Latent-Variable Modeling: Algorithms, Inference, and Applications

Thesis by
Armeen Taeb

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2020
Defended August 16th, 2019

Armeen Taeb
ORCID: 0000-0002-5647-3160

ACKNOWLEDGEMENTS

First and foremost, my sincere gratitude goes to my advisor and mentor Venkat Chandrasekaran for his unwavering support over the last six years. Venkat has been extremely generous with his time, and has helped transform me from the novice researcher at the beginning of my PhD to a more seasoned researcher by the end. Closely watching Venkat's style of research and teaching over the years, I have come to greatly admire his process: starting from first principles to develop elegant and simple solutions to mathematically hard problems. I hope to one day demonstrate such clarity in my thinking. When I first came to Caltech, I thought I would learn a thing or two about math and doing research. What I did not know was the extent to which I would experience personal growth, and a large component of this has been due to Venkat. He taught me to have conviction, to not rely on other's approval, to tackle problems logically, and to never lose my confidence throughout life's ups and downs. I will be forever indebted to Venkat for all that he is done for me, and look forward to continuing to learn from him for many years to come!

I am also grateful for the collaborators that I have had over these years: Parikshit Shah, JT Reager, Mike Turmon, and Andrew Stuart. Thanks to Parikshit Shah for having me as his intern at Yahoo, for giving a lot of useful insights, and bringing the non-Gaussian latent-variable graphical model project out from the dead! Thanks to JT for providing domain expertise for the reservoir work and being the most enthusiastic scientist I have come across! Thanks to Mike for serving as a mentor and being a role model on how to operate between theory and practice. Finally, thanks to Andrew for his kind support over the last couple of years, for having me as a TA for his course and a co-author of a book, and taking me to Switzerland, which set me up for my next career path. Outside of Caltech, I was extremely fortunate to make a connection with Rina Foygel Barber in my last year of PhD. I am immensely grateful for her hospitality during my visits to University of Chicago and providing many words of encouragement to keep me going.

Many thanks to the Caltech faculty and staff. Specifically, I am grateful to committee members Babak Hassibi, Lior Pachter, John Doyle, and Andrew Stuart for providing constructive feedback on this thesis and being helpful informal mentors. I also wish to thank Joel Tropp for his stellar teaching and serving as a role model for professionalism in academia. Thanks to the Resnick Sustainability Institute and in particular, Neil Fromer, for believing in me and funding my research. The

administrative staff at Caltech is superlative. Many thanks to Sydney Garstang, Maria Lopez, Carmen Nemer–Sirois, Sheila Shull, and Diana Bohler for all their help and kindness; my life at Caltech was so much smoother and joyful because of them.

I have been extremely fortunate to have incredible friends that have been with me throughout this journey. Thanks to Alex Turzillo, for being the first friend that understood and empathized with all my emotions! Alex stuck with me during hard times, often provided comedic relief to change my mood, and always had a loving and memorable way to not only ease a difficult situation, but leave me with something to think about. I greatly enjoyed all our backpacking trips and I look forward to many more years of friendship. Thanks to Jenny Somerville for teaching me to think of life as an adventure and to be comfortable being an explorer without a very concrete agenda in mind. I learned from her the value of empathy, and that it is OK to make a fool out of myself, because good things can come out of it! Thanks to Grace Huang for her kindness and graceful persona. She has often been the person that I would reach out to after a happy event, ranting about a bad day, or getting advice about relationships. Whatever it is, Grace has always been receptive, always gave me her time, and I always came out feeling much better. Most importantly, she is always keen to talk about our shared love for cats!

Thanks to Josh Hickernell for being a rock in my life since high school. I have enjoyed his unwavering friendship, the many tennis matches we have played, and various fun experiences in Los Angeles. Thanks to Hovey Yu for many fun movie nights and the archery outings! Thanks to Ania Baetica and Thomas Catanach for being great friends, and showing me a fine example of a healthy partnership! Thanks to Thomas Anderson for being a fun roommate and never failing to make me laugh with his crazy stories (tango and otherwise)! Thanks to Yong Sheng Soh for always being a sounding board for my research thoughts and for accompanying me in our many ridiculous adventures! Thanks to Fanny Yang for being caring, yet honest, and challenging me in many ways to become a better version of myself! Thanks to Kamyar Azizzadenesheli and Bamdad Hosseini for bringing me back to my Iranian roots and their always welcoming and warm demeanor. Thanks to Pasadena Toastmasters community for providing many words of encouragement and giving a platform for me to develop personally. Additionally, thanks to Jenny Butler for teaching me to be self reflective, for helping me explore and understand myself, and ultimately for guiding me towards becoming a healthier individual.

I am indebted to my family for their support. Thanks to my brother Ideen Taeb for reminding me to relax and giving me fashion tips throughout the years! Thanks to Mona Vajihollahi for being the most logical and straightforward person in our family to balance all the emotions! Thanks to my father for teaching me the value of hard work. Thanks to my mom for teaching me how to love, how to empathize, and how to deeply care about others. Thanks to my cat Betty, for providing a lot of emotional support throughout the years and reminding me that her needs often come ahead of mine! Finally, thanks to my sister Leili Taeb for instilling in me the importance of owning my responsibilities, being kind, and standing up for my values. Leili has stood by my side throughout these years and never hesitated to extend a hand when I needed help. This thesis would not have been possible without her, and is a testament to her kindness, patience, and sacrifice.

ABSTRACT

Many driving factors of physical systems are often latent or unobserved. Thus, understanding such systems crucially relies on accounting for the influence of the latent structure. This thesis makes advances in three aspects of latent-variable modeling: inference, algorithms, and applications. Specifically, we develop and explore latent-variable techniques that a) ensure interpretable and statistically significant models, b) can be efficiently optimized to identify best fit to data, and c) provide useful insights in real-world applications. The specific contributions of this thesis are:

- We employ a latent-variable graphical modeling technique to develop the first state-wide statistical model of the California reservoir network. With this model, we precisely characterize the system-wide behavior of the network to hypothetical drought conditions, and proposed guidelines for more sustainable reservoir management.

- Motivated by the previous application, we provide a geometric framework to assess the extent to which our latent variable model has learned true or false discoveries about the relevant physical phenomena. Our approach generalizes the classical notions of true and false discoveries in mathematical statistics that rely on the discrete structure of the decision space to settings where the decision space is continuous and more complicated. We highlight the utility of this viewpoint in problems involving subspace selection and low-rank estimation.

- We propose a convex optimization procedure to fit a latent-variable graphical model for generalized linear models. This framework provides a flexible approach to model non-Gaussian variables including Poisson, Bernoulli, and exponential variables. A particularly novel aspect of our formulation is that it incorporates regularizers that are tailored to the type of latent variables.

- We describe a computationally efficient framework to learn a latent-variable model with high-dimensional and non-iid data. This framework is based on factoriable precision operators that decouple the component associated with the observational dependencies and the component associated to interdependencies among the variables.

- We propose a convex optimization technique to provide semantics to latent variables of a factor model. This approach is based on linking auxiliary variables — chosen based on domain expertise — to these latent variables.

PUBLISHED CONTENT AND CONTRIBUTIONS

[TSC19] A. Taeb, P. Shah, and V. Chandrasekaran. "False Discovery and Its Control in Low-Rank Estimation". In: *arXiv* (2019). DOI: arXiv:1810.08595v1.
A.T. participated in the conception of the project, performed the analysis, and participated in the writing of the manuscript.

[TC18] A. Taeb and V. Chandrasekaran. "Interpreting Latent Variables in Factor Analysis via Convex Optimization". In: *Mathematical Programming* 167.1 (2018), pp. 129–154. DOI: 10.1007/s10107-017-1187-7.
A.T. participated in the conception of the project, performed the analysis, and participated in the writing of the manuscript.

[Tae+17] A. Taeb, J.T. Reager, M. Turmon, and V. Chandrasekaran. "A Satistical Model of the California Reservoir System". In: *Water Resources Research* 53.11 (2017), pp. 9721–9739. DOI: 10.1002/2017WR020412.
A.T. participated in the conception of the project, developed the methods, performed the data analysis, and participated in the writing of the manuscript.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Number *Page*

LIST OF TABLES

Chapter 1

INTRODUCTION

An overarching challenge in science and engineering is to develop concise and interpretable frameworks that characterize the relationships among a large collection of variables. As an example, in computational biology, a common scientific discovery involving a gene regulatory network is to determine how variation in one gene impacts the others genes in the network. In water resources, a complete understanding of the relationship among the different water entities in a network provides an important tool to enforce effective and sustainable policies. Finally, in imaging spectroscopy, characterizing the relationship among the spectral profile of patches in a scene is crucial for the accuracy of existing detection techniques (e.g., matched filters). A significant difficulty that arises with finding the statistical dependencies among a collection of variables is that we do not have sample observations of some of the relevant variables. These latent (hidden) variables complicate finding a concise representation, as they introduce confounding dependencies among the variables of interest. Consequently, significant efforts over many decades have been directed towards the problem of accounting for the effects of latent phenomena in statistical modeling via latent-variable techniques. Commonly employed latent-variable models include factor analysis, latent dirichlet allocation, mixture distributions, latent-variable graphical models, etc.

While the topic of latent-variable modeling has been widely studied in statistics, computer science, and optimization, many outstanding challenges remain at the interface of estimation, inference, and computation. Specifically, many latent-variable techniques rely on the data being generated from Gaussian distribution, which may be fundamentally incorrect in many applications. How do we reliably estimate parameters of a latent-variable model from data of various types? With respect to inference, how do we guarantee that our obtained model is an accurate representation of a physical phenomena given finite sample size? Finally, the data we face is often high-dimensional with a large number of observations. Can we develop optimization algorithms that scale and ensure that their solutions are statistically consistent? This thesis is an attempt at addressing these challenges with a strong emphasis on applications. Specifically, many of the specific problems that are tackled in thesis thesis are motivated by an application in water resources, which is described next.

1.1 Motivating Application

An application that has motivated many of the methodological advances made in this thesis is the California reservoir network. The reservoir network, consisting of 1530 reservoirs, is California's major defense against severe droughts, which is frequently experienced in the state. For four years, from 2012 to 2015, California was in a state of severe drought on par with the worst periods in the past 1,200 years [Agh+14]. The impact of the drought was exacerbated by a fundamental limitation in our ability to predict water levels in reservoirs. Which reservoir will dry up first? What is the likelihood of systemic failure (i.e. multiple large reservoirs exhausting)? Answering these sorts of questions would enable policy makers and water managers to mitigate the damage caused to California's 40 million residents.

Previous analysis has focused on the behaviour of a small collection of reservoirs using physical laws or via empirical techniques. Due to the size and complexity of the reservoir networks, these approaches have been difficult to carry out. Specifically, a challenge that must be overcome is to understand the influence of external factors on the reservoirs, as well as reservoir interdependencies, since one reservoir failing can negatively affect other reservoirs in the network. The external factors that strongly affect reservoirs may be measurable phenomena (e.g. precipitation and temperature), or hard-to-quantify influence of human operator. In other words, the external factors may be unobserved or latent. To that end, in Chapter 2, we employ a class of latent variable models, known as latent-variable graphical modeling, where the graph connections encode reservoir dependencies and the latent variables account for the external factors. All of these components are learned from data and are utilized to characterize the system-wide response of reservoirs. With this model in hand, we obtain a clearer picture of the demands placed on reservoirs during drought, and propose a practical guideline for policies that can lead to more sustainable water resources. The results of Chapter 2 correspond to the paper [Tae+17].

To the best of our knowledge, this is the first state-wide model of California reservoirs. However, we believe that even better models could be obtained by addressing some of the limitations of latent-variable modeling techniques that will not only serve reservoirs but other applications as well. More specifically, selecting a model from latent-variable modeling techniques often require tuning hyperparameters. In the reservoir work, these parameters were chosen via a heuristic

known as cross-validation, which does not offer any theoretical guarantees. Next, many latent-variable approaches assume that the variables of interest are Gaussian to produce computationally appealing and statistically accurate procedures. With reservoirs, we overcome this limitation by averaging daily observations to obtain monthly volumes, and validate that the Gaussianity assumption is reasonable after this preprocessing. Furthermore, latent-variable modeling techniques often assume that all observations are identically and independently distributed. With reservoirs, we removed seasonality and verified that the dependencies between observations is substantially reduced. Finally, latent variables produced from data are typically mathematical objects that without semantics. In the reservoir work, we employ a simple post-hoc correlation analysis to link potentially relevant auxiliary variables to learned latent variables to find matches.

1.2 Methodological Contributions

Motivated by these limitations, this thesis provides the following methodological contributions: an inference procedure to ensure that the model draws accurate inferences about some underlying phenomena, a convex optimization technique to identify a latent-variable graphical model for non-Gaussian variables, a mathematically rigorous and computationally efficient approach to handle non-iid and high-dimensional data, and finally, a convex optimization procedure to provide semantics to latent-variables. Throughout this thesis, we will explore how these methodologies will not only benefit reservoir modeling, but also applications in hyperspectral imaging, social networks, and collaborative filtering.

Below, we provide more details of the contributions of this thesis beyond the reservoir analysis. Details about related previous work are given in the relevant chapters. The research and results of Chapters 3 and 6 correspond to completed papers [TSC19] and [TC18], respectively. The work in Chapters 4 and 5 correspond to papers that are in preparation.

Chapter 3 - Inference in Low-rank Estimation Low-rank models are ubiquitous in latent variable modeling pipelines. In many of the applications in which they are employed, the row/column spaces of the low-rank matrix have some physical meaning or represent discoveries. As an example, with reservoirs in Chapter 2, they encode the effect of external factors (latent variables) on reservoirs; in hyperspectral imaging, they represent signature materials of an underlying scene; in radar, they

represent the direction of moving targets. Given the importance of row/column space structures, how do we evaluate the extent to which our model has learned true or false discoveries about the relevant phenomena?

A common approach to statistical model selection – particularly in scientific domains in which it is of interest to draw inferences about an underlying phenomenon – is to develop powerful procedures that provide control on *false discoveries*. Such methods are widely used in inferential settings involving variable selection, graph estimation, and others in which a discovery is naturally regarded as a discrete concept. However, this view of a discovery is ill-suited to many model selection and structured estimation problems in which the underlying decision space is not discrete. We describe a geometric reformulation of the notion of a discovery, which enables the development of model selection methodology for a broader class of problems. We highlight the utility of this viewpoint in problems involving subspace selection and low-rank estimation, with a specific algorithm to control for false discoveries in these settings. Concepts from algebraic geometry (e.g. tangent spaces to determinantal varieties) play a central role in the proposed framework.

Chapter 4 - Latent Variable Graphical Modeling: Beyond Gaussianity The algorithm to fit a latent-variable graphical model to reservoir volumes in Chapter 2 is appropriate when the variables are Gaussian. In many scientific and engineering applications, the set of variables one wishes to model strongly deviate from Gaussianity. Existing techniques to fit a graphical model to data suffer from one or more of these deficiencies: a) they are unable to handle non-Gaussianity, b) they are based on non-convex or computationally intractable algorithms, and c) they cannot account for latent variables. We develop a framework, based on Generalized Linear Models, that addresses all these shortcomings and can be efficiently optimized to obtain provably accurate estimates. A particularly novel aspect of our formulation is that it incorporates regularizers that are tailored to the type of latent variables: nuclear norm for Gaussian latent variables, max-2 norm for Bernoulli variables, and complete positive norm for Exponential variables. For each case, we provide a semidefinite relaxation and demonstrate that the associated norm yields a better sample complexity (than the nuclear norm) for similar computational cost. We further demonstrate the utility of our approach with data involving U.S. Senate voting record.

Chapter 5 - Model Selection with non-iid Data The data we observe and process is typically both non-iid and high-dimensional. As an example, reservoir volumes in

Chapter 2 exhibit significant temporal correlations so that the data is non-iid, and the reservoir network is large so that the data is high-dimensional. Existing techniques that model such complex datasets require $O(n^2p^6)$ computations (n : number of observations, p: number of variables), which is a significant bottleneck for large n; p. By appealing to ideas from Stochastic Partial Differential Equations (SPDE) and covariance selection, we provide a framework that blends temporal/spatial and network modeling in $O(np^2+n\log(n)p+p^6)$ computations. Using this methodology, we are able to efficiently obtain high-dimensional models with rich dependencies across observations. We apply our approach to signature detection in hyperspectral imaging and demonstrate improved performance over existing techniques.

Chapter 6 - Interpreting Latent Variables Via Convex Optimization Factor analysis is a prominent multivariate statistical modeling approach to identify the effects of (a small number of) latent variables on a set of observed variables. However, the latent variables in a factor model are purely mathematical objects that are derived from the observed phenomena, and they do not have any interpretation associated to them. A natural approach for attributing semantic information to the latent variables in a factor model is to obtain measurements of some additional plausibly useful covariates that may be related to the original set of observed variables, and to associate these auxiliary covariates to the latent variables. In this paper, we describe a systematic approach for identifying such associations. Our method is based on solving computationally tractable convex optimization problems, and it can be viewed as a generalization of the minimum-trace factor analysis procedure for fitting factor models via convex optimization. We analyze the theoretical consistency of our approach in a high-dimensional setting as well as its utility in practice via experimental demonstrations with real data.

Chapter 2

LATENT VARIABLE GRAPHICAL MODELING WITH APPLICATION TO RESERVOIR MODELING

As described in Chapter 1, many of the research questions that are tackled in this thesis stem from an application involving the statistical modeling of the California reservoirs. In this chapter, we will dive deep into this application, discuss the challenges that arise from modeling a system of reservoirs, and propose latent-variable methodologies that address these challenges. The results of this chapter are published in [Tae+17] and were developed jointly with John Reager, Michael Turmon and Venkat Chandrasekaran. The author contributed by performing data preprocessing, developing modeling framework & algorithms to analyze the data, and implementing the numerical methods to produce the final results. The description of the work contained in this chapter was written by the author.

2.1 Introduction

Motivation

The state of California depends on a complex water management system to meet wide-ranging water demands across a large, hydrologically diverse domain. As part of this infrastructure, California has constructed 1530 reservoirs having a collective storage capacity equivalent to a year of mean runoff from California rivers [Gra99]. The purpose of this system is to create water storage capacity and extend seasonal water availability to meet agricultural, residential, industrial, power generation, and recreational needs.

Major statewide California precipitation deficits during the years 2012—-2015 rivalled the most intense 4-year droughts in the past 1200 years [GA14]. The drought was punctuated by low snowpack in the Sierra Nevada, declining groundwater storage, and fallowed agricultural lands, in addition to significantly diminished reservoir levels [Agh+14; Fam14; How+14]. This sensitivity of the California reservoir network to external conditions (e.g. temperature, precipitation) has implications for statewide water and agricultural security. In this chapter, we seek a characterization of the relationships among the major California reservoirs and their sensitivity to statewide physical and economic factors, with a view to investigating and quantifying the likelihood of systemic catastrophes such as the simultaneous exhaustion of

multiple large reservoirs.

Such an analysis has been difficult to carry out on a system-wide scale due to the size and complexity of the reservoir network. In one direction, a body of work has focused on characterizing the behavior of a small collection of reservoirs using *physical laws* (e.g. [CL04; Chr+06; NW15]). Such approaches quickly become intractable in settings with large numbers of reservoirs whose complex management is based on multiple economic and sectoral objectives [How+14]. The hard-to-quantify influence of human operators and the lack of system closure have made the modeling and prediction of reservoir network behavior using physical equations challenging in hydrology and climate models [Sol+16]. In a different direction, numerous works have developed *empirical techniques* for modeling the behavior of a small number of reservoirs (e.g. [RW83; Pha89; NG91; NG93; BHH03; HE07; BP08; Wis+10; Che+15]). However, these methods are not directly applicable to modeling a large reservoir network, as the water levels of major reservoirs in California exhibit complex interactions and are statistically correlated with one another (as is demonstrated by our analysis). This necessitates a proper quantification of the complex dependencies among reservoirs in determining the systemic characteristics of the reservoir network.

The focus of this work is to develop a statewide model over the California reservoir network that addresses the following scientific questions:

1. What are the interactions or dependencies among reservoir holdings? In particular, how correlated are major reservoirs in the system?

2. Are there common external factors influencing the network globally? Could these external drivers cause a system-wide catastrophe?

To the best of our knowledge, our work is the first that attempts such a statewide characterization of the California reservoir network. The statewide external factors that we consider in our analysis include physical factors such as statewide PDSI and average temperature, and economic factors such as the consumer price index and the number of agricultural workers. The focus on these statewide external influences is driven by the global nature of our analysis; indeed, an exciting direction for further research is to complement our global model with local reservoir-specific factors to obtain an integrated picture of both systemic as well as local risks to the reservoir network.

Answering these questions for the California reservoir system raises a number of challenges, and it is important that any modeling framework that we consider addresses these challenges. First, reservoirs with similar hydrological attributes (e.g. altitude, drainage area, spatial location) tend to behave similarly. As an example, a pair of reservoirs that is approximately at the same altitude or in the same hydrological zone are more likely to have a stronger correlation than those in different altitudes / zones. Therefore, we seek a framework that ably models the complex heterogeneities in the reservoir system. A second challenge, which is in some sense in competition with the first one, is that compactly specified models are much more preferable to less succinct models, as concisely described models are often more interpretable and avoid problems associated with *over-fitting*. Finally, it is crucial that models with both of the preceding attributes have the additional feature that they can be identified in a computationally efficient manner.

Approach and Results

Gaussian *graphical models* offer an appealing and conceptually powerful framework with all the attributes just described. Graphical modeling is a prominent multivariate analysis technique that has been successfully employed in domains as varied as gene regulatory network analysis, social networks, speech recognition, and computer vision (see [Jor04] for a survey on graphical modeling). These models are defined with respect to graphs, with nodes of a graph indexing variables and the edges specifying statistical dependencies among these variables. In a reservoir modeling context, the nodes of the graph correspond to reservoirs and an edge between two reservoirs would describe the strength of the interaction between the levels of those reservoirs. Formally, the strength of an edge specifies the degree of conditional dependence between the corresponding reservoirs; in other words, this is the dependence between two reservoirs conditioned on all the other reservoirs in the network. Informally, an edge in a graphical model denotes the extent to which two reservoirs remain correlated even after accounting for the influence of all the other reservoirs in the network. We illustrate these points using a toy example of a graphical model over a collection of 8 reservoirs, shown in Figure 2.1(a). (This figure is purely for explanatory purposes rather than a factual representation of the complex dependencies among reservoirs, which we obtain in Section 3.) One can imagine that the reservoir volumes of Shasta (which is at a high elevation in northern California in the Sacramento hydrological zone) are independent of the reservoir Pine Flat and the reservoir Isabella (which are in southern California in the Tulare

hydrological zone) after conditioning on volumes of reservoirs in the central portion of the state (e.g. Black Butte, Lake Berrysa, New Melones, Buchanan, and Don Pedro). These relationships are encoded in a graphical model of Figure 1(a). In particular, note that Shasta has an edge linking it to each of the reservoirs {Black Butte, Lake Berrysa, Don Pedro, New Melones, Buchanan}, but does not have an edge connecting it to the reservoirs {Pine Flat, Isabella}. Figure 1(a) is, of course, a cartoon demonstration of a graphical modeling framework. In practice, identifying conditional dependencies between pairs of reservoirs in large networks such as the one considered in our work is a challenging problem, and we describe tractable approaches to learning such a graphical structure underlying the complex California reservoir system in a completely data-driven manner in Section 3. To the best of our knoweledge, this is the first work that applies graphical modeling techniques to model reservoirs or other water resources.

The graphical modeling framework provides a common lens for viewing two frequently employed statistical techniques. On the one hand, a classical approach for obtaining a multivariate Gaussian distribution over reservoir volumes is via a maximum likelihood estimator. This estimator has been widely used in various domains in the geophysical sciences for multivariate analysis of a collection of random variables [Wac03]. The model obtained by this maximum likelihood estimator is specified by a completely connected graphical structure, where all reservoirs are conditionally correlated given all other reservoirs. On the other hand, an independent reservoir model analyzes the behavior of an individual reservoir independently of the other reservoirs in the network. This model results in a fully disconnected graphical model. In this chapter, we learn a statistical graphical model over the reservoir network in a data-driven manner based on historical reservoir data. This model yields a sparse (yet connected) graphical structure describing the network interactions. We demonstrate that this model outperforms the model obtained via classical maximum likelihood estimator and an independent reservoir model. Thus, the reservoir behaviors are not independent of one another but can be specified with a moderate number of interactions. We demonstrate that a majority of these interactions are between reservoirs that are in the same basin or hydrological zone, and among reservoirs that have similar altitude and drainage area.

A natural question is whether some dependencies specified by the graphical model are due to a small number of external phenomena (drought, agricultural usage, Colorado river discharge, precipitation, etc.). For example, water held by a

collection of nearby reservoirs might be influenced by a common snowpack variable. Without observing this common variable, all reservoirs in this set would appear to have mutual links, whereas if snowpack is included in the analysis, the common behavior is explained by a link to the snowpack variable. Accounting for latent structure removes these *confounding* dependencies and leads to *sparser and more localized* interactions between reservoirs. Figure 2.1(b) illustrates this point. Latent variable graphical modeling offers a principled approach to quantify the effects of *external phenomena* that influence the entire reservoir network. In particular, this modeling framework uses observational data to (a) identify the number of global drivers (e.g. latent variables) that summarize the effect of external phenomena on the reservoir network, and (b) identify the residual reservoir dependencies after accounting for these global drivers. Our experimental results demonstrate that the reservoir network at a monthly resolution has two distinct global drivers, and residual dependencies persist after accounting for these global variables.

Latent variable graphical modeling obtains a mathematical representation of the global drivers of the reservoir network. One is naturally interested in linking these mathematical objects to real world signals (e.g. statewide Palmer Drought Severity Index, snowpack, consumer price index). We present an approach for associating semantics to these global drivers. We find that the statewide Palmer Drought Severity Index (PDSI) is highly correlated ($\rho \approx 0.88$) with one of the global drivers. PDSI is then included as a covariate in the *next* iteration of the graphical modeling procedure to learn a joint model over reservoirs and PDSI. Using this model, we characterize the system-wide behavior of the network to hypothetical drought conditions. In particular, we find that as PDSI approaches -5, there is a probability greater than 50% of simultaneous exhaustion of multiple large reservoirs. We further present an approach for identifying specific reservoirs in the network that are at high risk of exhaustion during extreme drought conditions. We find that the Buchanan and Hidden Dam reservoirs are at high risk and describe water management policies and practices that were enforced to prevent exhaustion.

2.2 Dataset and Model Validation

Our primary dataset consists of monthly averages of reservoir volumes, derived from daily time series of volumes downloaded from the California Data Exchange Center (CDEC). We also used secondary data for some covariates.

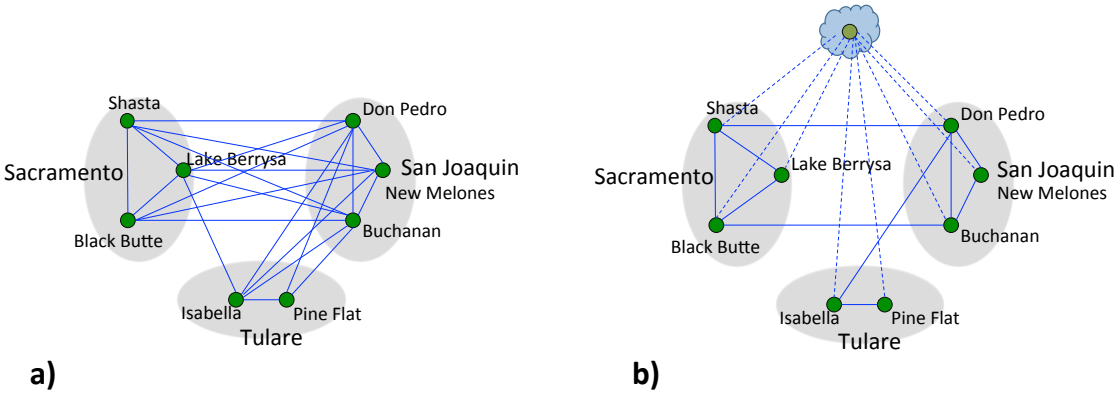


Figure 2.1: Graphical structure between a collection of 8 reservoirs without latent variables (a) and with latent variables (b). Green nodes represent reservoirs (variables) and the clouded green node represents latent variables. Solid blue lines represent edges between reservoirs and dotted edges between reservoirs and latent variables. The reservoirs have been grouped according to hydrological zones.

Reservoir Time Series & Preprocessing Techniques

As described in Section 1, there are 1,530 reservoirs in California. In this work, we perform statistical analysis on the largest 60 reservoirs in California. We apply our analysis to a subset of the reservoirs, as they have a large amount of historical data available. Our technique can be extended to a larger collection of reservoirs given sufficient data. For these 60 reservoirs, daily volume data is available during the period of study (January 2003 — November 2016). We excluded five reservoirs with more than half of their values undefined or zero, leaving 55 reservoirs. This list of daily values was inspected using a simple continuity criterion and approximately 50 specific values were removed or corrected. Corrections were possible in six cases because values had misplaced decimal points, but all other detected errors were set to missing values. The most common error modes were missing values that were recorded as zero volume, and a burst of errors in the Lyons reservoir during late October 2014 that seems due to a change in recording method at that time.

The final set of 55 reservoir volume time series spans 5083 days over the 167 months in the study period. It contains two full cycles of California drought (roughly, 2007 — 2008 and 2012 — 2015) and three cycles of wet period (2004 — 2006, 2009 — 2011, 2016). Four California hydrological zones are represented, with 25, 20, 6, and 4 reservoirs in the Sacramento, San Joaquin, Tulare, and North Coast zones, respectively.

We are interested in long-term reservoir behavior and thus model reservoir vol-

umes at a monthly time scale. In particular, we average the data from daily down to 167 monthly observations. The reservoir data exhibit strong seasonal components. As such, a seasonal adjustment step is performed to remove these predictable patterns, so that we can model deviations from the underlying trend in the reservoir behavior. Specifically the steps are as follows Let $\{\bar{y}^{(i)}\}_{i=1}^{n_{\text{train}}} \subset \mathbb{R}^{55}$ and $\{\bar{y}^{(i)}\}_{i=1}^{n_{\text{test}}} \subset \mathbb{R}^{55}$ be the averaged monthly reservoir volumes in the training and validation set respectively. Focusing on a reservoir r and the month of January, let $\mu_{\bar{y}_r}$ be the average reservoir level during January (obtained only from training observations). For each observation i in January, we apply the transformation: $\tilde{y}_r^{(i)} = \bar{y}_r^{(i)} - \mu_{\bar{y}_r}$. We repeat the same steps for all months. Furthermore, letting σ_r be the sample standard deviation of the training observations $\{\tilde{y}_r^{(i)}\}_{i=1}^{n_{\text{train}}}$, we produce unit variance observations with the transformation,$y_r^{(i)} = \frac{1}{\sigma_r^{1/2}} \tilde{y}_r^{(i)}$. Before being used in the fitting algorithms, each time series is also rescaled by its standard deviation so that each series has unit variance. We note that our statistical approach identifies correlations between reservoir volumes. Since correlation between two random variables is normalized by their respective variances, this transformation is appropriate. We repeat the same steps for all reservoirs to obtain the preprocessed reservoir observations $\{y^{(i)}\}_{i=1}^{n_{\text{train}}}$ and $\{y^{(i)}\}_{i=1}^{n_{\text{test}}}$.

