, 13(10):2173–2200, 2001.

Wermuth, N. Linear recursive equations, covariance selection, and path analysis. *Journal of the American Statistical Association*, pp. 963–972, 1980.

White, Halbert. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, pp. 1–25, 1982.

Wiesel, A. and Hero, A.O. Decomposable principal component analysis. *Signal Processing, IEEE Transactions on*, 57(11):4369–4377, 2009.

Wiesel, A. and Hero, A.O. Distributed covariance estimation in Gaussian graphical models. *Signal Processing, IEEE Transactions on*, 60(1):211–220, 2012.

Wiesel, A., Eldar, Y.C., and Hero, A.O. Covariance estimation in decomposable Gaussian graphical models. *Signal Processing, IEEE Transactions on*, 58(3):1482–1492, 2010.

Xu, Huan, Caramanis, Constantine, and Sanghavi, Sujay. Robust PCA via outlier pursuit. *Information Theory, IEEE Transactions on*, 58(5):3047–3064, 2012.

Yang, Eunho and Ravikumar, Pradeep. Dirty statistical models. In *Advances in Neural Information Processing Systems 26*, pp. 611–619. 2013.