With the exception of the Farmington reservoir (which has volume less than 10^8 m^3), the joint volume anomalies of the remaining 54 reservoirs (after preprocessing) are well-approximated by a multivariate Gaussian distribution. This is demonstrated by a Q-Q plot in Figure 2.2. Since a large amount of historical data is available for the Farmington reservoir, we have included it in our analysis. These observed properties suggest that the reservoir data is amenable to the multivariate Gaussian models we employ in this chapter.

Covariate Time Series

Latent variable graphical modeling identifies a mathematical representation of the global drivers of the reservoir network. We link these global drivers to real-world signals using ancillary data, i.e., *covariates*, which are observable variables, exogenous to the model, that may affect a large fraction of reservoirs. The particular covariates that we use are temperature (averaged values over California downloaded from NOAA), Palmer Drought Severity Index (averaged values over California downloaded from NOAA), hydroelectric power generation of California (downloaded from U.S. Energy Information and Administration), Colorado river discharge (averaged values downloaded from United States Geological Survey), and

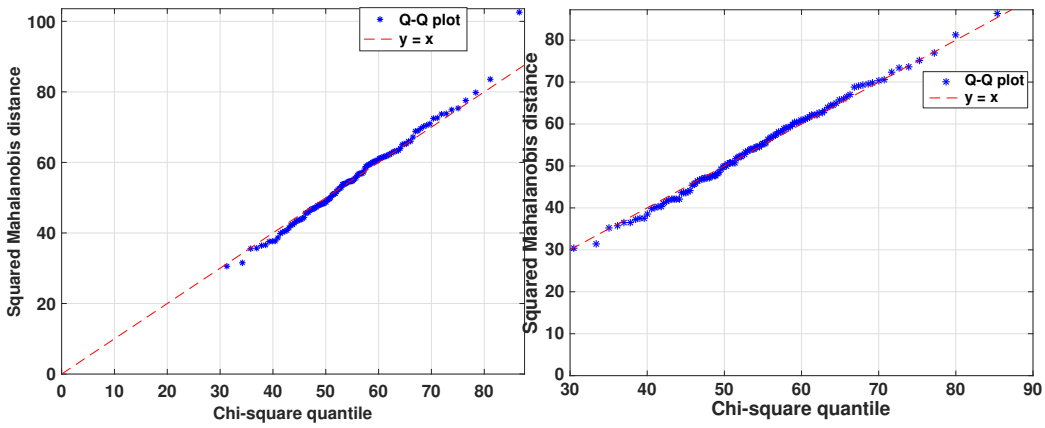


Figure 2.2: (a): Q-Q plot of the entire set of 55 reservoirs. (b): Q-Q plot of 54 reservoirs (excluding the Farmington reservoir). The Q-Q plots are against a multivariate Gaussian distribution. Notice that $y = x$ is a close approximation to the Q-Q plot implying that 54 reservoirs (excluding Farmington reservoir) are well approximated by a multivariate Gaussian distribution.

Sierra Nevada snow pack covariate (manually averaged in the Sierra Nevada region where the elevation is over 100 m, gridded observations downloaded from NOAA). Note that since we are interested in statewide covariates that exert influence over the entire network, these hydrological indicators were averaged over the state of California (or in the case of snowpack and Colorado river discharge, averaged over a large region in the Sierra Nevada and Colorado river respectively). In addition to these hydrological indicators, we use the following economic factors: statewide number of agricultural workers (downloaded from State of California Employment Development Department) and statewide consumer price index (downloaded from Department of Industrial Relations).

For each of the 7 covariates, we obtain averaged monthly observations from 2003—2016. We apply a time lag of two months to the covariates temperature, snowpack, Colorado river discharge, and Palmer Drought Severity Index (the reason for a two months lag is explained in Section 4.4). As with the reservoir time series, we remove seasonal patterns with a per-month average, and rescale to obtain unit variance variables.

Model Validation

To ensure that the model of the reservoirs is representative of reservoir behavior, we perform model validation using a technique known as *holdout validation*

[HTF09]. The objective of this technique is to produce models that are not overly tuned to the idiosyncrasies of observational reservoir data, so that these models are representative of future reservoir behavior. In a holdout validation framework, the available data is partitioned into a training set, and a disjoint validation set. The training set is used as input to a fitting algorithm to identify a model. The accuracy of this model is then validated by computing the average log-likelihood of the validation set with respect to the distribution specified by the model. Here, larger values of log-likelihood are indicative of better fit to data. For our experiments, we set aside monthly observations of reservoir volumes and covariates from January 2004 — December 2013 as a training set ($n_{\text{train}} = 120$) and monthly observations from January 2003 — December 2003 and January 2014 — November 2016 as a (disjoint) validation set ($n_{\text{test}} = 47$). Both the training and validation observations contain a signficiant amount of annual and inter-annual variability.

2.3 Dependencies Underlying the Reservoir Network

Method: Graphical Modeling

A common approach for fitting a graphical model to reservoirs is to choose the simplest model, that is, the sparsest network that adequately explains the observational data. Easing this taks, for Gaussian graphical models, the graphical structure is encoded in the sparsity pattern of the precision matrix (inverse covariance matrix) over the variables. Specifically, zeros in the precision matrix of a multivariate Gaussian distribution indicate absent edges in the corresponding graphical model. Thus, the number of edges in the graphical model equals the number of nonzeros of the precision matrix Θ. As an example, consider the toy graphical model in Figure 1(a). Suppose that the precision matrix Θ of size 8×8 is indexed according to the ordering {Shasta, Black Butte, Lake Beryssa, Isabella, Pine Flat, Don Pedro, New Melones, and Buchanan}. Then Θ has the following structure:

$$\Theta = \begin{pmatrix} \star & \star & \star & 0 & 0 & \star & \star & \star \\ \star & \star & \star & 0 & 0 & \star & \star & \star \\ \star & \star & \star & \star & 0 & \star & \star & \star \\ 0 & 0 & \star & \star & \star & \star & \star & \star \\ 0 & 0 & 0 & \star & \star & \star & 0 & \star \\ \star & \star & \star & \star & \star & \star & \star & \star \\ \star & \star & \star & \star & 0 & \star & \star & \star \\ \star & \star & \star & \star & \star & \star & \star & \star \end{pmatrix},$$

where $\star$ denotes a nonzero value. The intimate connection between a graphical structure and the precision matrix implies that fitting a sparse Gaussian graphical model to reservoir observational data is equivalent to estimating a sparse precision matrix Θ. Thus, the reservoirs are modeled according to the distribution $y \sim \mathcal{N}(0\,,\ \Theta^{-1})$, where Θ is sparse. Note that the preprocessing to remove climatology causes the mean to be zero. A natural technique to fit such a model to observational data is to minimize the negative log-likelihood (e.g. maximum likelihood estimation) of data while controlling the sparsity level of Θ. The log-likelihood function of the training observations $\mathcal{D}_{\text{train}} = \{y^{(i)}\}_{i=1}^{n_{\text{train}}} \subset \mathbb{R}^{55}$ (after removing some additive constants and scaling) is given by the concave function

$$\ell(\Theta; \mathcal{D}_{\text{train}}) = \log\det(\Theta) - \operatorname{tr}\left[\Theta \cdot \Sigma_n\right] \quad , \tag{2.1}$$

where $\Sigma_n = \frac{1}{n_{\text{train}}} \sum_{i=1}^{n_{\text{train}}} y^{(i)} y^{(i)\prime}$ is the sample covariance matrix. Thus, fitting a graphical model to $\mathcal{D}_{\text{train}}$ translates to searching over the space of precision matrices to identify a matrix Θ that is sparse and also yields a small value of $-\ell(\Theta; \mathcal{D}_{\text{train}})$. This formulation, however, is a computationally intractable combinatorial problem. Recent work [YL07; FHT08] has identified a way around this road block by using a convex relaxation:

$$\begin{aligned} \hat{\Theta} = \underset{\Theta \in \mathbb{S}^{55}}{\arg\min} \quad & -\ell(\Theta; \mathcal{D}_{\text{train}}) + \lambda \, \|\Theta\|_1 \\ \text{s.t.} \quad & \Theta \succeq 0 \quad . \end{aligned} \tag{2.2}$$

The notation $\mathbb{S}^{55}$ denotes the set of symmetric 55×55 matrices. The constraint $\succ 0$ imposes positive definiteness so that the joint distribution of reservoirs is non-degenerate. The regularization term $\|\cdot\|_1$ denotes the L_1 norm (element-wise sum of absolute values) that promotes sparsity in the matrix Θ. The L_1 penalty, and more broadly, regularization techniques, are widely employed in inverse problems in data analysis to overcome ill-posedness and avoid problems such as *over-fitting* to moderate sample size (see the textbooks/monographs [BG11; Wai14] and the references therein). These regularization approaches have proved to be valuable in many applications, including cameras [Dua+08], magnetic resonance imaging [Lus+08], gene regularity networks [ZK14], and radar [HS09].

The regularization parameter λ in (2.2) provides overall control of the trade-off between the fidelity of the model to the data and the complexity of the model. In particular, the program (2.2) with $\lambda = 0$ yields the familiar maximum likelihood covariance estimator. This estimator has a well-known closed form solution $\hat{\Theta} = \Sigma_n^{-1}$.

Generally, Σ_n^{-1} will not contain any zeros. This implies that the estimated graphical structure is fully connected with close fit to the training data $\mathcal{D}_{\text{train}}$. However, as explored in Section 3.2, this model may be over-tuned to the idiosyncrasies of the training observations $\mathcal{D}_{\text{train}}$ and will not generalize to future behavior of reservoirs (a phenomenon known as *over-fitting*). Larger values of λ yield a sparser graphical model with very large λ resulting in a completely disconnected graphical model where the reservoirs are independent of one another. Importantly, for any choice of $\lambda > 0$, (2.2) is a convex program with a unique optimum, and can be solved efficiently using general purpose off-the-shelf solvers [TTT16]. Further theoretical support of this estimator is presented in [Rav+11b].

We select the regularization parameter λ by *holdout validation*. In particular, for any choice of λ, we supply the training observations $\mathcal{D}_{\text{train}}$ to (2.2) to learn a graphical model and compute the average log-likelihood of this model on the validation set $\mathcal{D}_{\text{test}} = \{y^{(i)}\}_{i=1}^{n_{\text{test}}} \subset \mathbb{R}^{55}$. We sweep over all values of λ to choose the model with the best validation performance. Let the selected model (after holdout validation) be specified by the precision matrix $\hat{\Theta}$. As discussed earlier, the matrix $\hat{\Theta}$ specifies the structural properties of the graphical model of the network. An edge between reservoirs r and r' is present in the graph if and only if $\hat{\Theta}_{r,r'} \neq 0$, with larger magnitudes indicating stronger interactions. We denote the strength of an edge as the normalized magnitude of the precision matrix entry, that is,

$$s(r, r') = |\hat{\Theta}_{r,r'}| \big/ (\hat{\Theta}_{r,r} \hat{\Theta}_{r',r'})^{1/2} \geq 0. \tag{2.3}$$

The quantity $s(r, r')$ can be viewed as the partial correlation between reservoirs r and r', given all other reservoirs. In particular, a large $s(r, r')$ indicates that reservoirs r and r' are highly correlated even after accounting for the influence of all the other reservoirs in the network. A small value of $s(r, r')$ indicates that the reservoirs r and r' are weakly correlated conditioned on all the reservoirs. Finally, $s(r, r') = 0$ indicates that reservoirs r and r' are independent conditioned on all the remaining reservoirs.

Results: Graphical Model of Reservoir Network

In this section, we explore the properties of a graphical model over the reservoir network. As described in Section 3.1, we learn a graphical model by specifying a regularization parameter λ and supplying observations $\mathcal{D}_{\text{train}}$ to the convex program (2.2). We vary λ from 0 to 1 to identify a collection of graphical models. For $\lambda \geq 1$, the graphical model is completely disconnected and not of interest. For each graphical model, we measure the training performance as the log-likelihood of training

observation $\mathcal{D}_{\text{train}}$ and the validation performance as the log-likelihood of validation observations $\mathcal{D}_{\text{test}}$. Figure 2.3 illustrates the training and validation performances for different values of λ. Recall that $\lambda = 0$ corresponds to an unregularized maximum likelihood estimate and $\lambda = 1$ corresponds to independent reservoir model. We chose $\lambda = 0.23$ to obtain a graphical model with the best validation performance. Results of Figure 2.3 demonstrate that the training performance is a decreasing function of λ: smaller values of λ lead to a closer fit to training observations. However, small values of λ yield a high complexity model that fits the idiosyncrasies of the training data and thus suffers from over-fitting. This is evident from the poor validation performance of unregularized ML estimate (when $\lambda = 0$). The graphical model is the superior model since it has a better validation performance than the unregularized ML estimate and an independent reservoir model. Thus the reservoir behaviors are not independent but can be characterized by a moderate number of dependencies. In the supplementary material, we characterize the sensitivity of the graphical model to the choice of the regularization parameter λ.

Model	Training performance	Validation performance
unregularized ML estimate ($\lambda = 0$)	-23.91	-1140.4
independent reservoir model ($\lambda = 1$)	-83.23	-101.95
graphical model ($\lambda = 0.23$)	**-63.52**	**-85.54**

Table 2.1: Training and validation performances of unregularized maximum likelihood (ML) estimate, independent reservoir model, and graphical model. As larger values of log-likelihood are indicative of better performance, the graphical model is the superior model.

To demonstrate that the graphical model estimate does not vary significantly under small perturbations to λ, we also obtain graphical model estimates with $\lambda = 0.26$ and $\lambda = 0.20$ (Recall that the edge strengths in a graphical model contain the relevant information of the model). Figure 2.4(a) compares the edge strengths of the model with $\lambda = 0.23$ and the model with $\lambda = 0.20$. Furthermore, Figure 2.4(b) compares the edge strengths of the model with $\lambda = 0.23$ and the model with $\lambda = 0.26$. Evidently, strong edges persist across all models, with a few weak edges removed or added as λ is varied. The total number of edges in the graphical model when $\lambda = 0.20$, $\lambda = 0.23$, and $\lambda = 0.26$ is 295, 285, and 279 respectively. Furthermore, the quantity κ (defined in equation (4) of main paper) is 0.852, 0.859, and 0.862 for

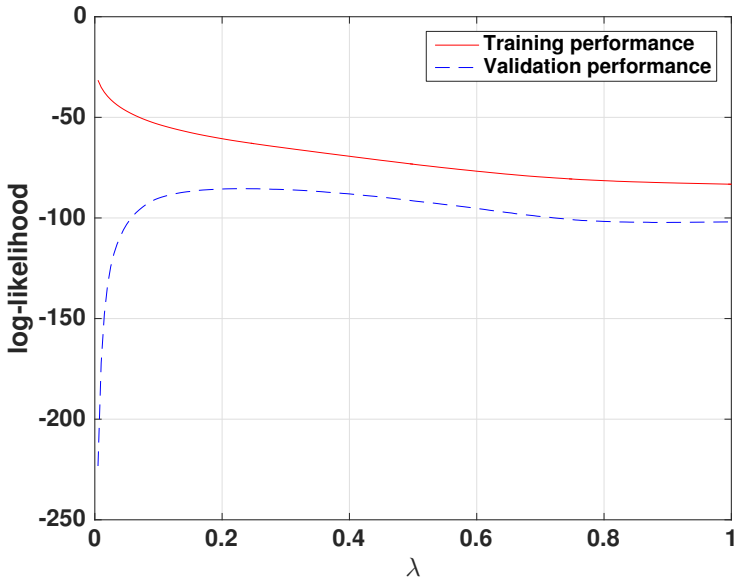


Figure 2.3: Training and validation performance of graphical modeling for different values of the regularization parameter λ. The training performance is computed as the average log-likelihood of training samples and the validation performance is computed as the average log-likelihood of validation samples.

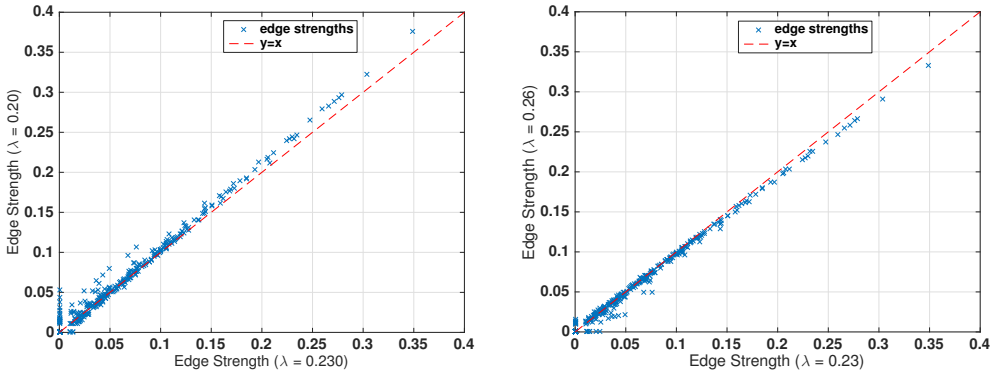


Figure 2.4: Sensitivity of the graphical model estimate to perturbations of λ around the optimal value $\lambda = 0.23$ (this choice of λ leads to optimal validation performance): we observe that strong edges in the original model are strong edges in the perturbed model (i.e., with perturbed λ) with approximately the same strength.

$\lambda = 0.20$, $\lambda = 0.23$, and $\lambda = 0.26$. These results suggest that our conclusions are not particularly sensitive to the choice of the regularization parameter, although we chose $\lambda = 0.23$ as it leads to the best validation performance.

We further explore the properties of the specified graphical model, consisting of 285 edges. Using relation (2.3), we compute the strength of the connections in the graphical structure. The upper triangle of Figure 2.5 shows the dependence

relationships between reservoirs in this graphical model. The five strongest edges in this graphical structure are between reservoirs Relief – Main Strawberry, Cherry – Hetch Hetchy, Invisible Lake – Lake Berryessa, Almanor – Davis, and Coyote Valley – Warm Spring. We show the geographical location of these pairs of reservoirs in Figure . The presence of these strong edges is sensible: each such edge is between reservoirs in the same hydrological zone, and 4 of these 5 edges are between pairs of reservoirs fed by the same river. The five most connected reservoirs in order Folsom Lake, Antelope river, Black Butte River, New Exchequer, and French Meadows, all of which are large reservoirs (volume $\geq 10^8\ m^3$). We show the five strongest connections to Folsom lake in Figure 2.6 , all of which are either connected or are in close proximity to the Sacramento River. As a point of comparison, the lower triangle of Figure 2.5 shows the graphical structure of the unregularized maximum likelihood estimate. This model yields a fully connected network.

Furthermore, we observe that a majority of interactions in this graphical model are among reservoirs that have similar drainage area (e.g. land where water falls off into reservoirs) and elevation. Figure 2.7 (a) shows a plot of the ratios of drainage areas between pairs of reservoirs connected via an edge and the strength of the connections. Figure 4(b) shows a plot of the ratios of altitudes between pairs of connected reservoirs and the strength of the connections. As a point of comparison, Figures 2.7(c) and 2.7(d) show similar metrics for the unregularized maximum likelihood estimate. Examining Figure 4, we observe that graphical modeling removes (or weakens) dependencies between reservoirs of vastly different drainage area or elevation. This is expected since reservoirs with substantially different drainage area or elevation are less likely to have similar variability.

We observe that a large portion of the strong interactions occur between reservoirs in the same hydrological zone, here denoted $h(r)$. To quantify this observation, we consider

$$\kappa = \frac{\sum_{r,r' \text{ and } h(r)=h(r')} s(r,r')}{\sum_{r,r'} s(r,r')} \quad , \tag{2.4}$$

the ratio of within-zone edge strength to total edge strength. The model we fit has $\kappa = 0.85$, so 85% of the total edge strength is between reservoirs in the same hydrological zones. In comparison, $\kappa = 0.46$ for an unregularized maximum likelihood estimate. Nevertheless, we notice some surprising connections between reservoirs that are geographically far apart. In the next section, we propose a framework to quantify the influence of external phenomena on the reservoir network.

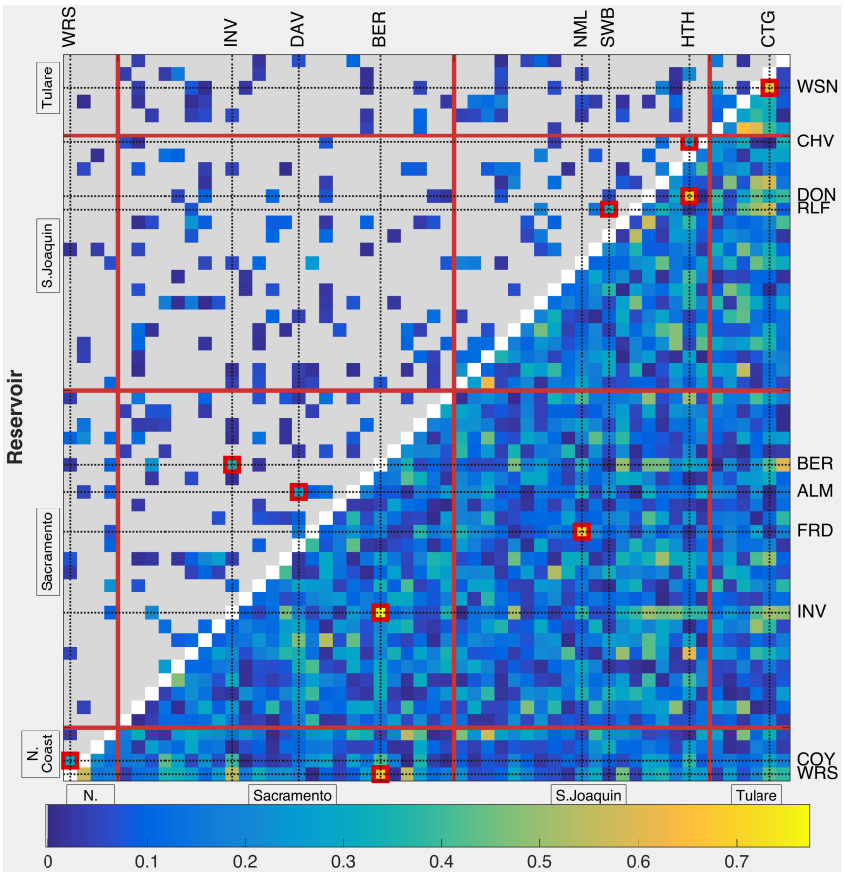


Figure 2.5: Linkages between reservoir pairs in the graphical model (upper triangle) compared with those of the unregularized maximum likelihood estimate (lower triangle). Connection strength $s(r, r')$ is shown in the image map, with unlinked reservoir pairs drawn in gray. The four hydrological zones are separated by red lines. Red boxes surround the five strongest connections in each model.

We further explore the effect of these external phenomena to remove the confounding relationships between geographically distant reservoirs.

2.4 Global Drivers of the Reservoir Network

We identified a graphical model over California reservoirs. Could some of these dependencies specified by the graphical model be due to external phenomena (e.g. global drivers)? In this section, we describe an approach, known as *latent variable graphical modeling*, that identifies the number and effect of global drivers on the reservoir network. Since these global drivers are not directly observed (although we later discuss an approach to link global drivers to real-world signals), we also denote them as *latent variables*.

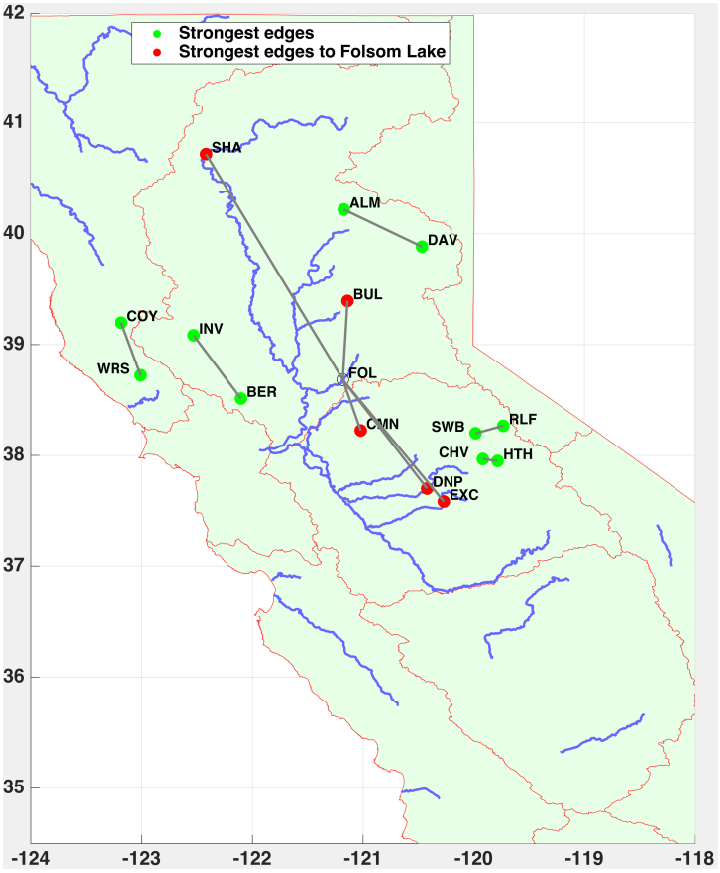
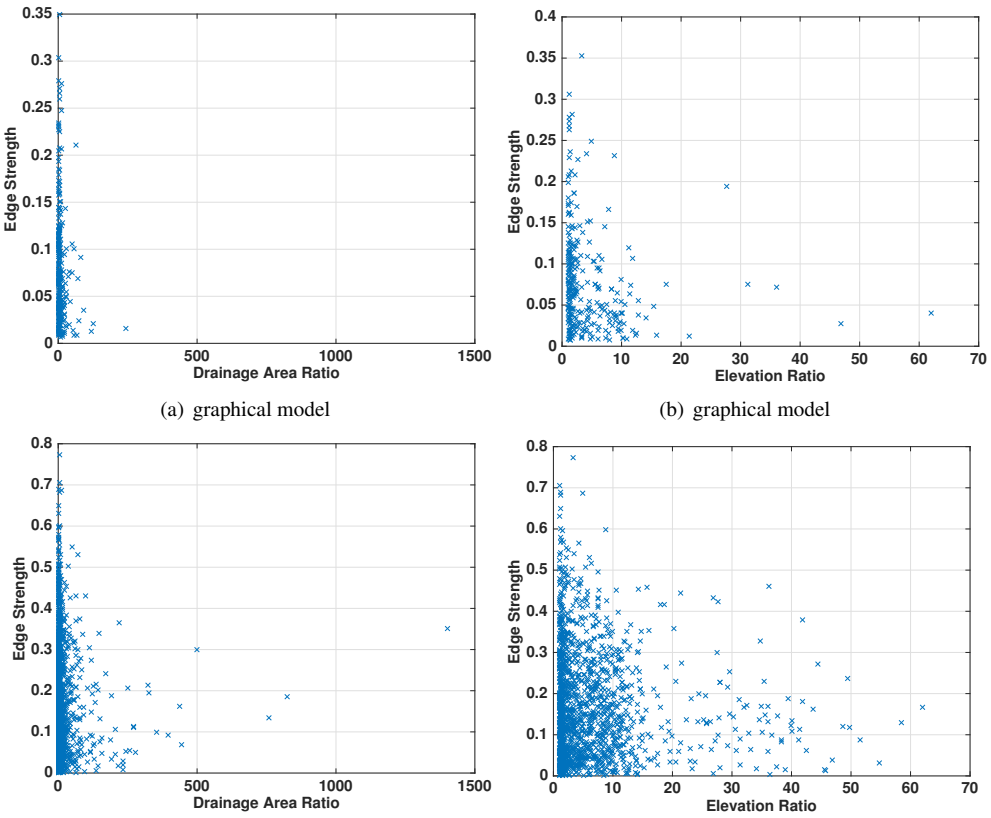


Figure 2.6: A schematic of California and its river network with some reservoir connections. Green nodes represent the 5 pairs of reservoirs with strongest edge strength in the graphical model. The red nodes represent the five strongest edges to Folsom Lake, which is the most connected reservoir in the network. The acronyms for the reservoirs are: WRS (Wishon), COY (Coyote Valley), INV (Indian Valley), BER (Lake Berryessa), SHA (Shasta), BUL (Bullards Bar), FOL (Folsom Lake), CMN (Camanche), DNP (Don Pedro), EXC (New Exchequer), ALM (Almanor Lake), DAV (Lake Davis), SWB (Main Strawberry), RLF (Relief), CHV (Cherry Valley), and HTH (Hetch-Hetchy).



(a) graphical model

(b) graphical model

(c) unregularized ML estimate

(d) unregularized ML estimate

Figure 2.7: a) Ratios of drainage areas between pairs of reservoirs connected with an edge and their corresponding edge strengths in a graphical model. b) Ratios of elevations of pairs of reservoirs connected with an edge and their corresponding edge strengths in a graphical model. c) Ratios of drainage areas between pairs of reservoirs connected with an edge and their corresponding edge strengths in an unregularized maximum likelihood (ML) estimate. d) Ratios of elevations of pairs of reservoirs connected with an edge and their corresponding edge strengths in an unregularized maximum likelihood (ML) estimate.

Method: Latent Variable Graphical Modeling

As shown by [CPW12], fitting a latent variable graphical model corresponds to representing the precision matrix of the reservoir volumes Θ as the difference $\Theta = S - L$, where S is sparse, and L is a low rank matrix. The matrix L accounts for the effect of external phenomena, and its rank is equal to the number of global drivers; these global drivers summarize the effect of external phenomena on the reservoir network. The matrix S specifies the residual conditional dependencies among the reservoirs after extracting the influence of global drivers. Moreover, the sparsity pattern of S encodes the residual graphical structure among reservoirs. As an example, consider the toy model shown in Figure 2.1(b). Suppose that the matrix S is indexed according to the ordering {Shasta, Black Butte, Lake Berrysa, Isabella, Pine Flat, Don Pedro, New Melones, and Buchanan}. Then S has the structure:

$$S = \begin{pmatrix} \star & \star & \star & 0 & 0 & 0 & 0 & \star \\ \star & \star & \star & 0 & 0 & \star & 0 & 0 \\ 0 & 0 & 0 & \star & \star & 0 & 0 & 0 \\ 0 & 0 & 0 & \star & \star & 0 & 0 & \star \\ 0 & 0 & 0 & \star & \star & 0 & 0 & 0 \\ 0 & \star & 0 & 0 & 0 & \star & \star & \star \\ 0 & 0 & 0 & 0 & 0 & \star & \star & \star \\ \star & 0 & 0 & \star & 0 & \star & \star & \star \end{pmatrix},$$

where $\star$ denotes a nonzero entry. Fitting a latent variable graphical model to reservoir volumes is to identify the simplest model, e.g. smallest number of global drivers and sparsest residual network, that adequately explains the data. In other words, we search over the space of precision matrices Θ that can be decomposed as $\Theta = S - L$ to identify a matrix S that is sparse, a matrix L that has a small rank, and also yields a small negative log-likelihood $-\ell(\mathcal{D}_{\text{train}}, S - L)$. As with the case of graphical modeling, this formulation is a computationally intractable combinatorial problem. Based on a recent work by [CPW12], a computationally tractable estimator is given by:

$$\begin{aligned} (\hat{S}, \hat{L}) = \underset{S,\, L \in \mathbb{S}^{55}}{\arg\min} \quad & -\ell(S - L; \mathcal{D}_{\text{train}}) + \lambda(\|S\|_1 + \gamma \operatorname{tr}(L)) \\ \text{s.t.} \quad & S - L \succ 0,\ L \succeq 0 \quad . \end{aligned} \tag{2.5}$$

The constraint $\succ 0$ imposes positive definiteness on the precision matrix estimate $S - L$ so that the joint distribution of reservoirs is non-degenerate. The constraint $\succeq 0$

imposes positive semi-definiteness on the matrix L (see [CPW12] for an explanation of this constraint). Here, $\hat{L}$ provides an estimate for the low-rank component of the precision matrix (corresponding to the effect of latent variables on reservoir volumes), and $\hat{S}$ provides an estimate for the sparse component of the precision matrix (corresponding to the residual dependencies between reservoirs after accounting for the latent variables).
The regularization parameter γ provides a trade-off between the graphical model component and the latent component. In particular, for very large values of γ, the convex program (2.5) produces the same estimates as the graphical model estimator (2.2) (that is, $\hat{L} = 0$ so that no latent variables are used). As γ decreases, the number of latent variables increases and correspondingly the number of edges in the residual graphical structure decreases; this is because latent variables account for a global signal common to all reservoirs. The regularization parameter λ provides overall control of the trade-off between the fidelity of the model to the data and the complexity of the model.

As before, the function $\|\cdot\|_1$ denotes the L_1 norm that promotes sparsity in the matrix S. The role of the trace penalty on L is to promote low-rank structure [Faz02]. As before, for $\lambda, \gamma \geq 0$, (2.5) is a convex program with a unique optimum that can be solved efficiently. Theoretical support for this estimator is presented in [CPW12].

Similar to the graphical model setting, we use the *holdout validation* technique to determine the number of global latent variables and edges in the graphical structure between reservoirs. Concretely, for a particular choice of λ, γ, we supply $\mathcal{D}_{\text{train}}$ as input to the program (2.5) to learn a latent variable graphical model and compute the average log-likelihood of this model on the validation set $\mathcal{D}_{\text{test}}$. We sweep over all possible choices of γ, λ and choose a set of parameters that yield the best validation performance.

Let the selected model (after holdout validation) be specified by the parameters $(\hat{S}, \hat{L})$. The matrix $\hat{L}$ denotes the effect of $k = \text{rank}(\hat{L})$ latent variables on the reservoir network. The matrix $\hat{S}$ encodes the residual graphical structure between reservoirs after incorporating k latent variables. We can quantify the strength of the edges of this graphical structure using the relation (2.3) with $\hat{\Theta}$ replaced with $\hat{S}$. Finally, we quantify the portion of the variability of the network explained by the

latent variables as follows: the model estimates the covariance matrix of reservoirs as $(\hat{S} - \hat{L})^{-1}$ so that $y \sim \mathcal{N}(0\,,\ (\hat{S} - \hat{L})^{-1})$. Given that the variance of a reservoir r is $\left[(\hat{S} - \hat{L})^{-1}\right]_{r,r}$, we denote the overall variance of the network as $\sum_{r=1}^{55}\left[(\hat{S} - \hat{L})^{-1}\right]_{r,r}$. The variance of reservoir r, conditioned on k latent variables, is given by $(\hat{S}^{-1})_r$. We thus denote the variance of the network conditioned on k latent variables by $\sum_{r=1}^{55}\left[\hat{S}^{-1}\right]_{r,r}$. Furthermore, we define the ratio

$$\delta(k) = \frac{\sum_{r=1}^{55}\left[(\hat{S} - \hat{L})^{-1} - \hat{S}^{-1}\right]_{r,r}}{\sum_{r=1}^{55}\left[(\hat{S} - \hat{L})^{-1}\right]_{r,r}} \tag{2.6}$$

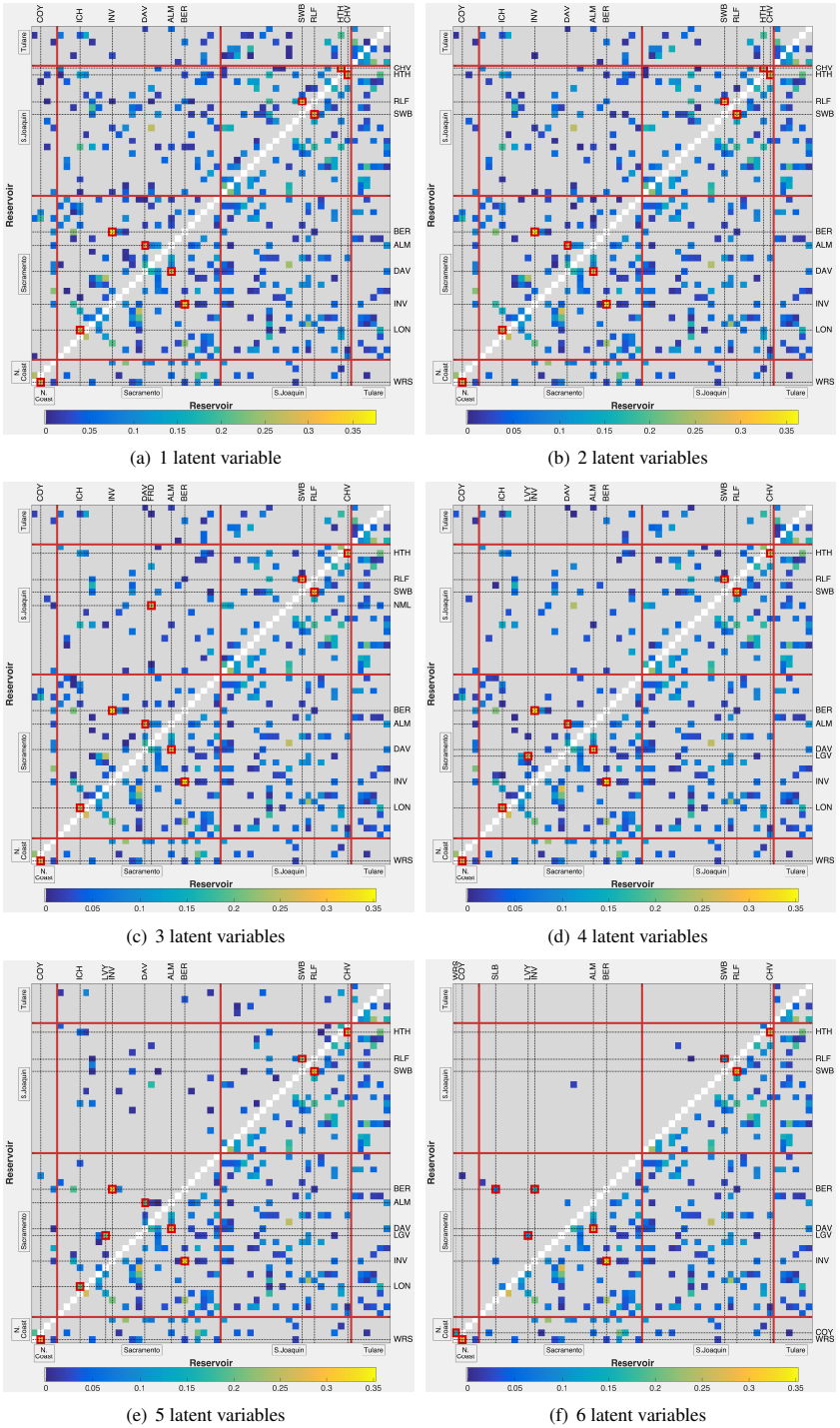
as the portion of the variability of the network explained by k latent variables.

Results: Accounting for Drivers of the Reservoir Network

We first explore the effect of global drivers on the connectivity of the reservoir network. Using observations $\mathcal{D}_{\text{train}}$ as input to the convex program (2.5), we vary the regularization parameters (λ, γ) to learn a collection of latent variables graphical models. Figure 2.8 shows the residual conditional graphical structure corresponding to each model. We observe that an increase in the number of latent variables leads to sparser structures and stronger inner-zone connections. Indeed, the ratios of inner zone edge strengths to total edge strength are $\kappa = 0.91$, $\kappa = 0.91$, $\kappa = 0.93$, $\kappa = 0.94$, $\kappa = 0.97$, and $\kappa = 0.99$ for models with 1, 2, 3, 4, 5, and 6 latent variables respectively. These results support the idea that latent variables extract global features that are common to all reservoirs, and incorporating them results in more localized interactions. The residual dependencies that persist (even after including several latent variables) can be attributed to unmodeled local variables.

Further, appealing to relation (2.6), the portion of the variability of the network explained by 1, 2, 3, 4, 5, and 6 latent variables is given by $\delta(1) = 0.23$, $\delta(2) = 0.25$, $\delta(3) = 0.28$, $\delta(4) = 0.31$, $\delta(5) = 0.32$, and $\delta(6) = 0.40$ respectively. Thus, the effect of latent variables on the network increases as we incorporate more of them in the model. Nonetheless, even 6 latent variables explain less than 50% of the reservoir variability, with the other portion attributed to residual conditional dependencies between reservoirs. Furthermore, this experiment suggests that both the influence of global latent variables and residual dependencies among reservoirs are important factors of the reservoir network variability.

We now focus on one of these latent variables. In particular, we choose the parameters (γ, λ) via holdout validation with the validation set $\mathcal{D}_{\text{test}}$ to learn a latent



(a) 1 latent variable (b) 2 latent variables

(c) 3 latent variables (d) 4 latent variables

(e) 5 latent variables (f) 6 latent variables

Figure 2.8: Linkages between reservoir pairs in the latent-variable sparse graphical model (upper triangle) with varying number of latent variables compared with those of the ordinary sparse graphical model model (lower triangle). Connection strength $s(r, r')$ is shown in the image map, with unlinked reservoir pairs drawn in gray. The four hydrological zones are separated by red lines. Red boxes surround the five strongest connections in each model.

variable graphical model consisting of two latent variables together with a residual graphical model (conditioned on the latent variables) having 171 edges. This is the model corresponding to Figure 2.8(b). Thus, the reservoir network consists of two global drivers, and some residual dependencies persist after accounting for their influence. The training and validation performance of this model (in terms of log-likelihood) are given by -62.11 and -85.87, respectively.

The conditional dependency relationships between reservoir pairs in this residual graphical structure are shown in the upper triangle of Figure 2.8(b). Comparing this graphical structure with the graphical structure without any latent variables (lower triangle of Figure 2.8(b)), accounting for the global drivers weakens or removes many connections between reservoirs: 134 are removed and 252 are weakened. Of the 134 edges removed, 94 are between reservoirs in different hydrological zones. Further, the latent variable graphical model has comparable model complexity and training/testing performance to the graphical model without latent variables. We conclude that many of the connections in the graphical model (without latent variables) are due to unmodeled global drivers and accounting for these variables leads to fewer remaining conditional dependencies.

Finally, of the 55 reservoirs in our system, 35 are used for sourcing hydroelectric power. In the graphical structure without latent variables, there are 154 pairwise edges between reservoirs that are used for generating hydroelectric power. Once the latent variables are incorporated, all but 15 of these edges are weakened or removed. This suggests that hydroelectric power is strongly correlated to one of the global drivers. We verify this hypothesis in the next section.

Method: Interpreting Latent Variables via Correlation Analysis

Latent-variable graphical modeling identifies a mathematical representation of the global drivers of the reservoir network. Naturally, one is interested in linking these mathematical variables to real-world signals to aid understanding of factors that globally affect the reservoir network. We propose an approach to give physical interpretations to the estimated global drivers. The high level intuition of this approach is to identify a space of all possible latent variable data termed *the latent space*. Then we compute the correlation of external covariates (the covariates we consider are in Section 2.2) with this space. Candidate covariates with high correlation are variables that globally influence the reservoir network.

Suppose we identified a latent variable graphical model with estimates $(\hat{S}, \hat{L})$ and

$k = \text{rank}(\hat{L})$. Let $z \in \mathbb{R}^k$ denote the latent variables (i.e. k global variables influencing the reservoir network) and $y \in \mathbb{R}^{55}$ denote reservoir volumes; further, partition the joint precision matrix of (y, z) as $\tilde{\Theta} = \begin{pmatrix} \tilde{\Theta}_y & \tilde{\Theta}'_{zy} \\ \tilde{\Theta}_{zy} & \tilde{\Theta}_z \end{pmatrix}$. A natural approximation for the observations of z given observations $\mathcal{D}_{\text{train}}$ is the conditional mean:

$$\tilde{z}^{(i)} = \mathbb{E}[z^{(i)} \mid y^{(i)}] = \tilde{\Theta}_z^{-1}\tilde{\Theta}_{zy}y^{(i)}. \tag{2.7}$$

If $\tilde{\Theta}_z$ and $\tilde{\Theta}_{zy}$ were explicitly known, the length-n_{train} observations $\{\tilde{z}^{(i)}\}_{i=1}^{n_{\text{train}}} \subset \mathbb{R}^k$ would provide an estimate of the latent variables given observations $\mathcal{D}_{\text{train}}$. As discussed in [CPW12], the low-rank component in the decomposition of the marginal precision matrix of y is $\hat{L} = \tilde{\Theta}'_{zy}\tilde{\Theta}_z^{-1}\tilde{\Theta}_{zy}$. However, even thought we have $\hat{L}$, this does not uniquely identify $\tilde{\Theta}_z^{-1}\tilde{\Theta}_{zy}$. Indeed, for any non-singular $A \in \mathbb{R}^{k\times k}$, one can transform $\tilde{\Theta}_z \to A\tilde{\Theta}_z A'$ and $\tilde{\Theta}_{zy} \to A\tilde{\Theta}_{zy}$ without altering $\hat{L}$. In terms of z, these observations imply that for any non-singular A, $\{A^{-1}\tilde{z}^{(i)}\}_{i=1}^{n_{\text{train}}}$ is an equivalent realization of the latent variable data: z is recoverable only up to a nonsingular transformation.

Nevertheless, the structure of the low-rank matrix $\hat{L}$ places a constraint on the effect of the latent variables z on y. Let $\tilde{Z} \in \mathbb{R}^{n\times k}$ denote a (non-unique) realization of latent variable observations. As we have seen, $\tilde{Z}A'^{-1}$ is an equivalent realization. The key *invariant* is the column-space of $\tilde{Z}$, a k-dimensional linear subspace of $\mathbb{R}^{n_{\text{train}}}$. We thus define the *latent space* to be the column-space of $\tilde{Z}$. We recover the latent space as follows: Let $Y \in \mathbb{R}^{n_{\text{train}}\times 55}$ denote observations of reservoir volumes, (2.7) becomes $\tilde{Z} = Y\tilde{\Theta}'_{zy}\tilde{\Theta}_z^{-1}$. Since the column-space of $Y\tilde{\Theta}'_{zy}\tilde{\Theta}_z^{-1}$ is equal to the column-space of $Y\hat{L}$, the basis elements of the latent space are given by the k left singular vectors of the matrix YL, which can be readily computed. We interpret the underlying latent variables by correlating each covariate with this latent space. The mathematical formulation of this correlation analysis is as follows: let $\mathcal{T} \subset \mathbb{R}^n$ with $\dim(\mathcal{T}) = k$ denote the latent space. Let $X_1 \in \mathbb{R}^{n_{\text{train}}}$ be the n_{train} observations of the covariate x_1 (normalized to have unit variance). The correlation of this covariate with the latent space is given by $\text{corr}(x_1) = \left\|\mathcal{P}_{\mathcal{T}}(X_1)\right\|_{\ell_2}$, where $\mathcal{P}_{\mathcal{T}}$ denotes the projection matrix onto the subspace $\mathcal{T}$. By definition, the quantity $\text{corr}(x_1)$ is between 0 and 1 with large values indicating that the covariate x_1 has a strong influence over the entire reservoir network.

Suppose we have identified a covariate x_1 that is highly correlated with the latent space. We can modify our technique to identify other covariates that are correlated with the latent space after taking away the effect of the covariate x_1. Taking this

effect away from further analysis is important since the covariates may be dependent on one another (e.g. PDSI and temperature). Let $U_1 D_1 V_1'$ be the reduced SVD of X_1 where $U_1 \in \mathbb{R}^{n_{\text{train}}}$, $D_1 \in \mathbb{R}$ and $V_1 \in \mathbb{R}$. Let $X_2 \in \mathbb{R}^{n_{\text{train}}}$ be the n_{train} observations of the covariate x_2. The correlation of a covariate x_2 with the latent space after taking away the effect of x_1 is given by: $\text{corr}_{x_1}(x_2) = \left\| (I - U_1 U_1')\mathcal{P}_{\mathcal{T}}(I - U_1 U_1')(X_2) \right\|_{\ell_2}$. If the quantity $\text{corr}_{x_1}(x_2)$ is large, then the covariate x_2 is strongly correlated to the second global statewide variable. We can once again take away the effect of the covariates x_1 and x_2 from the latent space, and find its correlation with another covariate x_3. Let $U_2 D_2 V_2'$ be the reduced SVD of $[X_1, X_2] \in \mathbb{R}^{n_{\text{train}} \times 2}$ where $U_2 \in \mathbb{R}^{n_{\text{train}} \times 2}$, $D_2 \in \mathbb{R}^{2 \times 2}$ and $V_2 \in \mathbb{R}^{2 \times 2}$. Let $X_3 \in \mathbb{R}^{n_{\text{train}}}$ be the n_{train} observations of the covariate x_3. The correlation of a covariate x_3 with the latent space after taking away the effect of x_1 and x_2 is given by $\text{corr}_{x_1,x_2}(x_3) = \left\| (I - U_2 U_2')\mathcal{P}_{\mathcal{T}}(I - U_2 U_2')(X_3) \right\|_{\ell_2}$. Similarly, if the quantity $\text{corr}_{x_1,x_2}(x_3)$ is large, then the covariate x_3 is strongly correlated to the third global driver. We can repeat this procedure to identify all the k global drivers influencing the reservoir network.

We make two remarks. First, the observations $\{y^{(i)}\}$ used in (2.7) to characterize the latent space need not be the same as the data employed to identify a latent variable graphical model using the estimator (5). In particular, to quantify the correlation of a covariate with the global drivers, we use observations $\{y^{(i)}\}$ in (2.7) that are of the same time scale and period as the data that is available for the covariate. As an example, if data for a particular covariate is only available from January 2005 - January 2016 at a monthly scale, we use monthly observations of y during the same time period in (2.7) to characterize the latent space, and subsequently link the observations of the covariate to this space. Second, we note that a subset of the authors of the present chapter have proposed an alternate approach for giving physical interpretation to the global drivers. This procedure is different that the one proposed in this chapter and is based on solving a convex optimization program [TC18] (see Chapter 6).

Results: Semantics for Global Drivers of the Reservoir Network

The latent variable graphical model identified two global drivers influencing the reservoir network. As described in Section 2.4, this yields a two dimensional latent space corresponding to all possible observations of the global drivers. To obtain real-world representation of these two global drivers, we link the two dimensional *latent space* to the 7 covariates described in Section 2.2. Recall from Section 2.2 that the covariates PDSI, Colorado river discharge, temperature, and snowpack had

a time lag of two months. The time lag for each of these covariates was selected to maximize their correlation with the latent space.

We find that the covariates PDSI and hydroelectric power have the largest correlations with $\rho = 0.88$ and $\rho = 0.80$, respectively. Secondary covariate influences are due to consumer price index, Colorado river discharge, Sierra Nevada snowpack (their correlations values are all less than $\rho = 0.5$) with little influence from the number of agricultural workers and temperature. We deduce that PDSI, being computed from variables like precipitation and temperature that control mass balance, is a forcing function on system-wide reservoir levels, while correlation of water levels with aggregate hydropower generation is a system-wide response to high reservoir levels across the network. We then take the effect of PDSI away from the latent space to find the correlation of the modified latent space with the remaining 6 covariates. We notice that the correlation of CPI (consumer price index) and Colorado river discharge with the latent space do not change very much, since they are unlikely to be structurally connected to PDSI. On the other hand, the correlation of number of agricultural workers, Sierra Nevada snowpack, hydroelectric power, and temperature are significantly reduced as they are largely dependent on PDSI. Nevertheless, all the 6 covariates have less than 0.5 correlation with the modified latent space. Further tests with additional covariates could yield candidates with strong influence over the reservoir network. The complete list of each covariate and its correlation with the latent space before and after removing PDSI is shown in Table 2.4.

Covariate	Correlation	Correlation after removing PDSI
Palmer Drought Severity Index (PDSI)	0.88	N/A
Hydroelectric power	0.80	0.09
Sierra Nevada snow pack	0.50	0.32
Consumer Price Index (CPI)	0.33	0.25
Colorado river discharge	0.29	0.23
Number of agricultural workers	0.17	0.03
Temperature	0.10	0.04

Table 2.2: Covariates and correlations with the latent space before and after removing PDSI

In the subsequent section, we describe an approach for incorporating PDSI as a covariate in the next iteration of graphical modeling to learn a joint distribution over reservoir volumes and PDSI. Since we identified one of the two global drivers influencing the network, we account for the presence of residual latent variables in

the modeling framework.

2.5 Systemic Dependency of the Network to Global Drivers

The previous experiment confirmed that the statewide PDSI signal is a strong forcing function on the entire reservoir network. For purposes of full generality, suppose that using the approach described in Section 4.3, we discovered a collection of covariates that are the global drivers of the reservoir network. We can extend our modeling framework to incorporate these covariates and characterize the behavior of the network subject to extreme values of these covariates.

Method: Conditional Latent Variable Graphical Modeling

Let $x \in \mathbb{R}^q$ be a collection of covariates that are global drivers of the reservoir network (in our setting, $q = 1$ and x is the PDSI variable). Since x can account for the effect of some of the global drivers, the distribution of y given x may still depend on a few *residual latent variables*. Therefore, we fit a latent variable graphical model to the conditional distribution of $y|x$. We term this modeling framework as *conditional latent variable graphical modeling*.

Let Σ be the join covariance matrix of $(y, x) \in \mathbb{R}^{55+q}$ and $\Theta = \Sigma^{-1}$ be the corresponding joint precision matrix partitioned as $\Theta = \begin{pmatrix} \Theta_y & \Theta_{yx} \\ \Theta'_{yx} & \Theta_x \end{pmatrix}$. The conditional precision matrix of y given x is equal to the submatrix Θ_y. Following the description of Section 4.1, fitting a latent variable graphical model to the distribution of y given x corresponds to decomposing the submatrix Θ_y as the difference $S_y - L_y$. The matrix L_y is the effect of residual latent variables on the reservoirs after regressing on the covariates x, and its rank is equal to the number of residual latent variables. The matrix S_y specifies the residual dependencies among reservoirs after accounting for x and residual latent variables. The sparsity pattern of S_y encodes the residual graphical structure among reservoirs.

Let $\mathcal{D}^+_{\text{train}} = \{(y^{(i)}; x^{(i)})\}_{i=1}^{n_{\text{train}}} \subset \mathbb{R}^{55+q}$ be the training set of reservoir volumes augmented with covariate data and let $\mathcal{D}^+_{\text{test}} = \{(y^{(i)}; x^{(i)})\}_{i=1}^{n_{\text{test}}} \subset \mathbb{R}^{55+q}$ be the corresponding validation set. A natural approach for fitting a conditional latent variable graphical model is to choose the simplest model, e.g. the smallest number of residual latent variables and sparsest residual graphical model, that adequately explains the data. Following a similar line of reasoning as the case of latent variable graphical modeling, we arrive at the following estimator for fitting a conditional

latent variable graphical model to the observations $\mathcal{D}^+_{\text{train}}$ [TC15]:

$$(\hat{\Theta}, \hat{S}_y, \hat{L}_y) = \underset{\substack{\Theta \in \mathbb{S}^{55+q} \\ S_y, L_y \in \mathbb{S}^{55}}}{\arg\min} \quad -\ell(\Theta; \mathcal{D}^+_{\text{train}}) + \lambda(\|S_y\|_1 + \gamma \operatorname{tr}(L_y))$$
$$\text{s.t.} \qquad \Theta \succ 0,\ \Theta_y = S_y - L_y,\ L_y \succeq 0. \tag{2.8}$$

The term $\ell(\Theta; \mathcal{D}^+_{\text{train}})$ is the Gaussian log-likelihood function over the variables (y, x), which after removing constants terms and scaling is given by

$$\ell(\Theta; \mathcal{D}^+_{\text{train}}) = \log\det(\Theta) - \operatorname{tr}\left[\Theta \cdot \Sigma^+_n\right] \quad ,$$

where $\Sigma^+_n = \frac{1}{n_{\text{train}}} \sum_{i=1}^{n_{\text{train}}} \begin{pmatrix} y^{(i)} \\ x^{(i)} \end{pmatrix} \begin{pmatrix} y^{(i)} \\ x^{(i)} \end{pmatrix}'$ is the sample covariance matrix of reservoirs and covariates. The program (6.3) with $\lambda = 0$ is the unregularized multivariate maximum likelihood estimator of reservoirs and covariates. For $\lambda, \gamma \geq 0$, the regularized maximum likelihood estimator of (6.3) is a convex program with a unique optimum and can be solved efficiently, similar to estimators (2.2) and (2.5). Theoretical support for this estimator is presented in [TC15]. We note that a conditional graphical model could also be obtained using other techniques, such as the convex program proposed by [FJM17].

We select the regularization parameters λ, γ in (6.3) via holdout validation with the testing set $\mathcal{D}^+_{\text{test}}$. Concretely, for a particular choice of λ, γ, we supply $\mathcal{D}^+_{\text{train}}$ as input to the program (6.3) to obtain a conditional latent variable graphical and validate the performance on the validation set $\mathcal{D}^+_{\text{test}}$. We perform this procedure as we vary λ, γ and choose the model with the best validation performance.

Suppose we obtain a conditional latent variable graphical model over $(y, x) \in \mathbb{R}^{55 \times q}$ with estimates $(\hat{\Theta}, \hat{S}_y, \hat{L}_y)$. We use this model to characterize the behavior of the network in response to the covariates x in the month of November (the analysis can be done for any month). Our metric for the behavior of the network is the *probability of simultaneous exhaustion*: the probability that the volumes of a collection of reservoirs drop below zero. Letting $\hat{\Sigma} = \hat{\Theta}^{-1}$, the composite variable $(y, x) \in \mathbb{R}^{55+q}$ is distributed as $(y, x) \sim \mathcal{N}(0, \hat{\Sigma})$. (Preprocessing to remove climatology causes the mean to be zero.) To determine the behavior of a collection of K reservoirs $\mathbf{r} = \{r_1, r_2, \ldots, r_K\}$ as the covariates x vary, we extract the $(K + q) \times (K + q)$ block of $\hat{\Sigma}$ corresponding to $y_{\mathbf{r}} \in \mathbb{R}^K$ and x, and recall that

$$y_{\mathbf{r}} \mid x \sim \mathcal{N}(\hat{\Sigma}_{y_{\mathbf{r}},x} \hat{\Sigma}_x^{-1} x,\ \hat{\Sigma}_{y_{\mathbf{r}}} - \hat{\Sigma}_{y_{\mathbf{r}},x} \hat{\Sigma}_x^{-1} \hat{\Sigma}_{x,y_{\mathbf{r}}}) \quad , \tag{2.9}$$

an instance of the standard expressions for the conditional mean and variance of these jointly Gaussian variables. Let the November climatology, subtracted during

preprocessing, for reservoir volume y_r ($r \in \mathbf{r}$) be μ_{y_r}, and the November climatology of x be $\mu_x \in \mathbb{R}^q$. Let the scaling used to make the time series of y_r have unit variance be a_{y_r} and the scaling matrix used to make the time series of each covariate to have unit variance be $a_x \in \mathbb{R}^{q \times q}$. Then, for $x = u$, the probability that at least k of K reservoirs have their volume drop below zero in November is:

$$P(A_K(k) \mid x = a_x(u - \mu_x)), \tag{2.10}$$

where $A_K(k)$ is the event that $y_r \leq -\mu_{y_r} a_{y_r}$ for at least k of the K reservoirs. The probability in (2.10), or that of any system-wide event, can be computed using Monte Carlo draws from the joint conditional distribution.

We can further use the model to identify "weak nodes" of the network: reservoirs that are at high risk of exhaustion. In particular, we compute the probability of each reservoir conditioned on PDSI, namely,

$$P(y_r < -\mu_{y_r} a_{y_r} \mid x = a_x(u - \mu_x)), \tag{2.11}$$

by applying (2.10) with $K = 1$.

Results: Network Behavior Under Drought

To obtain a system-wide response to drought, we follow the approach described in Section 2.5 to compute the probability of exhaustion of a collection of reservoirs conditioned on particular PDSI. We obtain this probability by learning a conditional latent variable graphical model over reservoir volumes and PDSI. This probability is computed for the month of November, when reservoirs are typically at their lowest, but the same calculation applies to any month. Since we applied a time lag of two months to the PDSI time series, these probabilities are computed based on September PDSI.

To learn a joint distribution, let $x \in \mathbb{R}$ denote PDSI and consider a conditional latent-variable graphical model over $(y, x) \in \mathbb{R}^{55+1}$. Using observations $\mathcal{D}^+_{\text{train}}$ (consisting of 55 reservoir volumes and PDSI values) and appropriate choice of regularization parameters λ, γ (using holdout validation), we fit a latent-variable graphical model to the conditional distribution $y \mid x$ via the estimator (6.3). The estimated model consists of 1 residual latent variable (e.g. $\text{rank}(\hat{L}_y) = 1$). Recall that the reservoir network consists of two global drivers. Evidently, by regressing away the effect of PDSI, we are left with one residual latent variable, which supports the observation that PDSI is a global driver of the reservoir network. It is plausible that a portion of the residual latent variable is due to management behavior.

The conditional latent-variable graphical modeling procedure also provides an estimate of a graphical model of the conditional distribution of y conditioned on PDSI (e.g. the matrix $\hat{S}_y$) — this graphical model consists of 206 edges. The training and validation performance of this model is -61.79 and -88.52, respectively. We now compute the systemic response to drought based on the conditional latent variable graphical model over reservoirs and PDSI. Appealing to relation (2.10), we can compute the probability that at least k of K reservoirs have their volume drop below zero in November. Here, we consider those res ervoirs having capacity of at least 10^8m^3 ($K = 31$). Of the 55 reservoirs in our dataset, 22 have capacity below 10^8 m^3. Two of the 33 remaining (Terminus and Success) are flood-control reservoirs: they are unique in that their volume routinely falls below 10% of capacity, independent of PDSI. Thus, we focus on the remaining 31 large reservoirs in what follows. We vary PDSI and compute (2.10) for selected values of k. Figure 2.9 indicates that with sustained precipitation deficits and a PDSI approaching -5, the probability that three or more of California's major reservoirs run dry is greater than 50%. This probability increase above 80% as PDSI drops to -6.

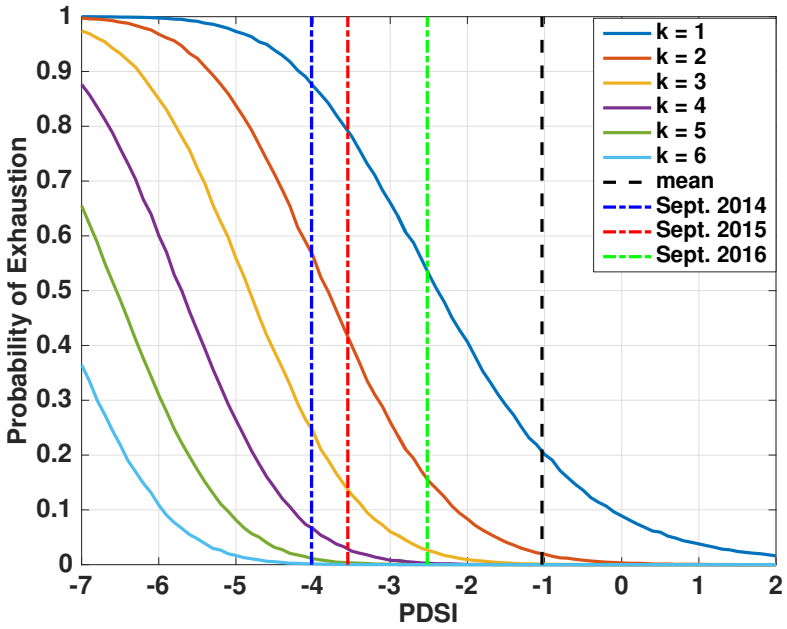


Figure 2.9: System-wide response to drought in a conditional latent variable graphical model: probability that at least k reservoirs out of 31 large reservoirs (with capacity $\geq 10^8\text{m}^3$) will have volume fall to zero, for a range of PDSI; Dashed black line: average September PDSI (September 2004—September 2015). Dashed blue line: September 2014 PDSI. Dashed red line: September 2015 PDSI. Dashed green line: September 2016 PDSI.

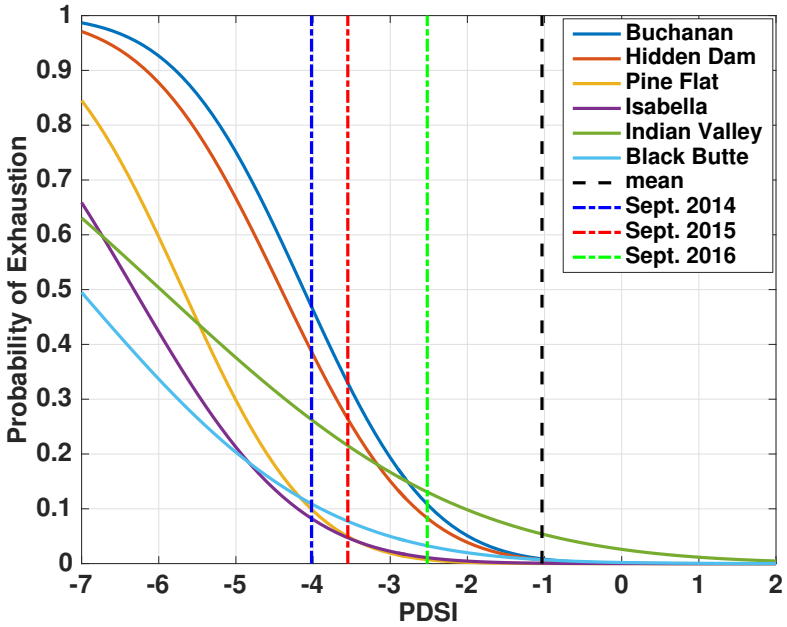


Figure 2.10: Individual reservoir responses to drought in a conditional latent variable graphical model: probability that six most-at-risk reservoirs out of 31 large reservoirs (with capacity $\geq 10^8\ m^3$) will have volume drop below zero; Dashed black line: average September PDSI (September 2004-September 2015). Dashed blue line: September 2014 PDSI. Dashed red line: September 2015 PDSI. Dashed green line: September 2016 PDSI.

Implications

The results of Figure 2.9 indicate that under severe drought conditions (e.g. small values of PDSI), there is a high risk of simultaneous exhaustion of multiple large reservoirs. To further investigate the implications of drought on reservoir conditions, we use (2.11) to compute the probability of exhaustion of each reservoir as a function of PDSI. Figure 2.5 shows those reservoirs (among 31 large reservoirs with capacity greater than $10^8\ m^3$) that were highly sensitive to PDSI. Evidently, these reservoirs are at high risk of exhaustion, and additionally, some have a greater sensitivity to small PDSI changes than others.

We focus on two reservoirs with highest risk of exhaustion: Buchanan and Hidden Dam reservoir. Stringent management practices, however, have prevented these reservoirs from running dry. Specifically, the Madera Irrigation District, which owns the water rights of the Hidden Dam reservoir, allowed for the release of very small amount of water during the drought period of 2014 – 2015. This is because the reservoir volume had reached the minimum pool of 5,000 acre feet (6.1 x $10^6\ m^3$, $\approx$ 5% of the total capacity) required for recreational purposes. The

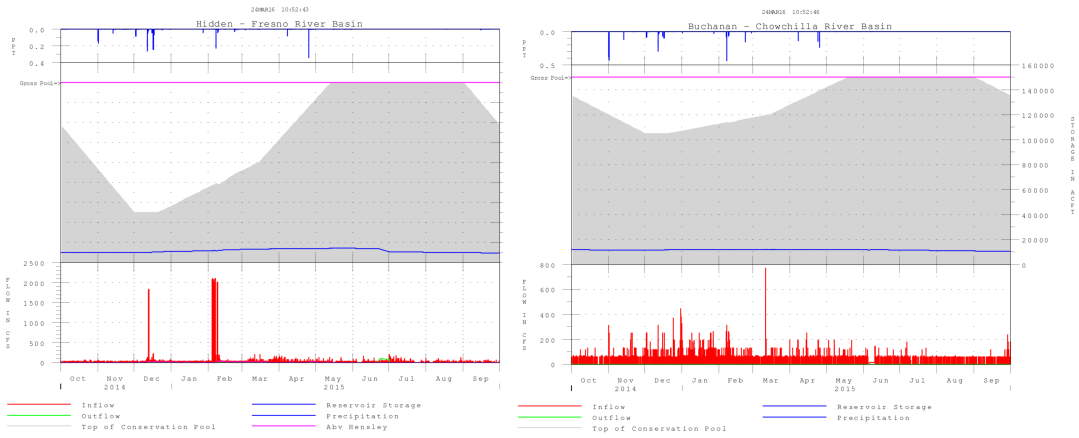


Figure 2.11: Inflows, outflows, precipitation, and water levels for the Buchanan and Hidden Dam reservoirs during the extreme drought period of 2014-2015. Notice that there was little precipitation, leading to marginal inflow of water into each reservoirs. Due to heavy management, there was little to no outflow of water from these reservoirs, preventing them from running dry. These figures are obtained from the Sacramento District Water Control Data System at `http://www.spk-wc.usace.army.mil/plots/california.html`.

Buchanan reservoir received a same degree of stringent management. During the $2014-2015$ period, the reservoir volume reached the minimum pool of $10,000$ acre feet (12.2 x 10^6 m^3, $\approx 6\%$ of the total capacity) required for recreational purposes. As a result, the Chowchilla Water District, which owns the water rights of the Buchanan reservoir, determined that no water will be released during the $2014-2015$ period. Figure 2.5 demonstrates the amount of water from precipitation into the Hidden Dam and Buchanan reservoirs, the total inflow, and the outflow as a consequence of the stringent management efforts.

Thus, at low reservoir volumes, the stringent management that these reservoirs receive results in their behavior deviating from the predictions of our model. To further highlight this distinction, we examine the historical reservoir volumes of Buchanan and Hidden Dam as a function of PDSI in Figure 2.5. As expected, there is a positive correlation between PDSI and reservoir volumes: smaller values of PDSI generally result in a lower volume. Suppose we restrict our attention to PDSI greater than -3. In this regime, the correlation of the Buchanan and Hidden Dam reservoirs with PDSI as obtained from our model is similar to the empirical historical average. On the other hand, for PDSI values less than -3 (corresponding to drought period 2014-2015), the empirical correlations are significantly reduced. Concretely, the empirical correlation of the Buchanan reservoir is a factor of $\approx 6/100$ of the value

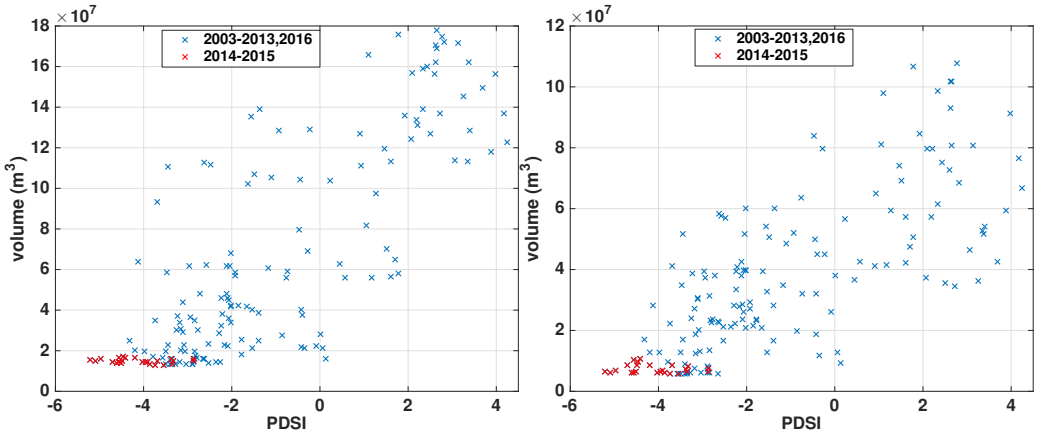


Figure 2.12: PDSI vs reservoir levels for the Buchanan and Hidden Dam reservoirs during the period of study (i.e. January 2003 to November 2016). Notice a positive correlation between PDSI and the reservoir volumes: smaller values of PDSI generally lead to lower reservoir volumes. During the 2014-2015 drought period (shown in red), the correlation is substantially reduced as a result of stringent management efforts.

estimated by our model. The empirical correlation of the Hidden Dam is a factor $\approx 2/5$ of the correlation estimated by our model. The significant reductions in these correlations for low PDSI values highlight the impact of the severe management practices. Our model is representative of the reservoir behavior in a “Business as Usual" (BAU) regime where heavy management practices have not been employed and therefore correlations of PDSI and reservoirs volumes are independent of PDSI value. Consequently, an alternative interpretation of our results is that Figure 2.9 provides an advanced guideline as to when strict reservoir management *needs* to be employed to leave the BAU regime — in effect breaking the correlation of PDSI and reservoir volumes — to prevent reservoir exhaustion. More specifically, we propose the following rule of thumb in situations where one may have advanced prediction of the PDSI value: if the exhaustion probabilities are low at the predicted value of PDSI, no heavy management effort is likely to be needed and the reservoir could be operated in a BAU setting. If these probabilities start to rise above 50%, this indicates trouble and that water managers *need* to prepare to leave the BAU regime.

To summarize, the proposed model characterizes the risk of exhaustion of large California reservoirs during extreme drought. The proposed methodology can be used to inform water managers of potential risks under typical management behavior. Additionally, the method used here can forecast other key events that precede reservoir exhaustion, such as when power generation is made impossible

as water levels drop below turbine inlets, or when water levels reach the minimum pool for recreational purposes.

2.6 Discussion and Future Directions

The California reservoir system is summarized by a complex, dynamic network of correlated time series that respond to a diverse set of global and local drivers, including both natural climate processes and human decision-making. Our objective was to develop the first statewide model of this complex network to address these scientific questions:

1. What are the interactions or dependencies among reservoir levels?

2. Are there common external factors influencing the network globally? Could these external drivers cause a system-wide catastrophe?

We appealed to a powerful modeling framework, known as graphical modeling, to address these questions. These models characterize the complex relationships among reservoirs, and can be learned efficiently based on solving a regularized maximum likelihood estimator. We identified a graphical model consisting of 285 edges over the reservoir network and demonstrated that $\approx 85\%$ of the dependencies are between reservoirs in the same hydrological zone. We observed that reservoirs with similar hydrological attributes (e.g. elevation and drainage area) tend to exhibit stronger dependencies. We further characterized Folsom Lake to be the most connected reservoir in the network, and demonstrated its strong dependencies with reservoirs connected to the Sacramento river. To address question 2, we quantified the influence of external phenomena on the network using an extension of the graphical modeling framework, known as *latent variable graphical modeling*. These models can be learned efficiently based on solving a generalization of the maximum likelihood estimator in the graphical modeling setting. Using historical reservoir data, we determined two global drivers influence the reservoir network at a monthly resolution, and proposed a novel methodology to obtain physical interpretation of these global drivers. We found that PDSI was highly correlated ($\rho \approx 0.88$) with one of the global drivers. We then used PDSI as a covariate in the *next* iteration of the graphical modeling procedure to characterize risks of system-wide catastrophe in response to hypothetical drought conditions. We also identified that Buchanan and Hidden Valley reservoirs are high susceptible to exhaustion.

The approach applied here to study reservoirs has the potential to be applicable across many complex data problems in the geosciences. The graphical modeling technique can be first used to model the complex network of variables. The model can be enhanced to account for global drivers (latent variables) that influence the entire network. Then a latent space summarizing all possible configurations of latent variable data can be estimated by model optimization. Candidate external forcing data can be linked to this latent space to find matches. Once a best match is found, the effect of this covariate can be taken away and other covariates could be tested to identify all the drivers of the global system variability. Then the latent variables could be included as covariates in a new iteration of the graphical modeling procedure to learn a joint model over the network variables and covariates. Using this model, the behavior of the network under extreme values of the global drivers can be characterized. This procedure has the additional value of directing and prioritizing observational efforts.

There are several interesting directions for future research. The analysis of this chapter was over a network of 55 major reservoirs in California. It would be interesting to obtain volumetric measurements of many more reservoirs (currently the amount of data available is insufficient for analysis on a larger set of reservoirs) and apply our procedure to obtain a model over this larger network; indeed, there is no other obstruction to carrying out a more extensive analysis with the methodology presented in this chapter. Further, the statistical framework developed in this chapter is focused on a global model of the reservoir network and the influence of statewide variables. An exciting direction for future investigation is to complement our modeling framework to account for local variables (e.g. local temperature, local precipitation, etc.). Specifically, associated with each reservoir, we can include a collection of local variables and apply our framework to the reservoir volumes after regressing on the local variables. As described, this procedure would model the reservoir network at both local and global scales.

Chapter 3

FALSE DISCOVERY AND ITS CONTROL IN LATENT-VARIABLE MODELS

The latent-variable graphical modeling technique that was employed to model California reservoirs in Chapter 2 involved identifying a low-rank matrix. This low-rank matrix encodes the effect of unmodeled phenomena on the reservoirs, and may be attributed to known physical quantities such as drought severity and precipitation (see Chapter 2 for more details). In other words, this low-rank matrix represents a *discovery*, and one seeks to ensure that this discovery closely matches the underlying physical process. More broadly, low-rank estimation is a common approach in data analysis, and in many scenarios, the estimated low-rank captures a physical phenomena. Naturally, the following question arises: how do we assess and control false discoveries in low-rank estimation? This chapter addresses this question.

The results in this chapter are published in [TSC19] and were completed jointly with Parikshit Shah and Venkat Chandrasekaran. The author contributed by developing a geometric notion of false discovery, introducing an algorithm to control for false discoveries, analyzing the theoretical performance of this algorithm, and providing numerical experiments supporting this approach. The description of the work contained in this chapter was written by the author and Venkat Chandrasekaran.

3.1 Introduction

Models described by low-rank matrices are ubiquitous in many contemporary problem domains. The reason for their widespread use is that low-rank matrices offer a flexible approach to specify various types of low-dimensional structure in high-dimensional data. For example, low-rank matrices are used to describe user preferences in collaborative filtering [Gol+92], small collections of end-member signatures in hyperspectral imaging [Man03], directions of moving targets in radar measurements [FL11], low-order systems in control theory [LV09], coherent imaging systems in optics [PK94], and latent-variable models in factor analysis [Sha82c]. In many of these settings, the row/column space structure of a low-rank matrix carries information about some underlying phenomenon of interest; for instance, in hyperspectral imaging for mineralogy problems, the column space represents the combined signatures of relevant minerals in a mixture. Similarly, the row/column

spaces of matrices obtained from radar measurements signify the directions of moving targets. Therefore, in inferential contexts in which low-rank matrices are estimated from data, it is of interest to evaluate the extent to which the row/column spaces of the estimated matrices signify true/false *discoveries* about the relevant phenomenon.

In seeking an appropriate framework to assess discoveries in low-rank estimation, it is instructive to consider the case of variable selection, which may be viewed conceptually as low-rank estimation with diagonal matrices. Stated in terms of subspaces, the set of discoveries in variable selection is naturally represented by a subspace that is spanned by the standard basis vectors corresponding to the subset of variables that are declared as significant. The number of true discoveries then corresponds to the dimension of the intersection between this 'discovery subspace' and the 'population subspace' (i.e., the subspace spanned by standard basis vectors corresponding to significant variables in the population), and the number of false discoveries is the dimension of the 'discovery subspace' minus the number of true discoveries. Generalizing this perspective to low-rank estimation, it is perhaps appealing to declare that the number of true discoveries is the dimension of the intersection of the estimated row/column spaces and the population row/column spaces, and the number of false discoveries is the dimension of the remaining components of the estimated row/column spaces. The difficulty with this approach is that we cannot expect any inference procedure to perfectly estimate with positive probability even a one-dimensional subspace of the population row/column spaces as the collection of these spaces is not discrete; in particular, the set of all subspaces of a given dimension is the Grassmannian manifold, whose underlying smooth structure is unlike that of the finite collection of coordinate subspaces that correspond to discoveries in variable selection. Therefore, the number of true discoveries would generically be zero. One method to improve upon this idea is to define the number of true discoveries as the dimension of the largest subspaces of the estimated row/column spaces that are within a specified angle of the population row/column spaces, and to treat the dimension of the remaining components of the estimated row/column spaces as the number of false discoveries. An unappealing feature of this second approach is that it depends on an extrinsic parameter, and minor perturbations of this parameter could result in potentially large changes in the number of true/false discoveries. In some sense, these preceding attempts fail as they are based on a sharp binary choice that declares components of the estimated row/column spaces exclusively as true or false discoveries, which is ill-suited to the

smooth structure underlying low-rank matrices.

As our first contribution, we develop in Section 3.2 a *geometric* framework for evaluating false discoveries in low-rank estimation. We begin by expressing the number of true/false discoveries in variable selection in terms of functionals of the projection matrices associated to the discovery/population subspaces described above; this expression varies smoothly with respect to the underlying subspaces, unlike dimensions of intersections of subspaces. Next, we interpret the discovery/population subspaces in variable selection as tangent spaces to algebraic varieties of sparse vectors. Finally, we note that tangent spaces with respect to varieties of low-rank matrices encode the row/column space structure of a matrix, and therefore offer an appropriate generalization of the subspaces discussed in the context of variable selection. Putting these observations together, we substitute tangent spaces with respect to varieties of low-rank matrices into our reformulation of discoveries in variable selection in terms of projection matrices, which leads to a natural formalism of the number of true/false discoveries that is suitable for low-rank estimation. We emphasize that although our definition respects the smooth geometric structure underlying low-rank matrices, one of its appealing properties is that it specializes transparently to the usual discrete notion of true/false discoveries in the setting of variable selection if the underlying low-rank matrices are diagonal.

Our next contribution concerns the development of a procedure for low-rank estimation that provides false discovery control. In Section 3.3, we generalize the 'stability selection' procedure of [MB10] for controlling false discoveries in variable selection. Their method operates by employing variable selection methods in conjunction with subsampling; in particular, one applies a variable selection algorithm to subsamples of a dataset, and then declares as discoveries those variables that are selected most frequently. In analogy to their approach, our algorithm — which we call 'subspace stability selection' — operates by combining existing low-rank estimation methods in conjunction with subsampling. Our framework employs row/column space selection procedures (based on standard low-rank estimation algorithms) on subsamples of a dataset, and then outputs as discoveries a set of row/column spaces that are 'close to' most of the estimated row/column spaces; the specific notion of distance here is based on our tangent space formalism. Building on the results in [MB10; SS13], we provide a theoretical analysis of the performance of our algorithm.

Finally, in Section 3.4 we contrast subspace stability selection with previous

methods in a range of low-rank estimation problems involving simulated as well as real data. The tasks involving real data are on estimating user-preference matrices for recommender systems and identifying signatures of relevant minerals in hyperspectral images. The estimates provided by subspace stability selection offer improvements in multiple respects. First, the row/column spaces of the subspace stability selection estimates are far closer to their population counterparts in comparison to other standard approaches; in other words, our experiments demonstrate that subspace stability selection provides estimates with far fewer false discoveries, without a significant loss in power (both false discovery and power are based on the definitions introduced in this paper). Second, in settings in which regularized formulations are employed, subspace stability selection estimates are much less sensitive to the specific choice of the regularization parameter. Finally, a common challenge with approaches based on cross-validation for low-rank estimation is that they overestimate the complexity of a model, i.e., they produce higher rank estimates (indeed, a similar issue arises in variable selection, which was one of the motivations for the development of stability selection in [MB10]). We observe that the estimates produced by subspace stability selection have substantially lower rank than those produced by cross-validation, with a similar or improved prediction performance.

The outline of this paper is as follows. In Section 3.2, we briefly review the relevant concepts from algebraic geometry and then formulate a false discovery framework for low-rank estimation. Our subspace stability selection algorithm is described in Section 3.3, with theoretical support presented in Section 3.3. In Section 3.4, we demonstrate the utility of our approach in experiments with synthetic and real data. We conclude with a discussion of further research directions in Section 3.5.

Related work We are aware of prior work for low-rank estimation based on testing the significance level of the singular values of an observed matrix (see, for example, [CTT17], [LL18], [SS18]). However, in contrast to our framework, these methods do not directly control deviations of row/column spaces, which carry significant information about various phenomena of interest in applications. Further, these previous approaches have limited applicability as they rely on having observations of all the entries of a matrix; this is not the case, for example, in low-rank matrix completion problems which arise commonly in many domains. In comparison, our methodology is general-purpose and is applicable to a broad range of low-rank estimation problems. On the computational front, our algorithm and its analysis

are a generalization of some of the ideas in [MB10; SS13]. However, the geometry underlying the collection of tangent spaces to low-rank matrices leads to a number of new challenges in our context.

Notation For a subspace $\mathbb{V}$, we denote projection onto $\mathbb{V}$ by $\mathcal{P}_{\mathbb{V}}$. Given a self-adjoint linear map $M : \bar{\mathbb{V}} \to \bar{\mathbb{V}}$ on a vector space $\bar{\mathbb{V}}$ and a subspace $\mathbb{V} \subset \bar{\mathbb{V}}$, the minimum singular value of M restricted to $\mathbb{V}$ is given by $\sigma_{\min}(\mathcal{P}_{\mathbb{V}} M \mathcal{P}_{\mathbb{V}}) = \inf_{x \in \mathbb{V} \setminus \{0\}} \frac{\|Mx\|_{\ell_2}}{\|x\|_{\ell_2}}$. We denote Kronecker product between two matrices A and B by $A \otimes B$. Finally, the nuclear norm is denoted by $\|\cdot\|_\star$.

3.2 A Geometric False Discovery Framework

We describe a geometric framework for assessing discoveries in low-rank estimation. Our discussion proceeds by first reformulating true/false discoveries in variable selection in geometric terms, which then enables a transparent generalization to the low-rank case. We appeal to elementary ideas from algebraic geometry on varieties and tangent spaces [Har95]. We also describe a procedure to obtain an estimate of a low-rank matrix given an estimate of a tangent space.

False Discovery in Low-Rank Estimation

The performance of a variable selection procedure $\hat{\mathcal{S}} \subset \{1, \ldots, p\}$, which estimates a subset of a collection of p variables as being significant, is evaluated by comparing the number of elements of $\hat{\mathcal{S}}$ that are also present in the 'true' subset of significant variables $\mathcal{S}^\star \subset \{1, \ldots, p\}$ — the number of true discoveries is $|\hat{\mathcal{S}} \cap \mathcal{S}^\star|$, while the number of false discoveries is $|\hat{\mathcal{S}} \cap \mathcal{S}^{\star c}|$. We give next a geometric perspective on this combinatorial notion. As described in the introduction, one can associate to each subset $\mathcal{S} \subset \{1, \ldots, p\}$ the coordinate aligned *subspace* $T(\mathcal{S}) = \{x \in \mathbb{R}^p \mid \text{support}(x) \subseteq \mathcal{S}\}$, where $\text{support}(x)$ denotes the locations of the nonzero entries of x. With this notation, the number of false discoveries in an estimate $\hat{\mathcal{S}}$ is given by:

$$\texttt{\#false-discoveries} = |\hat{\mathcal{S}} \cap \mathcal{S}^{\star c}| = \dim(T(\hat{\mathcal{S}}) \cap T(\mathcal{S}^\star)^\perp) = \text{trace}\left(\mathcal{P}_{T(\hat{\mathcal{S}})} \mathcal{P}_{T(\mathcal{S}^\star)^\perp}\right).$$

Similarly, the number of true discoveries is given by $\text{trace}\left(\mathcal{P}_{T(\hat{\mathcal{S}})} \mathcal{P}_{T(\mathcal{S}^\star)}\right)$. These latter reformulations in terms of projection operators have no obvious 'discrete' attribute to them. In particular, for any subspaces $\mathcal{W}, \tilde{\mathcal{W}}$, the expression $\text{trace}(\mathcal{P}_{\mathcal{W}} \mathcal{P}_{\tilde{\mathcal{W}}})$ is equal to the sum of the squares of the cosines of the principal angles between $\mathcal{W}$ and $\tilde{\mathcal{W}}$ [BG73]; as a result, the quantity $\text{trace}(\mathcal{P}_{\mathcal{W}} \mathcal{P}_{\tilde{\mathcal{W}}})$ varies smoothly with respect

to perturbations of $\mathcal{W}, \tilde{\mathcal{W}}$. The discrete nature of a discovery is embedded inside the encoding of the subsets $\hat{\mathcal{S}}, \mathcal{S}^\star$ using the subspaces $T(\hat{\mathcal{S}}), T(\mathcal{S}^\star)$. Consequently, to make progress towards a suitable definition of true/false discoveries in the low-rank case, we require an appropriate encoding of row/column space structure via subspaces in the spirit of the mapping $\mathcal{S} \mapsto T(\mathcal{S})$. Towards this goal, we interpret next the subspace $T(\mathcal{S})$ associated to a subset $\mathcal{S} \subset \{1, \ldots, p\}$ as a tangent space to an algebraic variety.

Formally, for any integer $k \in \{1, \ldots, p\}$ let $\mathcal{V}_{\text{sparse}}(k) \subset \mathbb{R}^p$ denote the algebraic variety of elements of $\mathbb{R}^p$ with at most k nonzero entries. Then for any point in $\mathcal{V}_{\text{sparse}}(k)$ consisting of exactly k nonzero entries at locations given by the subset $\mathcal{S} \subset \{1, \ldots, p\}$ (here $|\mathcal{S}| = k$), the tangent space at that point with respect to $\mathcal{V}_{\text{sparse}}(k)$ is given by $T(\mathcal{S})$. In other words, the tangent space at a smooth point of $\mathcal{V}_{\text{sparse}}(k)$ is completely determined by the locations of the nonzero entries of that point. This geometric perspective extends naturally to the low-rank case.

Consider the *determinantal variety* $\mathcal{V}_{\text{low-rank}}(r) \subset \mathbb{R}^{p_1 \times p_2}$ of matrices of size $p_1 \times p_2$ with rank at most r (here $r \in \{1, \ldots, \min(p_1, p_2)\}$). Then for any matrix in $\mathcal{V}_{\text{low-rank}}(r)$ with rank equal to r and with row and column spaces given by $\mathcal{R} \subset \mathbb{R}^{p_2}$ and $C \subset \mathbb{R}^{p_1}$, respectively, the tangent space at that matrix with respect to $\mathcal{V}_{\text{low-rank}}(r)$ is given by:

$$T(C, \mathcal{R}) \triangleq \{M_R + M_C \mid M_R, M_C \in \mathbb{R}^{p_1 \times p_2}, \text{row-space}(M_R) \subseteq \mathcal{R}, \text{column-space}(M_C) \subseteq C\}. \tag{3.1}$$

The dimension of $T(C, \mathcal{R})$ equals $r(p_1 + p_2) - r^2$ and the dimension of its orthogonal complement $T(C, \mathcal{R})^\perp$ equals $(p_1 - r)(p_2 - r)$. Further, the projection operators onto $T(C, \mathcal{R})$ and onto $T(C, \mathcal{R})^\perp$ can be expressed in terms of the projection maps onto C and $\mathcal{R}$ as follows:

$$\begin{aligned} \mathcal{P}_{T(C,\mathcal{R})} &= \mathcal{P}_C \otimes I + I \otimes \mathcal{P}_{\mathcal{R}} - \mathcal{P}_C \otimes \mathcal{P}_{\mathcal{R}} \\ \mathcal{P}_{T(C,\mathcal{R})^\perp} &= (I - \mathcal{P}_C) \otimes (I - \mathcal{P}_{\mathcal{R}}) = \mathcal{P}_{C^\perp} \otimes \mathcal{P}_{\mathcal{R}^\perp}, \end{aligned} \tag{3.2}$$

where $\otimes$ denotes a kronecker product. Consequently, the action of projection operators $\mathcal{P}_{T(C,\mathcal{R})}$ and $\mathcal{P}_{T(C,\mathcal{R})^\perp}$ on a matrix $M \in \mathbb{R}^{p_1 \times p_2}$ yields:

$$\mathcal{P}_{T(C,\mathcal{R})}(M) = \mathcal{P}_C M + M \mathcal{P}_{\mathcal{R}} - \mathcal{P}_C M \mathcal{P}_{\mathcal{R}} \;\; ; \;\; \mathcal{P}_{T(C,\mathcal{R})^\perp}(M) = \mathcal{P}_{C^\perp} M \mathcal{P}_{\mathcal{R}^\perp}.$$

In analogy to the previous case with variable selection, the tangent space at a rank-r matrix with respect to $\mathcal{V}_{\text{low-rank}}(r)$ encodes — and is in one-to-one correspondence with — the row/column space structure at that point. Indeed, estimating

the row/column spaces of a low-rank matrix can be viewed equivalently as estimating the tangent space at that matrix with respect to a determinantal variety. With this notion in hand, we give our definition of true/false discoveries in low-rank estimation:

Definition 1. *Let $C^\star \subset \mathbb{R}^{p_1}$ and $\mathcal{R}^\star \subset \mathbb{R}^{p_2}$ denote the column and row spaces of a low-rank matrix in $\mathbb{R}^{p_1 \times p_2}$; in particular,* $\dim(C^\star) = \dim(\mathcal{R}^\star)$. *Given observations from a model parametrized by this matrix, let $(\hat{C}, \hat{\mathcal{R}}) \subset \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$ be an estimator of the pair of subspaces $(C^\star, \mathcal{R}^\star)$ with* $\dim(\hat{C}) = \dim(\hat{\mathcal{R}})$. *Then the* expected false discovery *of the estimator is defined as:*

$$\text{FD} = \mathbb{E}\left[\text{trace}\left(\mathcal{P}_{T(\hat{C},\hat{\mathcal{R}})}\mathcal{P}_{T(C^\star,\mathcal{R}^\star)^\perp}\right)\right], \tag{3.3}$$

and the power *of the estimator is defined as:*

$$\text{PW} = \mathbb{E}\left[\text{trace}\left(\mathcal{P}_{T(\hat{C},\hat{\mathcal{R}})}\mathcal{P}_{T(C^\star,\mathcal{R}^\star)}\right)\right]. \tag{3.4}$$

The expectations in both cases are with respect to randomness in the data employed by the estimator, and the tangent spaces $T(\hat{C}, \hat{\mathcal{R}}), T(C^\star, \mathcal{R}^\star)$ are as defined in (3.1).

With respect to our objective of identifying a suitable notion of discovery for low-rank estimation, the definitions of FD and of PW possess a number of favorable attributes. These definitions do not depend on a choice of basis for the tangent space $T(C^\star, \mathcal{R}^\star)$. Further, for the reasons described above, small changes in row/column space estimates lead to small changes in the performance of an estimator, as evaluated by FD and PW. Despite these definitions respecting the smooth structure underlying low-rank matrices, they specialize transparently to the usual discrete notion of true/false discoveries in the setting of variable selection if the underlying low-rank matrices are diagonal. We also have that the expected false discovery is bounded as $0 \leq \text{FD} \leq \dim(T(C^\star, \mathcal{R}^\star)^\perp)$ and the power is bounded as $0 \leq \text{PW} \leq \dim(T(C^\star, \mathcal{R}^\star))$, which is in agreement with the intuition that the spaces $T(C^\star, \mathcal{R}^\star)$ and $T(C^\star, \mathcal{R}^\star)^\perp$ represent the total true and false discoveries, respectively, that can be made by any estimator. Similarly, we observe that $\text{FD} + \text{PW} = \mathbb{E}[\dim(T(\hat{C}, \hat{\mathcal{R}}))]$, which is akin to the expected total discovery made by the estimator $(\hat{C}, \hat{\mathcal{R}})$.

We note that the definition of FD may be modified to obtain an analog of the *false discovery rate* [YH95], which is of interest in contemporary multiple testing as well as in high-dimensional estimation:

$$\text{FDR} = \mathbb{E}\left[\frac{\text{trace}\left(\mathcal{P}_{T(\hat{C},\hat{\mathcal{R}})}\mathcal{P}_{T(C^\star,\mathcal{R}^\star)^\perp}\right)}{\dim(T(\hat{C}, \hat{\mathcal{R}}))}\right].$$

We focus in the present paper on procedures that control the quantity FD by generalizing the stability selection approach of [MB10], and we discuss in Section 3.5 some challenges associated with controlling FDR in low-rank estimation.

Finally, while the main focus of this paper is on a false discovery framework for low-rank estimation in which we seek reliable estimates of both the row and column spaces, the geometric perspective outlined here can be adapted to settings in which one only seeks an estimate of the column-space of a data matrix. (Such a problem arises in hyperspectral imaging, as illustrated in Section 3.4.) In such situations, the ideas described previously can be extended as follows:

$$\widetilde{\mathrm{FD}} = \mathbb{E}\left[\mathrm{trace}\left(\mathcal{P}_{\hat{C}}\mathcal{P}_{C^{\star\perp}}\right)\right] \;\; ; \;\; \widetilde{\mathrm{PW}} = \mathbb{E}\left[\mathrm{trace}\left(\mathcal{P}_{\hat{C}}\mathcal{P}_{C^{\star}}\right)\right] \;\; ; \;\; \widetilde{\mathrm{FDR}} = \mathbb{E}\left[\frac{\mathrm{trace}\left(\mathcal{P}_{\hat{C}}\mathcal{P}_{C^{\star\perp}}\right)}{\dim(\hat{C})}\right]. \tag{3.5}$$

Here $C^\star \subset \mathbb{R}^p$ represents the population column space and $\hat{C} \subset \mathbb{R}^p$ is an estimator. These expressions can be derived by considering tangent spaces with respect to quotients of the determinantal variety under certain equivalence relations; supplementary material Section A provides the details.

From Tangent Space to Parameter Estimation

While the main focus of this paper is on a framework to evaluate and control the expected false discovery of structure estimated from data (quantified by tangent spaces), in many practical settings (e.g. some of the prediction tasks with real datasets in Section 4), the ultimate object of interest is a parameter. As such, we adopt the following viewpoint: first produce a tangent space with few false discoveries, and then solve a parameter estimation problem restricted to this tangent space. In this subsection, we assume that a suitable tangent space has been obtained, and present a simple approach to solve a parameter estimation problem restricted to this tangent space. Starting with the low-rank setting, let $T(C, \mathcal{R}) \subset \mathbb{R}^{p_1 \times p_2}$ be a tangent space that corresponds to column and row spaces $C \subset \mathbb{R}^{p_1}, \mathcal{R} \subset \mathbb{R}^{p_2}$, and given a collection of observations $\mathcal{D}$, we wish to solve the following optimization problem:

$$\hat{L} = \underset{L \in \mathbb{R}^{p_1 \times p_2}}{\mathrm{argmin}}\, \mathrm{Loss}\,(L\,;\,\mathcal{D}) \;\text{ subject to }\; T(\text{column-space}(L), \text{row-space}(L)) \subseteq T(C, \mathcal{R}), \tag{3.6}$$

in which the decision variable L is constrained to have a tangent space that lies within the prescribed tangent space $T(C, \mathcal{R})$. Furthermore, this constraint may be

simplified as follows. Suppose that the subspaces $\mathcal{R}, C$ are of dimension k. Let $U_C \in \mathbb{R}^{p_1 \times k}$ and $U_R \in \mathbb{R}^{p_2 \times k}$ be any matrices with columns spanning the spaces C and $\mathcal{R}$, respectively. Then one can check that the set $\{U_C M U_R' \mid M \in \mathbb{R}^{k \times k}\}$ is precisely the collection of matrices whose tangent spaces are contained in $T(C, \mathcal{R})$. Consequently (3.6) may be reformulated as:

$$\hat{L} = \underset{L \in \mathbb{R}^{p_1 \times p_2},\ M \in \mathbb{R}^{k \times k}}{\operatorname{argmin}} \ \text{Loss}\ (L\ ;\ \mathcal{D}) \quad \text{subject to} \quad L = U_C M U_R'. \tag{3.7}$$

Note that the constraint here is linear in the decision variables L, M. Consequently, an appealing property of (3.7) is that if the loss function $\text{Loss}(\cdot\ ;\ \mathcal{D})$ is convex, then (3.7) is a convex optimization problem. For example, when $\text{Loss}(\cdot\ ;\ \mathcal{D})$ is the squared loss, an optimal solution can be obtained in closed form.

Next, we consider the subspace estimation problem, where the task is to find an appropriate projection of the data $\mathcal{D} = \{y_i\}_{i=1}^n \subseteq \mathbb{R}^p$ onto the estimated subspace $C \subseteq \mathbb{R}^p$. A natural optimization program to determine the projected data is

$$\{\hat{z}_i\}_{i=1}^n = \underset{\{z_i\}_{i=1}^n \subset \mathbb{R}^p}{\operatorname{argmin}} \ \text{Loss}\ \left(\{z_i\}_{i=1}^n\ ;\ \mathcal{D}\right) \quad \text{subject to} \quad \text{span}\left(\{z_i\}_{i=1}^n\right) \in C. \tag{3.8}$$

Similar to the low-rank case, the constraint can be simplified to $\text{concatenation}(\{z\}_{i=1}^n) = UM'$, where $U \in \mathbb{R}^{p \times k}$ is a matrix with columns spanning the subspace C and $M \in \mathbb{R}^{n \times k}$ is an additional decision variable.

3.3 False Discovery Control via Subspace Stability Selection

Building on the discussion in the preceding section, our objective is the accurate estimation of the tangent space associated to a low-rank matrix, as this is in one-to-one correspondence with the row/column spaces of the matrix. In this section, we formulate an approach based on the stability selection procedure of [MB10] to estimate such a tangent space. We will also describe how this method can be specialized for problems involving subspace estimation.

Stability selection is a general technique to control false discoveries in variable selection. The procedure can be paired with any variable selection procedure as follows: instead of applying a selection procedure (e.g. the Lasso) to a collection of observations, one instead applies the procedure to many subsamples of the data and then chooses those variables that are most consistently selected in the subsamples. The virtue of the subsampling and averaging framework is that it provides control over the expected number of falsely selected variables (see Theorem 1 in [MB10] and Theorem 1 in [SS13]). We develop a generalization of this framework in which

existing row/column space selection procedures (based on any low-rank estimation procedure) are employed on subsamples of the data, and then these spaces are aggregated to produce a tangent space that provides false discovery control.

Subsampling procedure: Although our framework is applicable with general subsamples of the data, we adopt the subsampling method outlined in [SS13] in our experimental demonstrations and our theoretical analysis; in particular, given a dataset $\mathcal{D}$ and a positive (even) integer B, we consider B subsamples or bags obtained from $B/2$ *complementary partitions* of $\mathcal{D}$ of the form $\{(\mathcal{D}_{2i-1}, \mathcal{D}_{2i}) : i = 1, 2, 3, \ldots, B/2\}$, where $|\mathcal{D}_{2i-1}| = |\mathcal{D}|/2$ and $\mathcal{D}_{2i} = \mathcal{D} \backslash \mathcal{D}_{2i-1}$.

Setup for numerical demonstrations: For our numerical illustrations in this section, we consider the following stylized low-rank matrix completion problem. The population parameter $L^\star \in \mathbb{R}^{70\times 70}$ is a rank-10 matrix with singular values (and associated multiplicities) given by $1(x3), 0.5(x5)$, and $0.1(x2)$, and with row/column spaces sampled uniformly at random according to the Haar measure. We are given noisy observations $Y_{i,j} = L^\star_{i,j} + \epsilon_{i,j}$ with $\epsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$ and $(i, j) \in \Omega$, where $\Omega \subset \{1, \ldots, 70\}^2$ is chosen uniformly at random with $|\Omega| = 3186$. The variance σ^2 is chosen to set the signal-to-noise (SNR) ratio (defined as $\mathbb{E}[\|L^\star\|_F / \|\epsilon\|_F]$) is at a desired level, and this is specified later. As our subsamples, we consider a collection of $B = 100$ subsets each consisting of $|\Omega|/2 = 1593$ entries obtained from 50 random complementary partitions of the data. On each subsample — corresponding to a subset $S \subset \Omega$ of observations with $|S| = 1593$ — we employ the following convex program [SS05; CR09]:

$$\hat{L} = \underset{L\in\mathbb{R}^{70\times 70}}{\operatorname{argmin}} \sum_{\{i,j\}\in S} \|(L - Y)_{i,j}\|_F^2 + \lambda \|L\|_\star, \tag{3.9}$$

and we report the tangent space $T(\texttt{column-space}(\hat{L}), \texttt{row-space}(\hat{L}))$ as the estimate associated to the subsample. Here $\lambda > 0$ is a regularization parameter (to be specified later) and $\|\cdot\|_\star$ is the nuclear norm (the sum of the singular values), which is commonly employed to promote low-rank structure in a matrix [Faz02]. We emphasize that our development is relevant for general low-rank estimation problems, and this problem is merely for illustrative purposes in the present section; for a more comprehensive set of experiments in more general settings, we refer the reader to Section 3.4.

Stable Tangent Spaces

The first step in stability selection is to combine estimates of significant variables obtained from different subsamples. This is accomplished by computing for each variable the frequency with which it is selected across the subsamples. We generalize this idea to our context via projection operators onto tangent spaces as follows:

Definition 2 (Average Projection Operator)**.** *Suppose $\hat{T}$ is an estimator of a tangent space of a low-rank matrix, and suppose further that we are given a set of observations $\mathcal{D}$ and a corresponding collection of subsamples $\{\mathcal{D}_i\}_{i=1}^{B}$ with each $\mathcal{D}_i \subset \mathcal{D}$. Then the* average projection operator *of the estimator $\hat{T}$ with respect to the subsamples $\{\mathcal{D}_i\}_{i=1}^{B}$ is defined as:*

$$\mathcal{P}_{avg} \triangleq \frac{1}{B}\sum_{i=1}^{B} \mathcal{P}_{\hat{T}(\mathcal{D}_i)}, \tag{3.10}$$

where $\hat{T}(\mathcal{D}_i)$ is the tangent space estimate based on the subsample $\mathcal{D}_i$.

Here $\mathcal{P}_{\mathsf{avg}} : \mathbb{R}^{p_1 \times p_2} \to \mathbb{R}^{p_1 \times p_2}$ is self-adjoint, and its eigenvalues lie in the interval $[0, 1]$ as each $\mathcal{P}_{\hat{T}(\mathcal{D}_i)}$ is self-adjoint with eigenvalues equal to 0 or 1. To draw a comparison with variable selection, the tangent spaces in that case correspond to subspaces spanned by coordinate vectors in $\mathbb{R}^p$ (with p being the total number of variables of interest) and the average projection operator is a diagonal matrix of size $p \times p$, with each entry on the diagonal specifying the fraction of subsamples in which a particular variable is selected. The virtue of averaging over tangent spaces estimated across a large number of subsamples is that most of the 'energy' of the average projection operator $\mathcal{P}_{\mathsf{avg}}$ tends to be better aligned with the underlying population tangent space. We illustrate this point next with an example.

Illustration: the value of averaging projection maps — Consider the stylized low-rank matrix completion problem described at the beginning of Section 3.3. To support the intuition that the average projection matrix $\mathcal{P}_{\mathsf{avg}}$ has reduced in energy in directions corresponding to $T^{\star\perp}$ (i.e., the orthogonal complement of the population tangent space), we compare the quantities $\mathbb{E}\left[\text{trace}\left(\mathcal{P}_{\mathsf{avg}}\mathcal{P}_{T^{\star\perp}}\right)\right]$ and $\mathbb{E}\left[\text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D})}\mathcal{P}_{T^{\star\perp}}\right)\right]$, where the expectation is computed over 100 instances. Generically speaking, the operator $\mathcal{P}_{\mathsf{avg}}$ is not a projection operator onto a tangent space and thus the quantity $\mathbb{E}\left[\text{trace}\left(\mathcal{P}_{\mathsf{avg}}\mathcal{P}_{T^{\star\perp}}\right)\right]$ is not a valid false discovery, rather it evaluates the average false discovery over the subsampled models. The second quantity, $\mathbb{E}\left[\text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D})}\mathcal{P}_{T^{\star\perp}}\right)\right]$, is based on employing the nuclear norm

regularization procedure on the full set of observations. The variance σ is selected so that SNR = $\{0.8, 1.6\}$. As is evident from Figure 3.1, $\mathbb{E}\left[\text{trace}\left(\mathcal{P}_{\text{avg}}\mathcal{P}_{T^{\star\perp}}\right)\right]$ is smaller than $\mathbb{E}\left[\text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D})}\mathcal{P}_{T^{\star\perp}}\right)\right]$ for the entire range of λ, with the gap being larger in the low SNR regime. In other words, averaging the subsampled tangent spaces reduces energy in the directions spanned by $T^{\star\perp}$.

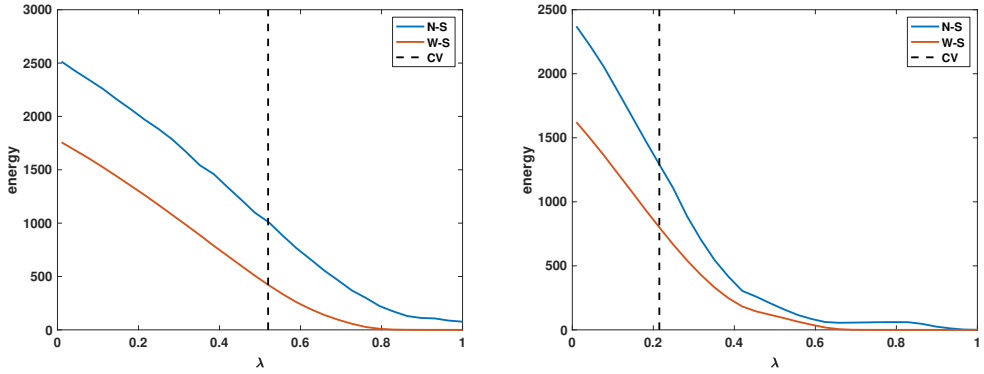


Figure 3.1: The quantities $\mathbb{E}\left[\text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D})}\mathcal{P}_{T^{\star\perp}}\right)\right]$ (in blue) and $\mathbb{E}\left[\text{trace}\left(\mathcal{P}_{\text{avg}}\mathcal{P}_{T^{\star\perp}}\right)\right]$ (in red) as a function of λ for SNR = 1.6 (right) and SNR = 0.8 (left) in the synthetic matrix completion setup. The cross-validated choice of λ is shown as the dotted black line. Here 'N-S' denotes no subsampling and 'W-S ' denotes with subsampling.

While the average projection aggregated over many subsamples appears to have less energy in $T^{\star\perp}$, this operator is not a proper projection. Thus it still remains for us to identify a single tangent space as our estimate from $\mathcal{P}_{\text{avg}}$. We formulate the following criterion to establish a measure of closeness between a single tangent space and the aggregate over subsamples:

Definition 3 (Stable Tangent Spaces)**.** *Suppose $\hat{T}$ is an estimator of a tangent space of a low-rank matrix, and suppose further that we are given a set of observations $\mathcal{D}$ and a corresponding collection of subsamples $\{\mathcal{D}_i\}_{i=1}^{B}$ with each $\mathcal{D}_i \subset \mathcal{D}$. For a parameter $\alpha \in (0, 1)$, the set of* stable tangent spaces *is defined as*

$$\mathcal{T}_\alpha \triangleq \left\{T \mid \sigma_{min}\left(\mathcal{P}_T\mathcal{P}_{\text{avg}}\mathcal{P}_T\right) \geq \alpha \text{ and } T \textit{is a tangent space to a determinantal variety}\right\}, \tag{3.11}$$

where $\mathcal{P}_{\text{avg}}$ is computed based on Definition 2.

As the spectrum of $\mathcal{P}_{\text{avg}}$ lies in the range $[0, 1]$, this is also the only meaningful range of values for α. The set $\mathcal{T}_\alpha$ consists of all those tangent spaces T to a determinantal variety such that the Rayleigh quotient of every nonzero element of T

with respect to $\mathcal{P}_{\mathtt{avg}}$ is at least α. To contrast again with variable selection, we note that both $\mathcal{P}_T$ and $\mathcal{P}_{\mathtt{avg}}$ are diagonal matrices in that case (and thus simultaneously diagonalizable). As a consequence, the set $\mathcal{T}_\alpha$ has a straightforward characterization for variable selection problems; it consists of subspaces spanned by any subset of standard basis vectors corresponding to variables that are selected as significant in at least an α fraction of the subsamples.

As averaging the tangent spaces obtained from the subsampled data reduces energy in the directions contained in $T^{\star\perp}$, each element of $\mathcal{T}_\alpha$ is also far from being closely aligned with $T^{\star\perp}$ (for large values of α). We build on this intuition by proving next that a tangent space estimator that selects any element of $\mathcal{T}_\alpha$ provides false discovery control at a level that is a function of α. In Section 6.1 we describe efficient methods to choose an element of $\mathcal{T}_\alpha$.

As a final remark, the ideas described here can be readily applied to subspace estimation problems. Specifically, the average projection operator $\mathcal{P}^{C}_{\mathtt{avg}}$ in (3.10) is the average of projection matrices onto column-space estimates obtained from $n/2$ subsamples. Furthermore, the stable subspace set (3.11) is the collection of subspaces $C \in \mathbb{R}^p$ that satisfy the criterion $\sigma_{\mathtt{min}}(\mathcal{P}_C \mathcal{P}^{C}_{\mathtt{avg}} \mathcal{P}_C) \geq \alpha$.

False Discovery Control of Stable Tangent Spaces: Theoretical Analysis

Setup: Suppose we have a rank r matrix $L^\star \in \mathbb{R}^{p_1 \times p_2}$ with associated tangent space $T^\star$, and we are given i.i.d. observations from a model parametrized by $L^\star$. The objective is to obtain an accurate estimate of $T^\star$. We intentionally keep our discussion broad so our results are relevant for a wide range of low-rank estimation problems, e.g., low-rank matrix completion, factor analysis, etc. Let $\hat{T}$ denote a tangent space estimator that operates on samples drawn from the model parametrized by $L^\star$. Let $\mathcal{D}(n)$ denote a dataset consisting of n i.i.d observations from this models; we assume that n is even and that we are given B subsamples $\{\mathcal{D}_i\}_{i=1}^{B}$ via complementary partitions of $\mathcal{D}(n)$.

In this section, we present a master theorem to control false discoveries of stable tangent spaces under the sole assumption that the dataset consists of n i.i.d observations. Under additional assumptions of exchangeability and better than random guessing, we specialize the master theorem to obtain a more refined false discovery bound that is similar in spirit to [MB10]. Finally, inspired by Theorem 1 of [SS13], we specialize our master theorem to produce a bag-independent false discovery bound that is valid for any $B \geq 2$. We note that the theoretical results in this section

naturally extend to settings where one only cares about the column-space of the data matrix. For precise statements in that setting, refer to the supplementary material Section A.

Before proceeding, we remark that a significant aspect in our analysis is the role played by the *commutator* between projection operators onto tangent spaces. Formally, the commutator between self-adjoint operators A, B is denoted $[A, B] = AB - BA$, and this function evaluates how far away A, B are from commuting with each other. If A and B are two projection operators associated to subspaces T_1 and T_2, one can check that $\left\|\left[\mathcal{P}_{T_1}, \mathcal{P}_{T_2}\right]\right\|_F^2 = \frac{1}{2}\sum_{i=1}^{\dim(T_1)} \sin(2\theta_i)^2$ and $\left\|\left[\mathcal{P}_{T_1}, \mathcal{P}_{T_2}\right]\right\|_2^2 = \frac{1}{4}\max_i \sin(2\theta_i)$ where $\{\theta_i\}_{i=1}^{\dim(T_1)}$ are the principal angles between T_1 and T_2. This feature in our analysis is a departure from the setting of variable selection in which the projection operators commute (e.g. $\theta_i = \{0, \frac{\pi}{2}\}$).

Theorem 4 (False Discovery Control of Subspace Stability Selection)**.** *Consider the setup described above. Let $\hat{T}(\mathcal{D}_j)$ denote the tangent space estimates obtained from each of the subsamples, and let $\mathcal{P}_{\mathsf{avg}}$ denote the associated average projection operator computed via* (3.10) *across B complementary bags. Fix any $\alpha \in (0, 1)$ and let T denote any selection of an element of the associated set $\mathcal{T}_\alpha$ of stable tangent spaces. Then we have that for any fixed collection $\{M_i\}_{i=1}^{\dim(T^{\star\perp})}$ of orthonormal basis set for $T^{\star\perp}$:*

$$\mathbb{E}\left[\text{trace}\left(\mathcal{P}_T \mathcal{P}_{T^{\star\perp}}\right)\right] \quad \leq \quad F + 4\sqrt{1-\alpha}\kappa_{bag} + 2(1-\alpha)\mathbb{E}[\dim(T)]. \tag{3.12}$$

Here, the quantities F and κ_{bag} are given by

$$\begin{aligned} F &= \min\{ \sum_{i=1}^{\dim(T^{\star\perp})} \mathbb{E}[\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}(M_i)\|_F]^2, \mathbb{E}[\text{trace}(\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\mathcal{P}_{T^{\star\perp}})^{1/2}]^2\} \\ \kappa_{bag} &= \mathbb{E}\Bigg[\min\Bigg\{ \sqrt{\dim(T)}\sqrt{\frac{1}{B}\sum_{j=1}^{B} \|[\mathcal{P}_{T^{\star\perp}}, \mathcal{P}_{\hat{T}(\mathcal{D}_j)}]\|_F^2} \\ &\quad , \quad \dim(T)\sqrt{\frac{1}{B}\sum_{j=1}^{B} \|[\mathcal{P}_{T^{\star\perp}}, \mathcal{P}_{\hat{T}(\mathcal{D}_j)}]\|_2^2} \Bigg\} \Bigg] \end{aligned}$$

and the expectation is with respect to randomness in the observations. The set $\mathcal{D}(n/2)$ denotes a collection of $n/2$ i.i.d. observations drawn from the model parametrized by $L^\star$.

The proof of Theorem 4 can be found in supplementary material Section A.1. Theorem 4 states that the expected false discovery of a stable tangent space is

bounded by the sum of three quantities. The first term F characterizes the quality of the estimator on subsamples consisting of n/2 observations. The terms $4\sqrt{1-\alpha}\kappa_{\text{bag}}$ and $2(1-\alpha)\mathbb{E}[\dim(T)]$ are functions of the user specified parameters α, number of bags B and a commutator term. As expected, choosing α closer to 1 leads to a smaller amount of false discovery and $\alpha > 1/2$ for (3.12) to be non-vacuous since the following bound always holds $\mathbb{E}\left[\text{trace}(\mathcal{P}_T\mathcal{P}_{T^{\star\perp}})\right] \leq \mathbb{E}[\dim(T)]$.

*R*emark 1: Both of the quantities $\sum_{i=1}^{\dim(T^{\star\perp})}\mathbb{E}[\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}(M_i)\|_F]^2$ and $\mathbb{E}[\text{trace}(\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\mathcal{P}_{T^{\star\perp}})^{1/2}]^2$ highlight the variance reduction due to bagging. For the ease of reading, we define shorthand notation: let $\beta \in \mathbb{R}^{\dim(T^{\star\perp})}$ with $\beta_i \triangleq \|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}(M_i)\|_F$ so that $\sum_{i=1}^{\dim(T^{\star\perp})}\mathbb{E}[\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}(M_i)\|_F]^2 = \text{trace}(\mathbb{E}[\beta]\mathbb{E}[\beta]')$, and let $\xi \triangleq \text{trace}(\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\mathcal{P}_{T^{\star\perp}})^{1/2}$. Jensen's inequality yields $\mathbb{E}[\xi]^2 \leq \mathbb{E}[\xi^2] = \mathbb{E}[\text{trace}(\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\mathcal{P}_{T^{\star\perp}})]$ where the improvement of bagging over just using $\mathcal{D}(n/2)$ once is precisely measured by $\text{var}(\xi)$. Once again, by Jensen's inequality, $\text{trace}(\mathbb{E}[\beta]\mathbb{E}[\beta]') \leq \mathbb{E}[\text{trace}(\beta\beta')] = \mathbb{E}[\text{trace}(\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\mathcal{P}_{T^{\star\perp}})]$, where the variance reduction is measured by $\text{trace}(\text{cov}(\beta))$. Naturally one may ask: what are the scenarios in which $\text{var}(\delta)$ and $\text{trace}(\text{cov}(\beta))$ are enhanced? Given a fixed $\mathbb{E}[\xi]$, the Bhatia–Davis inequality states that $\text{var}(\xi)$ is enhanced when the distribution of ξ concentrates around 0 and $\sqrt{\dim(T^{\star\perp})}$ (i.e. most discoveries are either true or false). Similarly, for any i, given a fixed $\mathbb{E}[\beta]$, $\text{trace}(\text{cov}(\beta))$ is enhanced when the distribution of β_i concentrates around 0 or 1 (i.e. the estimate $\hat{T}(\mathcal{D}(n/2))$ is mostly aligned with or orthogonal to $M_i \in T^{\star\perp}$). In Section 3.4, we use this intuition to provide synthetic experiments that illustrate the improvement (in terms of false discovery) of a stable tangent space over using the original estimator without subsampling.

Remark 2: In general, the terms $\sum_{i=1}^{\dim(T^{\star\perp})}\mathbb{E}[\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}(M_i)\|_F]^2$ and $\mathbb{E}[\text{trace}(\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\mathcal{P}_{T^{\star\perp}})^{1/2}]^2$ inside F_1, which measure the quality of the estimator, are incomparable. The quantity $\mathbb{E}[\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}(M_i)\|_F]^2$ is basis dependent and measures the energy of a tangent space estimate (obtained from $n/2$ observations) along each direction $M_i \in T^{\star\perp}$ and then aggregates. This metric is sensible in scenarios where a particular choice of $\{M_i\}_{i=1}^{\dim(T^{\star\perp})}$ is natural, such as variable selection where a coordinate basis has a clear interpretation. On the other hand, $\mathbb{E}[\text{trace}(\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\mathcal{P}_{T^{\star\perp}})^{1/2}]^2$ is basis independent and is more natural in problem settings when no particular choice of a basis is sensible.

*R*emark 3: The quantity κ_{bag} is an increasing function of the energy of the average commutator between the projection operators of the tangent spaces obtained from subsamples and $\mathcal{P}_{T^{\star\perp}}$. Recall that for principal angles $\{\theta_i\}_{i=1}^{\dim(T^{\star\perp})}$ between $\hat{T}(\mathcal{D}(n/2))$ and $T^{\star\perp}$, $\|[\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}, \mathcal{P}_{T^{\star\perp}}]\|_F^2 = \frac{1}{2}\sum_{i=1}^{\dim(T^{\star\perp})}\sin(2\theta_i)^2$ and

$\|[\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}, \mathcal{P}_{T^{\star\perp}}]\|_2^2 \leq \frac{1}{4}\max_i \sin(2\theta_i)$. In other words, the quantity κ_{bag} is small when the principal angles between $\hat{T}(\mathcal{D}(n/2))$ and $T^{\star\perp}$ are close to 0 or $\frac{\pi}{2}$. The involvement of the commutator term κ_{bag} highlights a key difference between variable selection and low-rank estimation. In particular, in the variable selection setting, projection operators onto tangent spaces commute (e.g. $\theta_i = \{0, \frac{\pi}{2}\}$) and as a result κ_{bag} vanishes. On the other hand, since the determinantal variety is locally smooth, the projection matrices onto tangent spaces of this variety will generically not commute. We discuss in Remark 4 that the commutativity property in the variable selection setting enables additional simplifications for obtaining even tighter bounds.

*R*emark 4: The bound (3.12) is also valid in the setting of variable selection. However, by exploiting the fact that projection matrices of tangent spaces to varieties of sparse vectors commute, one is able to choose a basis (the standard basis) that simultaneously diagonalizes all these matrices, which leads to certain simplifications as well as an eventual tighter bound on the expected false discovery in variable selection. Specifically, letting $\{M_i\}_{i=1}^{\dim(T^{\star\perp})}$ be collection of standard basis elements that form orthonormal basis set for $T^{\star\perp}$, one can modify the proof of Theorem 4 to obtain the following bound:

$$\begin{aligned}\mathbb{E}\left[\text{trace}\left(\mathcal{P}_T\mathcal{P}_{T^{\star\perp}}\right)\right] &\leq \sum_{i=1}^{\dim(T^{\star\perp})} \frac{\mathbb{E}\left[\left\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}(M_i)\right\|_F\right]^2}{2\alpha - 1} \\ &= \sum_{i=1}^{\dim(T^{\star\perp})} \frac{\mathbb{P}[i'\text{th null selected by } \hat{T}(\mathcal{D}(n/2))]}{2\alpha - 1}. \quad (3.13)\end{aligned}$$

We prove (3.13) in the supplementary material Section A, which holds in conjunction with (3.12). The second line here follows from the observations that $\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}$ is a diagonal projection matrix and that M_i is also diagonal with only one non-zero element on the diagonal. Thus, the simultaneous diagonalizability property also leads to the conceptually appealing interpretation that the overall expected false discovery for the special case of variable selection can be bounded in terms of the probability that the procedure Tb selects null variables on subsamples. The final expression (3.13) matches precisely Theorem 1 of [SS13]. As a final comparison between the low-rank estimation and variable selection settings, notice that once the commutator term F_3 vanishes in the variable selection setting, the dependence on α in (3.13) becomes multiplicative, as opposed to additive in (3.12) and (3.13).‘ In particular, in the low-rank case even if the estimator $\hat{T}$ performs exceedingly well on the subsamples, the expected false discovery may still be large depending on

the choice of α and $\dim(T^{\star\perp})$; in contrast, for variable selection if the estimator $\hat{T}$ performs exceedingly well on the subsamples, the expected false discovery is small provided α is close to 1. This distinction is fundamental to the geometry underlying the sparse and determinantal varieties. Specifically, in the low-rank case even if $\mathcal{P}_{\mathsf{avg}} \approx \mathcal{P}_{T^\star}$ the set of stable tangent spaces $\mathcal{T}_\alpha$ necessarily includes many tangent spaces that are near the population tangent space $T^\star$ but are not perfectly aligned with it. This is due to the fact that the collection of row/column spaces forms a Grassmannian manifold rather than a finite/discrete set. On the other hand, if $\mathcal{P}_{\mathsf{avg}} \approx \mathcal{P}_{T^\star}$ in variable selection, the only elements of the set of stable tangent spaces (for suitable α) are those corresponding to subsets of the true significant variables.

The false discovery bound (3.12) of Theorem 4 holds for tangent spaces with respect to any variety. We next refine (3.12) to exploit the structure of the determinantal variety to obtain a false discovery bound under additional assumptions. Specifically, we consider the following assumptions on the low-rank estimator and the data generation process:

$$\begin{aligned} &\text{Assumption 1: } \frac{\mathbb{E}\left[\operatorname{trace}\left(\mathcal{P}_{T^{\star\perp}}\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\right)\right]}{\dim(T^{\star\perp})} \leq \frac{\mathbb{E}\left[\operatorname{trace}\left(\mathcal{P}_{T^\star}\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\right)\right]}{\dim(T^\star)} \\ &\text{Assumption 2: distribution of } \left(\mathcal{P}_{\hat{C}(\mathcal{D}(n/2))} M \mathcal{P}_{\hat{R}(\mathcal{D}(n/2))}\right) \text{ is the same } \forall M \in T^{\star\perp} \\ &\qquad \text{with } \operatorname{rank}(M) = 1 \;\&\; \|M\|_F = 1 \end{aligned} \tag{3.14}$$

Assumption 1 states that the estimator's normalized power is greater than its normalized false discovery and Assumption 2 states that the energy of the estimate $\hat{T}(\mathcal{D}(n/2))$ onto any rank-1 element in the collection $\{M_i\}_{i=1}^{\dim(T^{\star\perp})}$ is the same in distribution. To better understand (3.14), it is instructive to consider its specialization in variable selection. Specifically, letting $\hat{T}(\mathcal{D}(n/2))$ and $T^\star$ be tangent spaces to the sparse variety, Assumption 1 reduces to precisely the "better than random guessing" assumption employed in [MB10]. With regards to Assumption 2, [MB10] place the condition that the variables $\{\mathbb{I}_{k\in\hat{S}(\mathcal{D}(n/2))}, k \in S^{\star c}\}$ are exchangeable. This assumption implies that the distribution of $\mathbb{I}_{k\in\hat{S}(\mathcal{D}(n/2))}$ is the same for all $k \in S^{\star c}$. Replacing $\operatorname{rank}(M) = 1$ with $\operatorname{card}(M) = 1$ in (3.14) so that each M is a standard basis element, Assumption 2 reduces to exactly the same condition. We demonstrate in supplementary material Section A that Assumptions 1 and 2 in (3.14) are satisfied in some natural model ensembles and estimators.

Under Assumptions 1 and 2 in (3.14), we next prove a refined false discovery

bound.

Proposition 1 (Refined False Discovery Control)**.** *Suppose that Assumptions 1 and 2 in* (3.14) *are satisfied. Let the average number of discoveries from $n/2$ observations be denoted by $q := \mathbb{E}[\dim(\hat{T}(\mathcal{D}(n/2)))]$. Then, for any rank $-$ 1 $M \in T^{\star\perp}$ with $\|M\|_F = 1$, the false discovery of a stable tangent space is bounded by:*

$$\mathbb{E}\left[\text{trace}\left(\mathcal{P}_T \mathcal{P}_{T^{\star\perp}}\right)\right] \leq \frac{q^2}{p_1 p_2} + f\left(\kappa_{indiv}\right) + 4\sqrt{1-\alpha}\kappa_{bag} + 2(1-\alpha)\mathbb{E}[\dim(T)] \quad (3.15)$$

where $\kappa_{indiv} := \mathbb{E}\left[\|[\mathcal{P}_{\text{span}(M)}, \mathcal{P}_{T^{\star\perp}}]\|_F\right]$ and $f(\kappa_{indiv}) = p_1 p_2 \kappa_{indiv}^2 + 2q\kappa_{indiv}$.

*R*emark 5: The proof of this proposition can be found in supplementary material Section A. The bound in (3.15) resembles Theorem 4 with $F \leq \frac{q^2}{p_1 p_2} + f\left(\kappa_{\text{indiv}}\right)$. We will next analyze each individual term in (1). The quantity q in (3.15) can be approximated by $q \approx \mathbb{E}[\text{trace}(\mathcal{P}_{\mathsf{avg}})$ and thus be tuned by the practitioner to be on the order $O(p_1 + p_2)$. Assuming that p_1 and p_2 are both large and in the same scale, then the first term in (3.15) will be on the order $O(1)$, which is a substantial reduction from q. The fourth term in (3.15) can be controlled by choosing α sufficiently close to 1 and noting that $\mathbb{E}[\dim(T)] \leq \frac{1}{\alpha}\mathbb{E}[\sigma_{\min}(\mathcal{P}_T \mathcal{P}_{\mathsf{avg}} \mathcal{P}_T)] \leq \frac{1}{\alpha}\mathbb{E}[\text{trace}(\mathcal{P}_{\mathsf{avg}})] \leq \frac{q}{\alpha}$. The second and third quantities are increasing functions of the commutator terms κ_{bag} and κ_{indiv} with the terms vanishing when $\kappa_{\text{ind}} = \kappa_{\text{bag}} = 0$. To get a handle of these quantities, the property $\|[\mathcal{P}_{S_1}, \mathcal{P}_{S_2}]\| \leq \frac{1}{2}$ for any two subspaces S_1 and S_2 leads to the bound $\kappa_{\text{bag}} \leq \frac{q}{2}$. Since the bound in (3.15) holds for any rank-1 $M \in T^{\star\perp}$ with $\|M\|_F = 1$, it suffices to find a single $M \in T^{\star\perp}$ with $\|M\|_F = 1$ to control κ_{indiv}. As such, we propose the following data-driven heuristic to approximate κ_{indiv}: let u be the smallest singular vector of $\mathcal{P}^{\mathcal{C}}_{\mathsf{avg}}$ with corresponding singular value δ_C, v be the smallest singular vector of $\mathcal{P}^{\mathcal{R}}_{\mathsf{avg}}$ with corresponding singular value δ_R, and $\tilde{M} = uv'/(\|v\|\|u\|)$. Setting $\delta = \max\{\delta_C, \delta_R\}$, one can check that there exists a rank-1 direction $M \in T^{\star\perp}$ such that the the cosine of the angle between $\text{span}(\tilde{M})$ and $\text{span}(M)$ is lower-bounded by $1 - (\mathbb{E}[\sqrt{1-\cos(\sigma_r)^2)}]^2 + 2(1 - \delta + \sqrt{1-\delta}))^2$, where σ_r is the maximum of the r-th principal angle between $C^\star$ and & $C(\hat{\mathcal{D}}(n/2)$ and between $\mathcal{R}^\star$ and & $\mathcal{R}(\hat{\mathcal{D}}(n/2)$. In other words, if the estimates $\hat{\mathcal{C}}(\mathcal{D}(n/2))$ and $\hat{\mathcal{R}}(\mathcal{D}(n/2))$ have good power, one can ensure that $\text{span}(\tilde{M})$ is a close approximation to $\text{span}(M)$ with the degree of proximity controlled by δ. We then obtain the following data-driven approximation $\kappa_{\text{indiv}} = \frac{1}{B}\sum_{j=1}^{B} \|[\mathcal{P}_{\hat{T}(\mathcal{D}_j)}, \mathcal{P}_{\text{span}(\tilde{M})}]\|_F$.

*R*emark 6: In the setting where all the low-rank matrices are diagonal (e.g. variable selection), the commutator terms $\kappa_{\text{bag}} = \kappa_{\text{indiv}} = 0$ since projection operators

commute. Furthermore, as described in Remark 4, by selecting the basis elements M_i carefully so that all the projection operators in these settings are simultaneously diagonalizable, the terms $\frac{q^2}{p_1 p_2} + 2(1-\alpha)\mathbb{E}[\dim(T)]$ can be modified to the multiplicative bound in terms of α. Specifically, the expected false discovery can be bounded by $\frac{q^2}{2(1-\alpha)p_1 p_2}$ which precisely matches Theorem 1 of [MB10].

Finally, the false discovery bound (3.12) in Theorem 4 depends prominently on the number of bags. Inspired by [SS13], we specialize master Theorem 4 to produce a false discovery bound that is independent of the number of bags.

Proposition 2 (Bag Independent Result)**.** *For any $B \geq 2$, false discovery of the stable tangent space is bounded by*

$$\mathbb{E}\left[\operatorname{trace}\left(\mathcal{P}_T \mathcal{P}_{T^{\star\perp}}\right)\right] \leq F + \frac{2q}{\alpha}(1-\alpha+\sqrt{1-\alpha}) \tag{3.16}$$

and under Assumptions 1 and 2 in (3.14)*, $F \leq \frac{q^2}{p_1 p_2} + f\left(\kappa_{indiv}\right)$.*

Remark 7: The proof of this result can be found in supplementary material Section A. Similar to the bound in Theorem 4, the bound (3.16) is a function of the quality of the estimator as well as α. In contrast, the bound (3.16) hold for any $B \geq 2$, and as a result, can be looser than (3.12). Specifically, the term $4\sqrt{1-\alpha}\kappa_{\text{bag}} + 2(1-\alpha)\mathbb{E}[\dim(T)]$ in (3.12) is bounded by $\frac{2q}{\alpha}(1-\alpha+\sqrt{1-\alpha})$, with the operating regime of (3.16) decreasing from $\alpha \geq \frac{1}{2}$ to $\alpha \gtrapprox 0.9$, as otherwise the bound exceeds q. We prove in supplementary material Section A that the operating regime can be increased to $\alpha \gtrapprox 0.84$ by replacing F in (3.16) with $\mathbb{E}[\operatorname{trace}(\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\mathcal{P}_{T^{\star\perp}})^{1/2}]^2$ and $\frac{2q}{\alpha}(1-\alpha+\sqrt{1-\alpha})$ with $\frac{3q}{\sqrt{2}\alpha}\sqrt{1-\alpha}$. Nonetheless, these bag independent results may be enlightening in the regimes where the signal strength is large so that high α may be considered.

Subspace Stability Selection Algorithm

As described in the previous subsection, every tangent space in $\mathcal{T}_\alpha$ provides control on the expected false discovery. The goal then is to select an element of $\mathcal{T}_\alpha$ to optimize power. A natural approach to achieve this objective is to choose a tangent space of largest dimension from $\mathcal{T}_\alpha$ to maximize the total discovery.

Consider the following optimization problem for each $r = 1, \ldots, \min\{p_1, p_2\}$:

$$T_{\texttt{OPT}}(r) = \underset{T \text{ tangent space to a point in } \mathcal{V}_{\text{low-rank}}(r)}{\operatorname{argmax}} \sigma_{\min}\left(\mathcal{P}_T \mathcal{P}_{\texttt{avg}} \mathcal{P}_T\right). \tag{3.17}$$

A conceptually appealing approach to select an optimal tangent space is via the following optimization problem:

$$T_{\texttt{OPT}} \in \operatorname*{argmax}_{T \in T_{\texttt{OPT}}(r) \cap \mathcal{T}_\alpha} r, \tag{3.18}$$

where by construction, the set $T_{\texttt{OPT}}(r) \cap \mathcal{T}_\alpha$ is non-empty if $\mathcal{T}_\alpha$ is a non-empty set. In the case of variable selection, this procedure would result in the selection of all those variables that are estimated as being significant in at least an α fraction of the bags, which is in agreement with the procedure of [MB10]. In our setting of low-rank estimation, however, we are not aware of a computationally tractable approach to solve the problem (3.17). The main source of difficulty lies in the geometry underlying the collection of tangent spaces to determinantal varieties. In particular, solving (3.17) in the case of variable selection is easy because the operators $\mathcal{P}_T, \mathcal{P}_{\texttt{avg}}$ are both diagonal (and hence trivially simultaneously diagonalizable) in that case; as a result, one can decompose (3.17) into a set of one-variable problems. In contrast, the operators $\mathcal{P}_T, \mathcal{P}_{\texttt{avg}}$ are not simultaneously diagonalizable in the low-rank case, and consequently there doesn't appear to be any clean separability in (3.17) in the general setting with determinantal varieties.

We describe next a heuristic to approximate (3.17). Our approximation entails computing optimal row-space and column-space approximations from the bags separately rather than in a combined fashion via tangent spaces. Specifically, suppose $\{(\hat{C}(\mathcal{D}_i), \hat{\mathcal{R}}(\mathcal{D}_i))\}_{i=1}^B$ denote the row/column space estimates from B subsamples $\{\mathcal{D}_i\}_{i=1}^B \subset \mathcal{D}$ of the data. We average the projection operators associated to these row/column spaces:

$$\mathcal{P}^{C}_{\texttt{avg}} = \frac{1}{B}\sum_{i=1}^{B} \mathcal{P}_{\hat{C}(\mathcal{D}_i)}, \qquad \mathcal{P}^{\mathcal{R}}_{\texttt{avg}} = \frac{1}{B}\sum_{i=1}^{B} \mathcal{P}_{\hat{\mathcal{R}}(\mathcal{D}_i)}. \tag{3.19}$$

Note that the average projection operator $\mathcal{P}_{\texttt{avg}}$ based on estimates from subsamples of tangent spaces to determinantal varieties is a self-adjoint map on the space $\mathbb{R}^{p_1 \times p_2}$. In contrast, the average operators $\mathcal{P}^{C}_{\texttt{avg}}$ and $\mathcal{P}^{\mathcal{R}}_{\texttt{avg}}$ are self-adjoint maps on the spaces $\mathbb{R}^{p_1}$ and $\mathbb{R}^{p_2}$, respectively. Based on these separate column-space and row-space averages, we approximate (3.17) as follows:

$$\begin{aligned} T_{\texttt{approx}}(r) = T\Bigg(& \operatorname*{argmax}_{C \subset \mathbb{R}^{p_2} \text{ subspace of dimension } r} \sigma_{\min}\left(\mathcal{P}_C \mathcal{P}^{C}_{\texttt{avg}} \mathcal{P}_C\right) \\ & , \operatorname*{argmax}_{\mathcal{R} \subset \mathbb{R}^{p_1} \text{ subspace of dimension } r} \sigma_{\min}\left(\mathcal{P}_{\mathcal{R}} \mathcal{P}^{\mathcal{R}}_{\texttt{avg}} \mathcal{P}_{\mathcal{R}}\right) \Bigg). \end{aligned} \tag{3.20}$$

The advantage of this latter formulation is that the inner-optimization problems of identifying the best row-space and column-space approximations of rank r can be computed tractably. In particular, the optimal column-space (resp. row-space) approximation of dimension r is equal to the span of the eigenvectors corresponding to the r largest eigenvalues of $\mathcal{P}^{C}_{\mathsf{avg}}$ (resp. $\mathcal{P}^{\mathcal{R}}_{\mathsf{avg}}$). We have that $\sigma_{\min}\left(\mathcal{P}_{T_{\mathtt{approx}}(r)}\mathcal{P}_{\mathsf{avg}}\mathcal{P}_{T_{\mathtt{approx}}(r)}\right) \leq \sigma_{\min}\left(\mathcal{P}_{T_{\mathtt{OPT}}(r)}\mathcal{P}_{\mathsf{avg}}\mathcal{P}_{T_{\mathtt{OPT}}(r)}\right)$ and we expect this inequality to be strict in general, even though tangent spaces to determinantal varieties are in one-to-one correspondence with the underlying row/column spaces. To see why this is the case, consider a column-space and row-space pair $(C, \mathcal{R}) \subset \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$, with $\dim(C) = \dim(\mathcal{R}) = r$. The collection of matrices $\mathcal{M}_C \subseteq \mathbb{R}^{p_1 \times p_2}$ with column-space contained in C has dimension $p_2 r$ and the collection of matrices $\mathcal{M}_{\mathcal{R}} \subseteq \mathbb{R}^{p_1 \times p_2}$ with row-space contained in $\mathcal{R}$ has dimension $p_1 r$. However, the tangent space $T(C, \mathcal{R}) \subset \mathbb{R}^{p_1 \times p_2}$, which is the sum of $\mathcal{M}_C$ and $\mathcal{M}_{\mathcal{R}}$ has dimension $p_1 r + p_2 r - r^2$. In other words, the spaces $\mathcal{M}_C, \mathcal{M}_{\mathcal{R}}$ do not have a transverse intersection (i.e. $\mathcal{M}_C \cap \mathcal{M}_{\mathcal{R}} \neq \{0\}$), and therefore optimal tangent-space estimation does not appear to be decoupled into (separate) optimal column-space estimation and optimal row-space estimation. Although this heuristic is only an approximation, it does yield good performance in practice, as described in the illustrations in the next subsection as well as in the experiments with real data in the Section 3.4. Further, our final estimate of a tangent space still involves the solution of (3.18) using the approximation (3.20) instead of (3.17). Consequently, we continue to retain our guarantees from Section 3.3 on false discovery control. The full procedure is presented in Algorithm 1.

The tuning parameter $\alpha \in [0, 1]$ in Algorithm 1 plays an important role in how much signal is selected by subspace stability selection. In our experience, the output of subspace stability selection is rather robust to α in moderate to high SNR settings. As a result, in all our experiments, we select α to equal 0.70. For a detailed analysis on the sensitivity to α, please refer to supplementary material Section 8.

Algorithm 1 Subspace Stability Selection Algorithm

1: **I**nput: A set of observations $\mathcal{D}$; a collection of subsamples $\{\mathcal{D}_i\}_{i=1}^B \subset \mathcal{D}$; a row/column space (equivalently, tangent space) estimation procedure $(\hat{\mathcal{C}}, \hat{\mathcal{R}})$; a parameter $\alpha \in (0, 1)$.

2: **O**btain Tangent Space Estimates: For each bag $\{\mathcal{D}_i, i = 1, 2, \ldots, B\}$, obtain row/column space estimates $\{(\hat{\mathcal{C}}(\mathcal{D}_i), \hat{\mathcal{R}}(\mathcal{D}_i))\}_{i=1}^B$ and set $\hat{T}(\mathcal{D}_i) = T(\hat{\mathcal{C}}(\mathcal{D}_i), \hat{\mathcal{R}}(\mathcal{D}_i))$.

3: **C**ompute Average Projection Operators: Compute the average tangent space projection operator $\mathcal{P}_{\mathsf{avg}}$ according to (3.10) and the average row/column space projection operators $\mathcal{P}^{\mathcal{R}}_{\mathsf{avg}}, \mathcal{P}^{\mathcal{C}}_{\mathsf{avg}}$ according to (3.19).

4: **C**ompute Optimal Row/Column Space Approximations: Compute ordered singular vectors $\{u_1, u_2, \ldots, u_{p_1}\} \subset \mathbb{R}^{p_1}$ and $\{v_1, v_2, \ldots, v_{p_2}\} \subset \mathbb{R}^{p_2}$ of $\mathcal{P}^{\mathcal{C}}_{\mathsf{avg}}$ and $\mathcal{P}^{\mathcal{R}}_{\mathsf{avg}}$, respectively. For each $r = 1, \ldots, \min\{p_1, p_2\}$, set $C^\star(r) = \mathrm{span}(u_1, \ldots, u_r)$ and $\mathcal{R}^\star(r) = \mathrm{span}(v_1, \ldots, v_r)$.

5: **T**angent Space Selection via (3.18): Let r_{S3} denote the largest r such that $T(C^\star(r), \mathcal{R}^\star(r)) \in \mathcal{T}_\alpha$.

6: **O**utput: Tangent space $T_{\mathsf{S3}} = T(C^\star(r_{\mathsf{S3}}), \mathcal{R}^\star(r_{\mathsf{S3}}))$.

Computational Cost of Algorithm 1 — We do not account for the cost of obtaining the row/column space estimates $\{(\hat{\mathcal{C}}(\mathcal{D}_i), \hat{\mathcal{R}}(\mathcal{D}_i))\}_{i=1}^B$ on each subsample in Step 2, and focus exclusively on the cost of combining these estimates via Steps 3 – 5. In Step 3, the computational complexity of computing the average projection maps $\mathcal{P}^{\mathcal{R}}_{\mathsf{avg}}, \mathcal{P}^{\mathcal{C}}_{\mathsf{avg}}$ requires $O(B \max\{p_1, p_2\}^2)$ operations and computing the average tangent space projection map $\mathcal{P}_{\mathsf{avg}}$ requires $O(Bp_1^2p_2^2)$ operations. Step 4 entails the computation of two singular value decompositions of matrices of size $p_1 \times p_1$ and $p_2 \times p_2$, which leads to a cost of $O(\max\{p_1, p_2\}^3)$ operations. Finally, in Step 5, to check membership in $\mathcal{T}_\alpha$ we multiply three maps of size $p_1p_2 \times p_1p_2$ and compute the singular value decomposition of the result, which requires a total of $O(p_1^3p_2^3)$ operations. Thus, the computational cost of Algorithm 1 to aggregate estimates produced by B bags is $O(\max\{Bp_1^2, Bp_2^2, Bp_1^2p_2^2, p_1^3, p_2^3, p_1^3p_2^3\})$.

Although the scaling of Algorithm 1 is polynomial in the size of the inputs, when either p_1 or p_2 is large the overall cost due to terms such as $p_1^3p_2^3$ may be prohibitive. In particular, the reason for the expensive terms $Bp_1^2p_2^2$ and $p_1^3p_2^3$ in the final expression is due to computations involving projection maps onto tangent spaces (which belong to $\mathbb{R}^{p_1p_2}$). We describe next a modification of Algorithm 1 so that the resulting procedure only consists of computations involving projection maps onto row and column spaces (which belong to $\mathbb{R}^{p_2}$ and $\mathbb{R}^{p_1}$ respectively).

Modification of Algorithm 1 and Associated Cost — The inputs to this modified procedure are the same as those of the original procedure. We modify Step 3 of Algorithm 1 by only computing the average row/column space projection maps $\mathcal{P}^{\mathcal{R}}_{\texttt{avg}}, \mathcal{P}^{\mathcal{C}}_{\texttt{avg}}$. Let $\mathcal{P}^{\mathcal{C}}_{\texttt{avg}} = U\Gamma U'$ and let $\mathcal{P}^{\mathcal{R}}_{\texttt{avg}} = V\Delta V'$ be the singular value decomposition computations of Step 4. We modify Step 5 of Algorithm 1 to choose the largest $r'_{\texttt{S3}}$ so that $\Gamma_{r'_{\texttt{S3}},r'_{\texttt{S3}}} \geq \alpha$ and $\Gamma_{r'_{\texttt{S3}},r'_{\texttt{S3}}} \geq \alpha$. One can check that the cost associated to this modified procedure is $O(\max\{Bp_1^2, Bp_2^2, p_1^3, p_2^3\})$.

This modified method has the property that the row and column spaces are individually well-aligned with the corresponding averages from the subsamples; the following result shows that the resulting tangent space belongs to a set of stable tangent spaces:

Proposition 3 (Modified Algorithm 1 Satisfies Subspace Stability Selection Criterion)**.** *Let $T_{\texttt{S3-modified}}$ be the output of the modified Algorithm 1 with input parameter α. Then, $T_{\texttt{S3-modified}} \in \mathcal{T}_{1-4(1-\alpha)}$.*

Proposition 3 guarantees that our modification of Algorithm 1 continues to provide false discovery control. We use this modified approach in some of our larger experiments in Section 3.4. The proof of this proposition can be found in supplementary material Section 7.

Finally we remark that in subspace estimation problems (see Section 2.1), the subspace stability selection can be readily employed to find a stable tangent space. In particular, recall from Section 3.3 that the stability selection criterion (3.11) reduces to finding C such that $\sigma_{\texttt{min}}\left(\mathcal{P}_C \mathcal{P}^{\mathcal{C}}_{\texttt{avg}} \mathcal{P}_C\right) \geq \alpha$. Naturally, a projection operator $\mathcal{P}_C$ that satisfies the criterion above can be obtained via singular-value thresholding. Furthermore, this subspace estimate is optimal according to (3.18).

Further Illustrations

In the remainder of this section, we explore various facets of Algorithm 1 via illustrations on the synthetic matrix completion problem setup described at the beginning of Section 3.3. For further demonstrations of the utility of subspace stability selection with real data, we refer the reader to the experiments of Section 4.

Illustration : α vs. $r_{\texttt{S3}}$ — The threshold parameter α determines the eventual optimal rank $r_{\texttt{S3}}$, with larger values of α yielding a smaller $r_{\texttt{S3}}$. To better understand this relationship, we plot in Figure 3.2 $\sigma_{\texttt{min}}(\mathcal{P}_{T_{\texttt{S3}}} \mathcal{P}_{\texttt{avg}} \mathcal{P}_{T_{\texttt{S3}}})$ as a function of $r_{\texttt{S3}}$ for a

large range of values of the regularization parameter λ and SNR = $\{0.4, 0.8, 1.2, 50\}$. Each curve in the different plots corresponds to a particular value of $r_{\mathtt{S3}}$, with the solid curves representing $r_{\mathtt{S3}} = 1, \ldots, 10$ and the dotted curves representing $r_{\mathtt{S3}} = 11, \ldots, 70$. As smaller values of $r_{\mathtt{S3}}$ lead to larger values of $\sigma_{\mathtt{min}}(\mathcal{P}_{T_{\mathtt{S3}}}\mathcal{P}_{\mathtt{avg}}\mathcal{P}_{T_{\mathtt{S3}}})$, the curves are ordered such that the top curve corresponds to $r_{\mathtt{S3}} = 1$ and the bottom curve corresponds to $r_{\mathtt{S3}} = 70$. We first observe that for a fixed $r_{\mathtt{S3}}$, the associated curve is generally decreasing as a function of λ. For large values of λ, both signal and noise are substantially reduced due to a significant amount of regularization. Conversely, for small values of λ, both signal and noise are present to a greater degree in the estimates on each subsample; however, the averaging procedure reduces the effect of noise, which results in high-quality aggregated estimates for smaller values of λ. Next, we observe that the curves indexed by $r_{\mathtt{S3}}$ cluster in the high SNR regime, with the first three corresponding to $r_{\mathtt{S3}} = 1, 2, 3$, the next five corresponding to $r_{\mathtt{S3}} = 4, \ldots, 8$, the next two corresponding to $r_{\mathtt{S3}} = 9, 10$, and finally the remaining curves corresponding to $r_{\mathtt{S3}} > 10$. This phenomenon is due to the clustering of the singular values of the underlying population $L^\star$. On the other hand, for low values of SNR, the clustering is less pronounced as the components of $L^\star$ with small singular values are overwhelmed by noise.

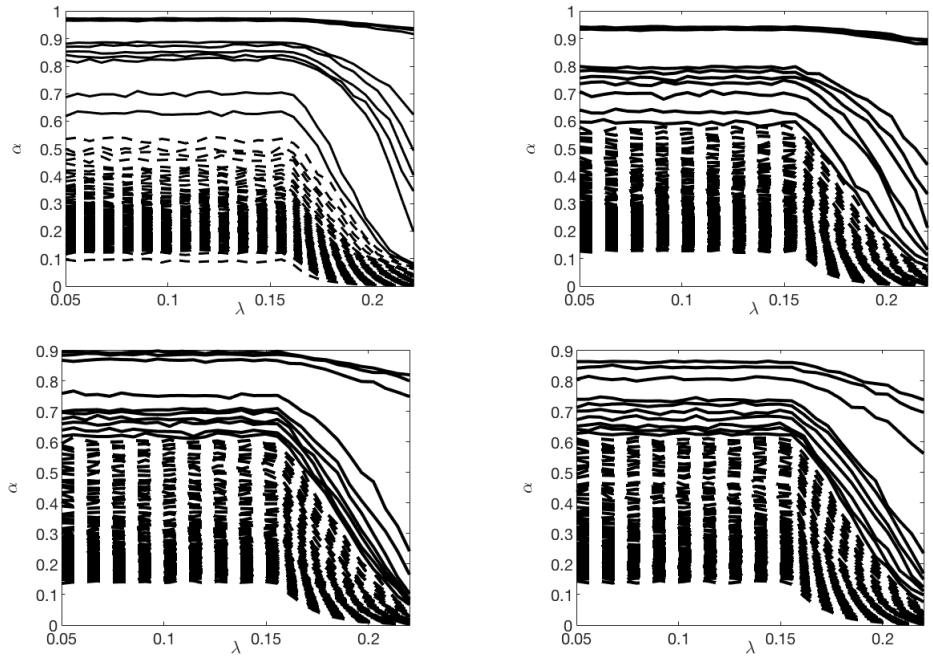


Figure 3.2: Relationship between $r_{\mathtt{s3}}$ and α in Algorithm 1 for a large range of λ and SNR = $\{0.4, 0.8, 1.2, 50\}$.

Illustration: subspace stability selection reduces false discovery — Next, we demonstrate that subspace stability selection produces a tangent space which is different and usually of a higher quality (e.g. smaller expected false discovery) than the base estimator applied to the full dataset. We choose the noise level so that SNR takes on one of the values in $\{1.5, 2, 2.5, 3\}$. On the one hand, we employ the procedure (3.9) on a subset of 2231 observations (the training set) of the full set of 3186 observations and the remaining subset of 955 observations constitute the test set. We use cross-validation to identify an optimal choice $\lambda^\star$ of the regularization parameter. The estimate produced by (3.9) on the training set for this choice of $\lambda^\star$ is recorded as the output of the non-subsampled approach. On the other hand, the estimator (3.9) with the choice $\lambda^\star$ is used in conjunction with $\alpha = 0.7$ to produce a subspace stability selection tangent space via Algorithm 1. For each of the four choices of SNR, we run 100 experiments and average to find an empirical approximation to the expected false discovery (3.3). Table 3.1 compares the expected false discovery (with one sigma statistics) of the non-subsampled approach to that of the subspace stability selection procedure for the different problem settings. Evidently, subspace stability selection yields a much smaller amount of false discovery compared to not employing subsampling.

Method	No Subsampling	Subspace stability selection
SNR = 1.5	1274.6 ± 78.8	107.6 ± 11.5
SNR = 2	1532.8 ± 68.5	89.7 ± 16.9
SNR = 2.5	1573.5± 71.2	87.9 ± 18.7
SNR = 3	1417 ± 63.5	87.9 ± 19.4

Table 3.1: False discovery of subspace stability selection vs a non-subsampled approach on the stylized matrix completion problem. The maximum possible amount of false discovery is $\dim(T^{\star\perp}) = (70 - 10)^2 = 3600$.

At this stage, it is natural to wonder whether the source of the improved false discovery control provided by subspace stability selection over not using subsampling is simply due to the non-subsampled approach providing estimates with a larger rank? In particular, as an extreme hypothetical example, the zero-dimensional space is a stable tangent space and has zero expected false discovery, and more generally lower-rank tangent-space estimates are likely to have smaller expected false discovery. Thus, is subsampling better primarily because it produces lower-rank

estimates? To address this point in our stylized setup, we consider a population $L^\star$ with associated incoherence parameter equal to 0.8 [1]. We sweep over the regularization parameter λ, and we compare the following two estimates: first, the estimate $\hat{L}$ obtained via (3.9) and then truncated to its first three singular values, and subsampled estimates obtained via Algorithm 1 with r_{S3} set to three. The choice of three here is motivated by the fact that the population low-rank matrix $L^\star$ has three large components. We perform this comparison for SNR = $\{0.8, 1.6\}$ and describe the results in the plots in Figure 3.3. In the high SNR regime, the performances of the subsampled and the non-subsampled approaches are similar. However, in the low SNR regime, subspace stability selection yields a tangent space with far less false discovery across the entire range of regularization parameters. Further, subspace stability selection provides a fundamentally different solution that cannot be reproduced simply by selecting the "right" regularization penalty in (3.9) applied to the entire dataset.

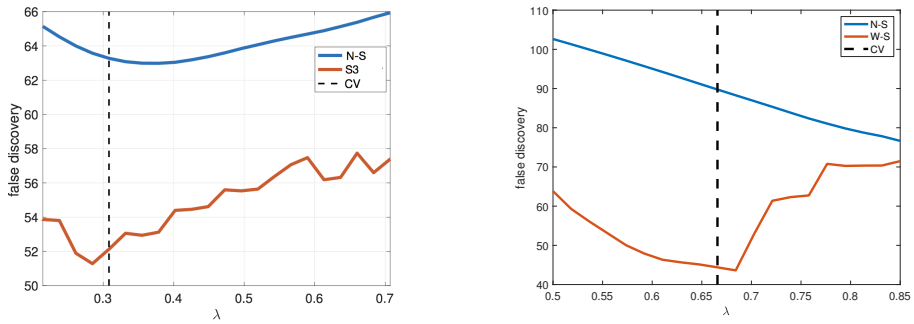


Figure 3.3: False discovery of subspace stability selection vs a non-subsampled approach with SNR = 1.6, 0.8. Here, we choose a rank-3 approximation of the non-subsampled approach and $r_{S3} = 3$ in Algorithm 1 of subspace stability selection. The maximum possible amount of false discovery is $\dim(T^{\star\perp}) = (70 - 10)^2 = 3600$. Furthermore,'N-S' denotes no subsampling and 'S3' denotes subspace stability selection.

Similar behavior is also observed when the solution $\hat{L}$ is truncated at a different rank. As an example, with SNR = 0.8, we choose λ via cross-validation and truncate $\hat{L}$ at rank $r = 1, 2, \ldots, 5$ and compare its false discovery to the estimate produced by subspace stability selection with $r_{S3} = r$ (shown in Table 2).

[1] The incoherence of a matrix M is $\max_i \max\{\|\mathcal{P}_{\text{col-space}(M)}(e_i)\|_2^2, \|\mathcal{P}_{\text{row-space}(M)}(e_i)\|_2^2\}$, where e_i is the i'th standard basis vector, and it plays a prominent role in various analyses of the low-rank matrix completion problem [CR09].

Method	rank = 1	rank = 2	rank = 3	rank = 4	rank = 5
No subsampling	20.4	48.1	89.7	146.7	218.8
Subspace stability selection	12.4	25.6	44.3	70.4	109

Table 3.2: False discovery of subspace stability selection vs a non-subsampled approach with SNR = 0.8 and rank of the estimate set to vary from 1 to 5. The maximum possible amount of false discovery is $\dim(T^{\star\perp}) = 3600$.

Illustration: stability of tangent spaces to small changes in regularization parameter— Finally, we note that in settings in which regularization is employed, the estimate can be extremely sensitive to the choice of regularization parameter. For example, in nuclear-norm regularized formulations such as (3.9), small changes to the parameter λ can often lead to substantial changes in the optimal solution. A virtue of subspace stability selection is that the estimates that it provides are generally very stable to small perturbations of λ. To formalize this discussion, given two tangent spaces T and $\tilde{T}$, we consider the quantity

$$\mu(T, \tilde{T}) \triangleq 1 - \frac{\text{trace}\,(\mathcal{P}_T \mathcal{P}_{\tilde{T}})}{\max\{\dim(T), \dim(\tilde{T})\}},$$

which measures the degree to which T and $\tilde{T}$ are misaligned. If $T = \tilde{T}$, then $\mu(T, \tilde{T}) = 0$, and on the other hand, $T \subseteq \tilde{T}^{\perp}$ would yield $\mu(T, \tilde{T}) = 1$. Hence, larger values of $\mu(T, \tilde{T})$ are indicative of greater deviations between T and $\tilde{T}$. We use this metric to compare the stability of the non-subsampled approach with subspace stability selection. In our stylized setup, we choose the noise level so that $SNR = 4$ and we select $\lambda = 0.03$ (based on cross-validation). Letting T be the tangent space of the estimator (3.9) with $\lambda = 0.03$ and $\tilde{T}$ with $\lambda = 0.05$, we find that $\mu(T, \tilde{T}) = 0.23$. Setting $\alpha = 0.7$ with $B = 100$ complementary bags and computing the same metrics for the outputs of subspace stability selection, we find that $\mu(T, \tilde{T}) = 0.003$. This contrast is observed for many other SNR levels.

3.4 Experiments

In this section, we demonstrate the utility of subspace stability selection in providing false discovery control both with synthetic and real data. We consider the following types of low-rank estimation problems:

1. Low-rank linear measurements and matrix completion: We consider noisy linear functions of a low rank matrix $L^{\star} \in \mathbb{R}^{p_1 \times p_2}$ of the form $Y_i \approx \langle \mathcal{A}_i, L^{\star} \rangle$, $i = 1, \ldots, n$ where each $\mathcal{A}_i \in \mathbb{R}^{p_1 \times p_2}$. In the linear measurement setting, $\mathcal{A}_i$ is a general matrix, and in the matrix completion setting, $\mathcal{A}_i$ will be zeros

everywhere except a single entry which is equal to 1. The matrix completion problem is similar to the one considered in the stylized demonstrations of Section 3.3. One point of departure from that discussion in the present section is that in experiments where the dimensions p_1, p_2 are large, employing the nuclear norm regularized estimator (3.9) on each subsample is impractical. Instead, we use on each subsample the following non-convex formulation:

$$(\hat{U}\,,\ \hat{V}) = \underset{U\in\mathbb{R}^{p_1\times k}, V\in\mathbb{R}^{p_2\times k}}{\operatorname{argmin}} \quad \sum_{i\in S}(Y_i - \langle \mathcal{A}_i, UV'\rangle)^2 + \lambda\,(\|U\|_F^2 + \|V\|_F^2). \tag{3.21}$$

where $\|U\|_F^2 + \|V\|_F^2$ is a surrogate for the nuclear norm penalty in (3.9), $\lambda > 0$ is a regularization parameter, and $S \subset \{1,\dots,p_1\}\times\{1,\dots,p_2\}$ is the set of observed indices. By construction, $\hat{L} = \hat{U}\hat{V}'$ is constrained to have rank at most k, and this rank can be adjusted by appropriately tuning λ. Fixing U (resp. V) the above problem is convex in V (resp. U), and thus a commonly employed approach in practice is alternating least-squares (ALS) [SJ05] .

2. Factor analysis: We observe samples $\{Y^{(i)}\}_{i=1}^n \subset \mathbb{R}^p$ of a random vector and we identify a factor model that best explains these observations, i.e., a model in which the coordinates of the observed vector are independent conditioned on a small number $k \ll p$ of latent variables. In other words, our objective is to approximate the sample covariance of $\{Y^{(i)}\}_{i=1}^n$ by a covariance matrix that is decomposable as the sum of a diagonal matrix and a low-rank matrix. Using the Woodbury Inversion Lemma, we have that the precision matrix can be decomposed as a diagonal matrix minus a low-rank matrix. The virtue of working with precision matrices is that the the log-likelihood function is concave with respect to this parametrization. On each subsample, we use the following estimator [Sha82c]:

$$\begin{aligned}(\hat{D},\hat{L}) = \underset{L\in\mathbb{S}^p, D\in\mathbb{S}^p}{\operatorname{argmin}} \quad & -\log\det(D-L) + \operatorname{trace}\left(\left(\frac{1}{|S|}\sum_{i\in S} Y^{(i)}Y^{(i)\prime}\right)(D-L)\right) \\ + \quad & \lambda\,\operatorname{trace}(L). \\ \text{subject to} \quad & D - L \succ 0,\ L \succeq 0,\ D \text{ is diagonal.}\end{aligned} \tag{3.22}$$

Here trace$(\cdot)$ is the restriction of the nuclear norm to symmetric positive-semidefinite matrices.

Synthetic Simulations

We explore the role of commutator in the false discovery bound of Theorem 4 in a stylized matrix denoising setting. Specifically, we generate a population low-rank matrix $L^\star \in \mathbb{R}^{p\times p}$ with $p = 200$, the rank of $L^\star$ is set to 6, the nonzero singular values are set to $\{120, 100, 80, 30, 20, 10\}$, and the row and column spaces sampled uniformly from the Steifel manifold. Letting $U^\star Q V^{\star\prime}$ be the full SVD of $L^\star$, we obtain n noisy measurements of $L^\star$ of the form $Y_i = L^\star + \delta[\gamma U^\star D_i V^{\star\prime} + \epsilon_i]$ for $i = 1, 2, \ldots, n$, where D_i is a diagonal matrix with entries chosen iid from a normal distribution and $\epsilon_i \in \mathbb{R}^{p\times p}$ is a normal Gaussian matrix with iid entries. The parameter $\delta > 0$ controls the signal-to-noise ratio and the parameter $\gamma > 0$ controls the commutator term inside Theorem 4 . In particular, large γ leads to smaller commutator term since all the measurements Y_i and $L^\star$ are nearly simultaneously diagonalizable. Geometrically, this corresponds to the principals angles between $T^{\star\perp}$ and $\hat{T}(\mathcal{D}(n/2))$ concentrating around 0 and $\pi/2$.

We vary γ in the range $[30, 20]$ and for each γ, we chose δ so that SNR = 0.15. Here, SNR $= \mathbb{E}\left[L^\star\|_2/\|\lambda[\gamma U^\star D_i V^\star + \epsilon_i]\|_2\right]$. We obtain $n = 2p$ measurements as input to supply to the estimator that performs hard thresholding on the average $\bar{Y} = \sum_{i=1}^n Y_i$. Specifically, letting $U_Y D_Y V_Y'$ be the SVD of $\bar{Y}$, the estimate for a choice of $\lambda > 0$ is given to be $\hat{L} = U_Y (D_Y)_\lambda V_Y'$ where $(D_Y)_\lambda$ sets the diagonal entries of D_y that are less than λ to zero and leaves the rest unchanged. Since the population rank is 6, we select λ to obtain a rank 6 output. We apply subspace stability selection with $\alpha \in [0.9, 0.98]$ and $B = 50$ complementary bags, and we obtain an empirical approximation of the expected false discovery over 100 trials. Since the population model is known, the quantities inside Theorem 4 are readily obtainable. We note that as we have knowledge of the noise structure of the population model, we set the orthonormal basis elements $\{M_i\}_{i=1}^{\dim(T^{\star\perp})}$ needed to compute the term F in (3.12) to be $\{U^\star_{:,r+i} V^{\star\prime}_{:,r+j}\}_{i,j=1}^{p-r}$. Figure 3.4(a,b) compares the achieved false discovery of subspace stability selection with Theorem 4, average number of discoveries of subspace stability selection (e.g. $\mathbb{E}[\dim(T)]$, and simply using the entire data once without any subsampling. The results of Figure 3.4 suggest that in settings where the commutator terms are not too large, our theorem bound is valid, non-vacuous and effective: they produce a smaller false discovery than the average number of discoveries of subspace stability selection and it gives a smaller value than the estimator that uses the full data with no subsampling. Notice that subspace stability selection is substantially better than the full data approach or the theorem bound, as $L^\star$ has three strong components, and subspace stability selection teases those

away from the noise and yields a small false discovery. We believe that bridging the gap between the achieved false discovery bound of subspace stability selection and Theorem 4 requires more careful bag dependent analysis, which is an interesting avenue for future research. Figure 3.4(c,d) demonstrates the Theorem 4 utility when λ is conservatively chosen so that a rank-10 estimate is selected.

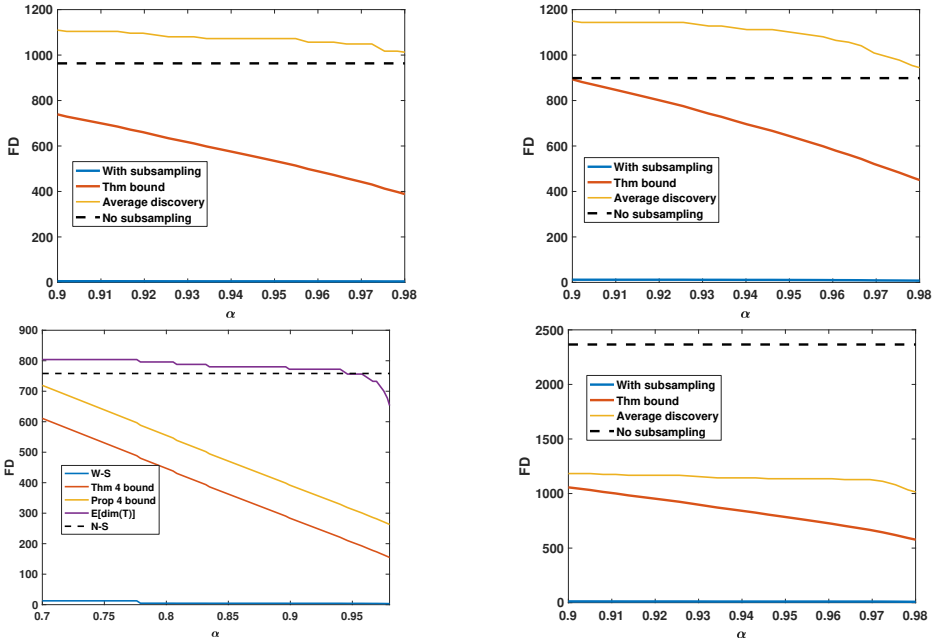


Figure 3.4: top left: $\gamma = 30$; rank sel. $= 6$ and $\|[\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}, \mathcal{P}_{T^{\star\perp}}]\|_F \approx 41$;top right: $\gamma = 20$; rank sel. $= 6$ and $\|[\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}, \mathcal{P}_{T^{\star\perp}}]\|_F \approx 91$; bottom left: $\gamma = 30$; rank sel. $= 10$ and $\|[\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}, \mathcal{P}_{T^{\star\perp}}]\|_F \approx 69$; bottom right: $\gamma = 30$; rank sel. $= 10$ and $\|[\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}, \mathcal{P}_{T^{\star\perp}}]\|_F \approx 114$. False discovery of subspace stability selection as a function of α for matrix denoising setting. The blue curve is false discovery obtained by subspace stability selection; the red curve is Theorem 4 bound; the yellow curve is average dimension of the selected tangent space; and the dotted line is false discovery from using entire data. Subspace stability selection has small but nonzero false discoveries. As an example, for $\gamma = 20$, rank selected $= 6$, and $\alpha = 0.9$, subspace stability selection chooses on average a rank-3 model with 11.7 false discoveries. Here $\dim(T^{\star\perp}) = 37636$.

Next, we explore the false discovery and power attributes of subspace stability selection in different noise and rank regimes. We consider the linear Gaussian measurement setting described earlier with $p = 60$, rank of $L^\star$ in the set $\{1, 2, 3, 4\}$, the nonzero singular values set to 1, and the row and column spaces sampled uniformly from the Steifel manifold. These observation noise level is tuned so that SNR lies in the set $\{1, 2, 3, 4, 5\}$. A fraction $6p^2/10$ are used as training data for the estimator (3.21) with λ chosen via holdout validation with a validation set of size

$3p^2/20$ and the rank constraint k set to 10. With this choice of λ, we evaluate the expectation and standard deviations of false discovery and the power empirically over 100 trials. As a point of comparison, we set $\alpha = 0.7$ with $B = 100$ complementary bags and compute the same metrics based on subspace stability selection. Figure 3.5(a) demonstrates the performance of the non-subsampled approach and subspace stability selection for all the problem settings. For settings where either the false discovery standard deviation normalized by expected value or the power standard deviation normalized by expected value is greater than 0.01, we plot the expected value with a cross and the one sigma around the mean with a rectangle. Evidently, for most problem instances, subspace stability selection yields a solution with a significantly smaller amount of false discovery without much loss in power.

We repeat a similar experiment in the matrix completion setting where $L^\star \in \mathbb{R}^{p\times p}$ with $p = 100$, rank in the set $\{1, 2, 3, 4\}$, row and column spaces chosen uniformly from Steifel manifold. We select a fraction $7/10$ of the total entries are chosen uniformly at random as the observation set Ω so that $|\Omega| = 7p^2/10$. These observations are corrupted with Gaussian noise with variance selected so that SNR is in the range $\{0.5, 0.875, 1.25, 1.625, 2.00\}$. We use these observations as input to the estimator (3.21), with λ selected based on holdout validation on a $n_{\text{test}} = 7/20p^2$ validation set. Figure 3.5(b) compares the performance of the non-subsampled approach and subspace stability selection computed empirically over 100 iterations. Several settings in Figure 3.5 experience a significant loss in power using the subspace stability selection procedure. Those precisely correspond to models with high rank and low SNR regime where some components of the signal are overwhelmed by noise. To control false discoveries in these settings, subspace stability selection filters out filters out some of the signal and as a result yields a small power.

Experimental Results on Real Datasets

Collaborative filtering

In collaborative filtering, one is presented with partially filled user-preference matrices in which rows are indexed by users and columns by items, with each entry specifying a user's preference for an item. The objective is to infer the unobserved entries. As discussed in Section 3.1, such user-preference matrices are often well-approximated as low-rank, and therefore a popular approach to collaborative filtering is to frame it as a problem of low-rank matrix completion, and solve this problem based either on the convex relaxation (3.9) or the non-convex approach

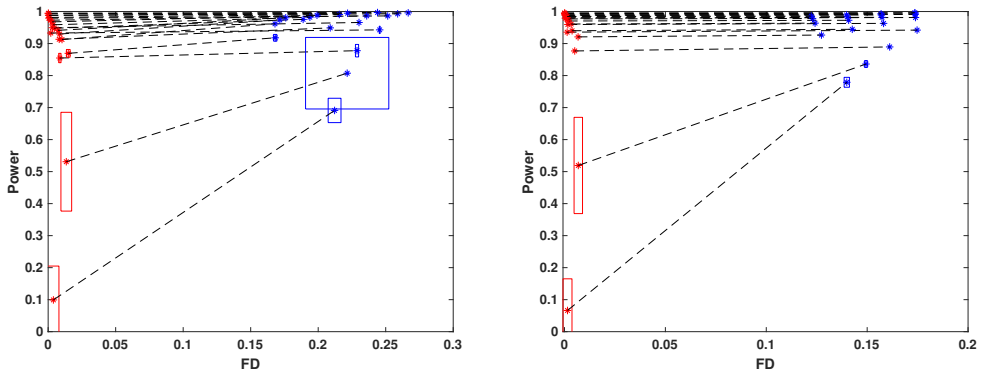


Figure 3.5: False discovery vs power with (a) matrix completion and (b) linear measurements over 20 different problem instances (varying rank and noise level). Blue crosses corresponds to the performance of the non-subsampled approach and red crosses correspond to subspace stability selection with $\alpha = 0.7$. For the instances where standard deviation divided by mean is greater than 0.01, we show one sigma rectangle around the mean. The lines connect dots corresponding to the same problem instance. Both the false discovery and the power are normalized by dividing the expressions (3.3) and (3.4) by $\dim(T^{\star\perp})$ and $\dim(T^{\star})$, respectively.

(3.21) via ALS. We describe experimental results on two popular datasets in collaborative filtering: 1) the Amazon Book-Crossing dataset (obtained from `http://www2.informatik.uni-freiburg.de/~cziegler/BX/`) of which we consider a portion consisting of $p_1 = 1245$ users and $p_2 = 1054$ items with approximately 6% of the ratings (integer values from 1 to 10) observed, and 2) the Amazon Video Games dataset (obtained from `http://jmcauley.ucsd.edu/data/amazon/`) of which we consider a portion consisting of $p_1 = 482$ users and $p_2 = 520$ items with approximately 3.5% of the ratings (integer values from 1 to 5) observed. In each case, we partition the dataset as follows: we set aside 85% of the observations as a training set, 10% of the observations as a holdout validation set, and the remaining 5% as an evaluation set to assess the performance of our learned models.

As these problems are relatively large in size, we employ ALS on the non-convex formulation (3.21) with $k = 80$ (the upper bound on the rank) and we apply the modification of Algorithm 1 for subspace stability selection. Finally, to obtain estimates of low-rank matrices (as this is the eventual object of interest in collaborative filtering) we use the formulation (3.7) given estimates of tangent spaces. We set $\alpha = 0.7$ and $B = 100$ complementary bags. Figure 3.6 illustrates the mean squared error of ALS and subspace stability selection on the holdout set for these two datasets for a range of values of the regularization parameter λ. For both datasets, we observe that subspace stability selection yields models with better

MSE on the holdout set over the entire range of regularization parameters. On the Book-Crossings dataset, we further note that at the cross-validated λ, the rank of the estimate obtained from the non-subsampled approach is 80 (i.e., the maximum allowable rank) with the first three singular values equal to $4329, 135.4, 63.1$. The MSE of this model on the evaluation set is equal to 0.83. On the other hand, at the cross-validated λ subspace stability selection yields a rank-2 model with an MSE of 0.81 on the evaluation set. Thus, we obtain a much simpler model with subspace stability selection that also offers better predictive performance. Similarly, for the Amazon Video Games dataset, the rank of the estimate obtained from the non-subsampled approach is 39 with the first five singular values equal to $1913.5, 49.4, 43.6, 28.4, 27.4$, with an MSE of 0.87 on the evaluation set. On the other hand, subspace stability selection yields a rank-4 solution with a much smaller MSE of 0.74 on the evaluation set. Finally, we observe for both datasets that subspace stability selection is much more stable across the range of regularization parameters. Thus, subspace stability selection is far less sensitive to the particular choice of λ, which removes the need for fine-tuning λ.

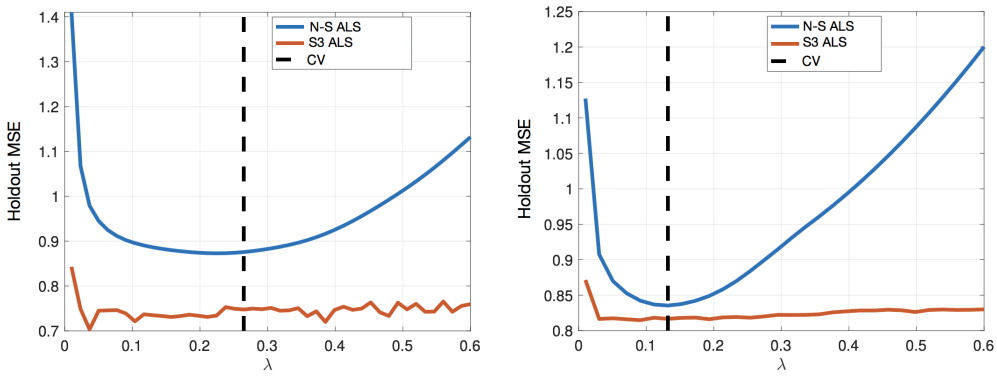


Figure 3.6: Collaborative filtering: MSE on holdout set of non-subsampled approach (denoted 'N-S' and colored in blue) and subspace stability selection (denoted 'S3' and colored in red). Dotted black line represents the cross-validated choice of λ with the non-subsampled approach.

Hyperspectral unmixing

Here we give an illustration with real hyperspectral imaging data in which the underlying population parameters are known based on extensive prior experiments. In this problem, we are given a hyperspectral image $Y \in \mathbb{R}^{p_1 \times p_2}$ consisting of p_1 frequency bands and p_2 pixels, where $Y_{i,j}$ is the reflectance of the

j'th image pixel to the i'th frequency band. The spectral unmixing problem aims to find $W \in \mathbb{R}^{p_1 \times k}$ (called the endmember matrix) and $H \in \mathbb{R}^{k \times p_2}$ (called the abundance matrix) so that $Y \approx WH$, where $k \ll \min(p_1, p_2)$ is the number of endmembers [Man03]. Of particular interest is the k-dimensional column-space of W, which corresponds to the space spanned by the k endmembers that are present in the image. We discuss two natural hyperspectral unmixing problems that arise commonly in practice. We focus on the Urban dataset (obtained from `http://www.escience.cn/people/feiyunZHU/Dataset_GT.html`), a hyperspectral image consisting of 307×307 pixels, each of which corresponds to a $2 \times 2m^2$ area with 210 wavelengths ranging from $400nm$ to $2500nm$. Following previous analyses of this dataset, we remove 48 noisy channels to obtain 162 wavelengths and select a 30×25 patch (equal to 750 pixels) shown in Figure 3.7(a). In the selected patch, there are a total of 3 endmembers (shown in Figure 3.7(b)), with one strong signal and two weak signals.

In many settings, obtaining a complete hyperspectral image of a scene may be

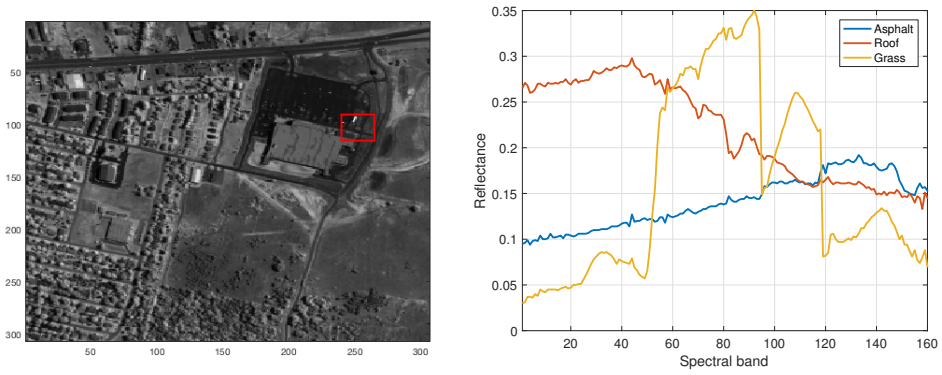


Figure 3.7: Urban hyperspectral image (left) and spectra of three materials present in the image (right). The data and the population spectra are obtained from `http://www.escience.cn/people/feiyunZHU/Dataset_GT.html`.

costly, and it is of interest to accurately reconstruct a hyperspectral image from partial observations. This problem may be naturally formulated as one of low-rank matrix completion. As with other application domains in which problems are reformulated as low-rank matrix completion, ALS applied to the non-convex formulation (3.21) is especially popular in hyperspectral unmixing. To simulate such a hyperspectral unmixing problem, we randomly subsample 10% of the hyperspectral data in the patch as training data. We further select another 10% of the remaining data as a holdout validation set. We compare the amount of false discovery of a non-subsampled approach and subspace stability approach, with k

conservatively chosen to be equal to 20 in the ALS procedure in each case. Due to the scale of this problem being large, we use the modification of Algorithm 1 (with $\alpha = 0.7$ and $B = 100$ complementary bags) described in Section 3.3 for subspace stability selection. As the column space of the low-rank estimate is the principal object of interest for endmember detection, the quantities of interest for evaluating performance are based on (3.5): $\overline{\text{FD}} = \mathbb{E}\left[\text{trace}\left(\mathcal{P}_{\text{col-space}(W^\star)^\perp}\mathcal{P}_{\text{col-space}(\hat{W})}\right)\right]$ and $\overline{\text{PW}} = \mathbb{E}\left[\text{trace}\left(\mathcal{P}_{\text{col-space}(W^\star)}\mathcal{P}_{\text{col-space}(\hat{W})}\right)\right]$. Here, the expectation is with respect to the randomness in the selection of the 10% training data, $W^\star \in \mathbb{R}^{162\times 3}$ is the matrix consisting of the spectra of the three endmemebers in Figure 3.7(b)), and $\hat{W}$ is the estimated matrix. We find a cross-validated choice of $\lambda = 1$ from one random selection of training data. With this λ and over 100 random trials in the selection of training data, no subsampling ALS produces on average rank-20 estimate with $\overline{\text{FD}} = 0.1$ dim(col-space($W^{\star\perp}$)) and $\overline{\text{PW}} = 0.97$ dim(col-space($W^\star$)). In contrast, for the same $\lambda = 1$, subspace stability selection (operating on tangent spaces T_n(col-space($\hat{W}$)) produces on average rank-2.86 with $\overline{\text{FD}} = 0.0007$ dim(col-space($W^{\star\perp}$)) and $\overline{\text{PW}} = 0.91$ dim(col-space($W^\star$)). Furthermore, even if λ is set large enough (for example, $\lambda = 29$) so that the non-subsampled ALS estimate has on average rank equal to 2.52, the false discovery estimate is $\overline{\text{FD}} = 0.007$ dim(col-space($W^\star$)$^\perp$), which is still far larger than the amount of false discovery of subspace stability selection.

A different type of hyperspectral unmixing problem arises if the observations are corrupted by noise. In particular, based on the decomposition $Y \approx WH$, the outer product YY' is well approximated by a low-rank matrix. Thus, another natural approach for endmember detection is to perform factor analysis by viewing each column of Y (i.e., an entire collection of wavelengths corresponding to each pixel) as an observation and approximating the sample covariance of these observations as the sum of diagonal and low-rank matrices. The row/column spaces of the low-rank component (which is symmetric, hence the row and column spaces are the same) serve as estimates of the subspace spanned by the endmembers. We obtain $\{Y^{(i)}\}_{i=1}^{750} \subset \mathbb{R}^{162}$ spectral observations of the 750 total pixels by applying white noise to the population parameters with the noise level chosen so that SNR = 0.78. We then set aside 80% of the data as training data for the estimator (3.22), which is solved using LogDetPPA solver [TTT16]. We set aside the remaining 20% as a holdout validation set. Employing the estimator (3.22) without subsampling and with λ chosen via cross-validation and expectations computed over 100 yields false

discovery FD = 0.04 $\dim(T^{\star\perp})$ and power PW = 0.48 $\dim(T^{\star})$. (Here $T^{\star}$ represents the population tangent space.) On the other hand, subspace stability selection with $\alpha = 0.7$ and $B = 100$ complementary bags yields a tangent space estimate with a false discovery and power FD = 0.015 $\dim(T^{\star\perp})$ and PW = 0.69 $\dim(T^{\star})$, respectively. Evidently, subspace stability selection yields a substantial decrease in the amount of false discovery as well as an improvement in power.

3.5 Conclusions and Future Directions

In this paper, we describe a geometric framework for assessing false discoveries in low-rank estimation. The proposed framework has many appealing properties including that it is a natural generalization of false discovery in variable selection. We further describe the subspace stability selection algorithm to provide false discovery control in the low-rank setting. This procedure is a generalization of the stability selection method of [MB10]. The method is general and we demonstrate its utility with both synthetic and real datasets in a range of low-rank estimation tasks.

There are several interesting directions for further investigation that arise from our work. First, Algorithm 1 from Section 6.1 outputs an estimate that does provide false discovery control, but it is unclear whether this is the most powerful procedure possible. In particular, it is of interest to obtain an optimal solution to the problem (3.17), or to prove that Algorithm 1 computes a near-optimal solution. Next, a significant topic of contemporary interest in variable selection — especially when there are a large number of possible predictors — is to control for the false discovery rate. In Section 3.2 we gave a formulation of false discovery rate in the low-rank setting, and it is natural to seek procedures that provide false discovery rate control in settings with high-dimensional matrices. One obstacle that arises with this effort is that every proof of false discovery rate control of a variable selection method (of which we are aware) relies strongly on the simultaneous diagonalizability of the projection matrices associated with the population tangent space and the estimated tangent space (when translated to the geometric viewpoint of our paper). Finally, the geometric framework developed in this paper for assessing false discovery is potentially relevant beyond the specific setting of low-rank estimation. For example, our setup extends naturally to latent-variable graphical model selection as well as low-rank tensor estimation, both of which are settings in which the underlying geometry is similar to that of low-rank estimation. More broadly, the perspective presented here may be useful in addressing many other structured estimation problems.

Chapter 4

LATENT VARIABLE GRAPHICAL MODELING FOR GENERALIZED LINEAR MODELS

The latent-variable modeling of the reservoir system in Chapter 2 assumed that the reservoir volumes and the latent variables are well-approximated by a Gaussian distribution. In Chapter 2, we validated the Gaussianity assumption to be sensible. However, in many settings, the observed or latent variables deviate strongly from Gaussianity. As an example, with voting records, the data typically consists of binary values, and with gene expressions, the data contains count values. Motivated by these applications, in this chapter, we address the challenge of latent variable graphical modeling beyond Gaussian variables.

The results in this chapter will be submitted in a paper that is in preparation. This work was joint with Parikshit Shah and Venkat Chandrasekaran. The author contributed by helping develop the modeling framework and associated parameter estimation algorithm, as well as producing numerical experiments. The description of the work contained in this chapter was written by the author.

4.1 Introduction

Graphical modeling is a commonly employed technique for identifying dependencies among a collection of variables. The task of finding a graphical model underlying a collection of variables is made difficult by the presence of latent variables.

In Chapter 2, we considered a latent-variable graphical model for the reservoir network to account for the presence of latent variables. This modeling framework and the associated convex estimator rely on the property that the observed and latent variables are jointly Gaussian. In many settings, the data may not be well approximated by Gaussian distributions. As an example, the latent variables underlying the reservoirs may be policies (or regulations) and these variables are better described by categorical variables. As another example, we analyze in Section 4.4 Level III Breast Cancer miRNA expression data and U.S. Senate 108th voting record that strongly deviate from Gaussianity. Specifically, RNA sequence information is often represented by count data that take on positive integer values, and voting record

dataset consist of yes or no (binary variables) to a collection of bills. In all these application settings, there is a need to develop modeling techniques and efficient algorithms to identify a latent variable graphical model underlying non-Gaussian variables.

Our Contributions

The challenge with modeling beyond non-Gaussian variables is that the state of the art methods suffer from at least one of these deficiencies:

1. They are unable to handle non-Gaussianity
2. They rely on non-convex or computationally intractable algorithms
3. They cannot account for the presence of latent variables

We address all three challenges based on *Generalized Linear Models* (GLM's). This class of distributions provide a flexible generalization of ordinary linear regression so that the response variables have a non-Gaussian distribution [MN83]. In particular, GLM's relate a linear model of a collection of covariates to the response variables via a link function. In Section 4.2, we describe an exponential family class where the conditionals (a single observed variable conditioned on the remaining observed variables and latent variables) are distributed as a GLM. Our modeling approach — parameterized by sparse matrix encoding graph structure among observed variables and low-rank matrix encoding the effect of latent variables on observed variables — provides a flexible framework to model Gaussian, Bernoulli, Poison, and Exponential variables. In Section 4.3, we describe a convex algorithm to identify the parameters of this model. A key ingredient in formulating this estimator is approximating the maximum-likelihood estimator by the *pseudo-likelihood*, which is commonly employed in statistical modeling to achieve computationally efficient estimator at the cost of statistical efficiency [Bes75]. Finally, we provide extensive experimental demonstrations showing the utility of our approach with synthetic data as well as real data consisting of 108th voter's US Senate voter records dataset and Level 3 breast cancer miRNA expression data.

Related works:

Our modeling class of conditional Generalized Linear Models is similar to [Yan+15; CWS15]. However, unlike these previous works, our model accounts for the presence of unobserved variables. Furthermore, the algorithm to fit to this

model is inspired by the neighborhood selection approach originally introduced in [MB06] and subsequently employed in various graphical modeling frameworks such as [Yan+15; CWS15], among others. Neighborhood selection solves an optimization problem for each node to find the connected edges. However, since the effect of latent variables are global, our algorithm stitches each of these separate optimization problems together and solves a joint problem. This type of joint neighborhood selection was considered in [HT09] for speedup in the graphical Lasso.

4.2 Modeling Framework

We consider the pairwise conditional graphical model among the collection of p random variables $x \in \mathbb{R}^p$ conditioned on $z \in \mathbb{R}^r$ latent variables, where the nodewise conditional distribution is specified by a GLM. In particular, we posit that the distribution of $x|z$ is from the following class:

$$\mathscr{P} = \left\{ p(x|z) = \exp\left(\left[\alpha^T x + \sum_{s=1}^{p} \Lambda_{s,s} f(x_s) \right] + \frac{1}{2} x^T K x + x^T B z + A(K, B, \alpha, \Lambda) \right) \right\} \quad (4.1)$$

where $\left[\alpha^T x + \sum_{s=1}^{p} \Lambda_{s,s} f(x_s)\right]$ represents the node potential with $\alpha \in \mathbb{R}^{p\times 1}$ and $\Lambda \in \mathbb{S}^p$ a diagonal matrix with positive diagonal entries. The matrix $K \in \mathbb{R}^{p\times p}$ encodes the conditional independency properties of the graph, so that $x_s \perp\!\!\!\perp x_t | z, x_{-s,t} \leftrightarrow K_{s,t} = 0$. The model class (4.1) assumes that the dependency structure K does not depend on the specific configuration of the latent variables. We further assume without loss of generality that $K_{s,s} = 0$ for all i as the component corresponding to the diagonal can be absorbed into $f(x_s)\Lambda_{s,s}$. Furthermore, the matrix $B \in \mathbb{R}^{p\times k}$ encodes the effect of latent variables on observed variables. Finally, the quantity $A(K, B, \alpha, \Lambda)$ is a normalization constant that is a convex function of its parameters. The model class (4.1) without the latent terms has been considered in previous works to identify a graphical model in GLM's [CWS15; Yan+15].

Given a distribution from the model class (4.1), the nodewise conditional distribution, i.e. the conditional distribution of each variable conditioned on the other observed variables and latent variables has the form

$$\mathscr{P}_{\text{cond}} = \left\{ p(x_s | x_{\sim s}, z) = \exp\Big(\Lambda_{s,s} f(x_s) + x_s \eta_s - D(\eta_s) \Big) \right\}, \quad (4.2)$$

where $\eta_s = \alpha_s + e_s^T K x + e_s^T B z$ and $D(\eta_s; \Lambda_{s,s})$ is a normalization term

$$D(\eta_s; \Lambda_{s,s}) = \log \left[\int \exp\{\Lambda_{s,s} f(x) + x\eta_s\} \mathrm{d}x \right].$$

It is straightforward to check that $\mathbb{E}[x_s|x_{\sim s}, z] = D'(\eta_s; \Lambda_{s,s})$ and $\text{var}[x_s|x_{\sim s}, z] = D''(\eta_s; \Lambda_{s,s}) \geq 0$, which proves that $D(\cdot)$ is a convex function of its input. Assuming that Λ is known (which is sensible for many variable types), we have that (4.2) is a generalized linear model with a linear predictor η_s and the link function $D'(\eta_s; \Lambda_{s,s})$.

Instantiations of Model

We consider joint distributions where the conditional distribution of observed variables conditioned on the latent variables is either a Gaussian, Ising, Poisson, or Exponential pairwise graphical model with the the node-wise conditional distribution of the form (4.2). These distributions all fall inside our modeling framework (4.1) with the following restrictions on the parameters:

1. Gaussian: $f(x_s) = x_s^2$, $D(\eta_s; \Lambda_{s,s}) = \frac{\Lambda_{s,s}^{-1}\eta_s^2}{2} + \log(2\pi\Lambda_{s,s}^{-1})$, and Λ represents a diagonal matrix with each element encoding the inverse of the conditional variance.
2. Ising: $f(x_s) = 0$, $D(\eta_s) = \log(\exp(\eta_s) + \exp(-\eta_s))$, and $\Lambda = \mathcal{I}$.
3. Poisson: $f(x_s) = -\log(x_s)$, $D(\eta_s) = \exp(\eta_s)$ with $K \leq 0$, and $\Lambda = \mathcal{I}$.
4. Exponential: $f(x_s) = 0$, $D(\eta_s) = -\log(\eta_s)$ with $K \leq 0$ and $\alpha > 0$, and $\Lambda = \mathcal{I}$.

The element wise positivity or negativity conditions on the parameters K, B, α in the Poisson or Exponential setting are enforced so that the corresponding distributions are normalizable [CWS15].

4.3 Pusedo-Likelihood Estimator

Our objective is to fit a conditional graphical model (of either Gaussian, Ising, Poisson, or Exponential) conditioned on some latent variables from the class (4.1). Specifically, let the model parameters (including latent variables instantiations) be denoted by $\theta = (K, B, \Lambda, \alpha) \subset \mathbb{R}^{p\times p} \times \mathbb{R}^{p\times k} \times \Lambda \in \mathbb{R}^{p\times p} \times \alpha \in \mathbb{R}^p$, and $\{(x^{(i)}, z^{(i)})\}_{i=1}^n \subset \mathbb{R}^p \times \mathbb{R}^k$ be n iid realizations of the observed and latent variables. Then the maximum-likelihood estimator with respect to θ and latent observations $\{z^{(i)}\}_{i=1}^n$ is given by:

$$(\hat{\theta}, \{\hat{z}^{(i)}\}_{i=1}^n) = \arg\min_{\theta} \quad \sum_{i=1}^n -\log\left[\Pr\left(x^{(i()}|z^{(i)}\right)\right], \tag{4.3}$$

where

$$-\log\left[\Pr\left(x^{(i()}|z^{(i)}\right)\right] = \sum_{s=1}^p \Lambda_{s,s} f(x_s) + \alpha^T x^{(i)} + \sum_{s=1}^p {x^{(i)}}^T K x^{(i)} + {x^{(i)}}^T B z^{(i)} + A(K; B; \alpha).$$

There are two main challenges to solving (4.3). First, the normalization constant $A(K; B; \Lambda; \alpha)$ is intractable to compute in high-dimensions. As an example, for Bernoulli variables, the complexity of computing $A(K; B; \Lambda; \alpha)$ is on the order $O(2^p)$, which is a bottleneck when p is large. Second, both the parameter B and the latent variable observations $\{z\}_{i=1}^n$ are unknown which lead to a non-convex estimator.

We begin with addressing the first challenge and assume that the observations $\{z\}_{i=1}^n$ are known and let $\hat{\theta}_{\texttt{MLE}}$ be the maximum-likelihood estimate from solving (4.3) without the decision variables $\{z\}_{i=1}^n$. To circumvent the computational complexity of exact inference with a full likelihood function, [Bes75] introduced a pseudo-likelihood approximation of the maximum likelihood. In particular, the pseudo-likelihood approximates the joint distribution as the *product* of node-wise conditional distributions so that $\Pr(x|z) \approx \prod_{s=1}^p \Pr(x_s|z)$. Then the pseudo-likelihood estimator is given by

$$\arg\min_{\theta} \sum_{i=1}^{n} \sum_{s=1}^{p} \log \left[\Pr\left(x_s^{(i)} | x_{\sim s}^{(i)}; z^{(i)} \right) \right], \tag{4.4}$$

where

$$-\log \left[\Pr\left(x_s^{(i)} | x_{\sim s}^{(i)}; z^{(i)} \right) \right] = -\Lambda_{s,s} f(x_s) - x_s^{(i)} M_{s,i}(\theta) + D(M_{s,i}(\theta))$$

with $M_{s,i}(\theta) = \alpha_s + e_s^T K x^{(i)} + e_s^T B z^{(i)}$. Recall from Section 4.2 that M is a linear predictor of the parameters θ and $D(\cdot)$ is a convex link function, so that (4.4) is a convex program that can be solved efficiently and optimally. The psuedo-likelihood leads to a computational and statistical tradeoff. In particular, [LJ08] show that psuedo-likelihood estimator remains asymptotically consistent and normal, but is usually statistically less efficient the maximum-likelihood estimator — a sacrifice for computational efficiency.

Arriving at the estimator (4.4) was based on the assumption that the observations $\{z^{(i)}\}_{i=1}^n$ are known. We now remove this assumption. Let $X \in \mathbb{R}^{p \times n}$ and $Z \in \mathbb{R}^{r \times n}$ (unknown) be the concatenation of observed and latent variables. Then, an equivalent formulation of (4.4) is given by:

$$\begin{aligned} \underset{K,\alpha,B,Z,M}{\text{argmin}} \quad & \ell(X; M), \\ \text{subject-to} \quad & M = KX + BZ + \alpha \mathbf{1}' \end{aligned}$$

where $\ell(X; M) = \frac{1}{n} \sum_{i=1}^n \sum_{s=1}^p -X_{s,i} M_{s,i} + D(M_{s,i})$. A remaining challenge is that both the observations Z as well as B are unknown; in fact generically, we can only

identify the column space of $L = BZ$, or the effect of latent variables on observed (in settings where Z have structure such as signed values, additional aspects of L may be recovered). To make the problem tractable, we must exploit structure about the latent variable observations. Note that assuming $r \ll \min\{p, n\}$, L has small rank. Hence, we replace BZ with L and impose low-rank structure on L. Furthermore, the matrix K encoding the conditional dependency structure (conditioned on latent variables) is expected to be sparse. As such, we also impose a sparsity structure on K leading to the following regularized pseudo-likelihood estimator:

$$\begin{aligned}(\hat{L}, \hat{K}, \hat{\alpha}, \hat{M}) = \underset{L,K,\alpha,M}{\operatorname{argmin}} \quad & \ell(X; M) + \lambda(\|K\|_1 + \gamma\|L\|_\star). \\ \text{subject-to} \quad & M = KX + L + \alpha\mathbf{1}'; \;\; K_{i,j} = 0 \iff K_{j,i} = 0; \;\; K_{s,s} = 0\end{aligned}$$

Here, the regularization parameter $\lambda \geq 0$ controls the tradeoff between fit to data and complexity of model and the regularization parameter $\lambda \geq 0$ controls the tradeoff between the sparsity of K and rank of L. The constraint $K_{i,j} = 0 \iff K_{j,i} = 0$ ensures that the conditional dependency structure is consistent. The constraint $K_{i,j} = 0 \iff K_{j,i} = 0$ is difficult to optimize over. We overcome this challenge by relaxing the constraint $K_{i,j} = 0 \iff K_{j,i} = 0$ by the symmetry constraint $K = K^T$ to obtain the following convex optimization estimator we employ in this paper

$$\begin{aligned}(\hat{L}, \hat{K}, \hat{\alpha}, \hat{M}) = \underset{L\in\mathbb{R}^{p\times n}, K\in\mathbb{S}^p, \alpha\in\mathbb{R}^p, M\in\mathbb{R}^{p\times n}}{\operatorname{argmin}} \quad & \ell(X; M) + \lambda(\|K\|_1 + \gamma\|L\|_\star). \\ \text{subject-to} \quad & M = KX + L + \alpha\mathbf{1}'; \;\; K_{s,s} = 0\end{aligned} \tag{4.5}$$

A few remarks are in the order:

Remark 1: Recall from section 4.2 that the parameters K, B, α lie in a restricted space for certain distributions. These restrictions can be naturally added in the estimator (4.5). As an example, in the setting where both the observed and latent variables are Poisson, the constraints $K \leq 0$ and $B \leq 0$ must be added to (4.5),

Remark 2: In the settings where $\Lambda = \mathcal{I}$, the symmetrization constraint in (4.5) is not a relaxation of $K_{i,j} = 0 \iff K_{j,i} = 0$ since K in (4.1) can be taken to be symmetric without loss of generality. In the Gaussian setting under different conditional variances so that $\Lambda \neq \mathcal{I}$, an alternate approach to ensure consistency in conditional dependency structure is the following AND-OR post-processing step [MB06]:

$$\hat{E}_{\text{or}} = \{(s,t), \hat{\Theta}_{s,t} \neq 0 \vee \hat{\Theta}_{t,s} \neq 0\} \;\; ; \;\; \hat{E}_{\text{and}} = \{(s,t), \hat{\Theta}_{s,t} \neq 0 \wedge \hat{\Theta}_{t,s} \neq 0\}$$

Remark 3: There is an intimate connection between the loss-function $\ell(X; M)$ in (4.5) and the Bregman divergence. Recall that the Bregman divergence of two matrices P and Q with respect to a function ψ is given by:

$$d_\Psi(P, Q) = \Psi(P) - \Psi(Q) - \text{trace}\left(\nabla_Q(P - Q)\right).$$

The loss function $\ell(X; M)$ for each distributional setting is equal to $d_\Psi(X, f(M))$ for appropriate choice of function $\psi : \mathbb{R}^{p\times n} \to \mathbb{R}$ and map $f : \mathbb{R}^{p\times n} \to \mathbb{R}^{p\times n}$:

1. Gaussian: $\Psi(P) = \frac{1}{2n}\|P\|_F^2$ so that $d_\Psi(P, Q) = \frac{1}{2n}\|P - Q\|_F^2$ and f is the identify map.

2. Poisson: $\Psi(P) = \frac{1}{n}\sum_i \sum_j P_{i,j}\log(P_{i,j}) - P_{i,j}$, $d_\Psi(P, Q) = \frac{1}{n}\sum_i \sum_j P_{i,j}\log(P_{i,j}/Q_{i,j}) - P_{i,j} + Q_{i,j}$ which is the generalized relative entropy, and f is the exponential map.

3. Exponential: $\Psi(P) = \frac{1}{n}\sum_i \sum_j -\log(P_{i,j}) - 1$, $d_\Psi(P, Q) = \frac{1}{n}\sum_i \sum_j P_{i,j}/Q_{i,j} - \log(P_{i,j}/Q_{i,j}) - 1$ and f is the reciprocal map.

4. Ising: $\Psi(P) = \frac{1}{n}\sum_i \sum_j -\log(P_{i,j}) - 1$, $d_\Psi(P, Q) = \frac{1}{n}\sum_i \sum_j P_{i,j}/Q_{i,j} - \log(P_{i,j}/Q_{i,j}) - 1$ and f is the identity map.

Tailored Latent Variable Regularizers

Latent variables often have structure, such as signed values for Bernoulli variables, or non-negative values for Poisson and exponential variables. These structures translate to additional constraints on $L = BZ$ on top of low-rankness. Specifically, the structures on L can be represented mathematically via a weighted sum of *atoms*[Cha+12]:

$$L = \sum_{i=1}^{p} c_i a_i b_i^T \;\; ; \;\; a_i b_i^T \in \mathcal{A}. \tag{4.6}$$

Here $\mathcal{A}$ consists of a collection of atoms that is a compact subset of $\mathbb{R}^{p\times n}$ and $\alpha_i \geq 0$ are a set of coefficients. Next we describe the atomic set $\mathcal{A}$ for different latent-variable types

- **G**aussian: $\mathcal{A} = \{ab^T \mid a \in \mathbb{R}^p, b \in \mathbb{R}^n, \|a\|_2 = 1, \|b\|_2 = 1\}$
- **B**ernoulli: $\mathcal{A} = \{ab^T \mid a \in \mathbb{R}^p, b \in \mathbb{R}^n, \|a\|_2 = 1, b = \{\pm 1\}^n/\sqrt{n}\}$
- **P**oisson: $\mathcal{A} = \{ab^T \mid a \in \mathbb{R}^p, b \in \mathbb{Z}_+^n, \|a\|_2 = 1\}$

- **E**xponential: $\mathcal{A} = \{ab^T \mid a \in \mathbb{R}^p, b \in \mathbb{R}^n_+, \|a\|_2 = 1\}$

A natural approach to induce the appropriate structure underlying L is the atomic norm $\|L\|_{\mathcal{A}}$ [Cha+12]:

$$\|L\|_{\mathcal{A}} = \inf \left\{ \sum_{ab^T \in \mathcal{A}} c_{ab} \; : \; L = \sum_{ab' \in \mathcal{A}} c_{ab} ab^T, c_{ab} \geq 0 \text{ for all } ab^T \in \mathcal{A} \right\} \quad (4.7)$$

The set $\mathcal{A}$ is non-convex for each distributional setting. Hence, we consider outer relaxations $\tilde{\mathcal{A}}$ to $\mathcal{A}$ so that $\mathcal{A} \subseteq \tilde{\mathcal{A}}$. We next describe the outer convex relaxations that we employ in each distributional setting to find a convex norm function $\|L\|_{\tilde{\mathcal{A}}}$.

- Gaussian: we consider the relaxation:

$$\tilde{\mathcal{A}} = \left\{ M \in \mathbb{R}^{p \times n} \mid \exists W_1 \in \mathbb{S}^p, W_2 \in \mathbb{S}^n \text{ such that } ; \begin{pmatrix} W_1 & M \\ M^T & W_2 \end{pmatrix} \geq 0 \right.$$
$$\left. ; \quad \frac{1}{2}\left(\text{trace}(W_1) + \text{trace}(W_2)\right) \leq 1 \right\}.$$

  Consequently, plugging in $\tilde{\mathcal{A}}$ into (4.7) yields:

$$\|L\|_{\tilde{\mathcal{A}}} = \min \left\{ \frac{1}{2}\text{trace}(W_1) + \frac{1}{2}\text{trace}(W_2) \right.$$
$$\left. : \quad \begin{pmatrix} W_1 & L \\ L' & W_2 \end{pmatrix} \geq 0 \; : \; \text{for } W_1 \in \mathbb{S}^p, W_2 \in \mathbb{S}^n \right\}.$$

  By definition, $\|L\|_{\tilde{\mathcal{A}}} \leq \|L\|_{\mathcal{A}}$ and furthermore $\|L\|_{\tilde{\mathcal{A}}} = \|L\|_\star$.

- Bernoulli: we consider the relaxation:

$$\tilde{\mathcal{A}} = \left\{ M \in \mathbb{R}^{p \times n} \mid \exists W_1 \in \mathbb{S}^p, W_2 \in \mathbb{S}^n \text{ such that } ; \begin{pmatrix} W_1 & M \\ M^T & W_2 \end{pmatrix} \geq 0 \right.$$
$$\left. ; \quad \text{trace}(W_1) \leq 1 \; ; \; (W_2)_{i,i} = \frac{1}{n} \text{ for all i} \right\}.$$

  Consequently, plugging in $\tilde{\mathcal{A}}$ into (4.7) yields:

$$\|L\|_{\tilde{\mathcal{A}}} = \min \left\{ \frac{1}{2}\text{trace}(W_1) + \frac{n}{2}\max(W_2)_{i,i} \right.$$
$$\left. : \quad \begin{pmatrix} W_1 & L \\ L' & W_2 \end{pmatrix} \geq 0 \; : \; \text{for } W_1 \in \mathbb{S}^p, W_2 \in \mathbb{S}^n \right\}.$$

  By definition, $\|L\|_{\tilde{\mathcal{A}}} \leq \|L\|_{\mathcal{A}}$ and it is straightforward to check that $\|L\|_\star \leq \|L\|_{\tilde{\mathcal{A}}}$. In other words, $\|\cdot\|_{\tilde{\mathcal{A}}}$ is more appropriate than the nuclear norm.

- Poisson and Exponential: A straightforward relaxation of $\mathcal{A}$ is

$$\mathcal{A}_{\text{compl. pos. cone}} = \left\{ M \in \mathbb{R}^{p\times n} \mid \exists W_1 \in \mathbb{S}^p, W_2 \in \mathbb{S}^m \text{ such that} \right.$$
$$; \quad \begin{pmatrix} W_1 & M \\ M^T & W_2 \end{pmatrix} \succeq 0;\ \frac{1}{2}\left(\text{trace}(W_1) + \text{trace}(W_2)\right) \leq 1$$
$$\left. ; \quad W_2 \in \text{Complete Positive Cone} \right\},$$

where the complete positive cone represents the collection of matrices $C \in \mathbb{S}^n$ that can be represented by the outer product $C = QQ^T$ for nonnegative Q. The complete positive cone is non-convex so we consider a further relaxation using the property that complete positive cone $\subseteq$ PSD Cone$\cap$Non-negativity Cone:

$$\tilde{\mathcal{A}} = \left\{ M \in \mathbb{R}^{p\times n} \mid \exists W_1 \in \mathbb{S}^p, W_2 \in \mathbb{S}^p \text{ such that } ; \begin{pmatrix} W_1 & M \\ M^T & W_2 \end{pmatrix} \succeq 0 \right.$$
$$\left. ; \quad \frac{1}{2}\left(\text{trace}(W_1) + \text{trace}(W_2)\right) \leq 1 \ ; \ W_2 \succeq 0 \right\}.$$

Consequently, plugging in $\tilde{\mathcal{A}}$ into (4.7) yields:

$$\|L\|_{\tilde{\mathcal{A}}} = \min \left\{ \frac{1}{2}\left(\text{trace}(W_1) + \text{trace}(W_2)\right) \right.$$
$$\left. : \quad \begin{pmatrix} W_1 & L \\ L' & W_2 \end{pmatrix} \succeq 0 \ : \ \text{ for } W_1 \in \mathbb{S}^p, W_2 \in \mathbb{S}^n_+ \right\}.$$

By definition, $\|L\|_{\tilde{\mathcal{A}}} \leq \|L\|_{\mathcal{A}}$ and it is straightforward to check that $\|L\|_\star \leq \|L\|_{\tilde{\mathcal{A}}} \leq \|L\|_{\mathcal{A}_{\text{comp. pos. cone}}} \leq \|L\|_{\mathcal{A}}$. In other words, the norm $\|\cdot\|_{\tilde{\mathcal{A}}}$ is strictly tighter than the nuclear norm.

Remark 4 In addition to the structure in latent variable observations Z, one might have a-priori structure on the matrix $B^\star$ as well. As an example, recall in the setting where all observed and latent variables are Poisson, the model (4.1) is normalizable if $B \leq 0$. This additional structure can be exploited by adding the constraint $W_1 \geq 0$ in the optimization problem for $\|L\|_{\tilde{\mathcal{A}}}$.

4.4 Experiments

In this section, we demonstrate the utility of latent variable graphical modeling for GLM's with synthetic data as well as real data involving voting record dataset and miRNA expression levels.

Consistency Simulations with Synthetic Data

We begin by giving experimental evidence for the utility of our proposed latent variable graphical modeling procedure on synthetic examples. Specifically, we generate from two types of distributions from the model class (4.1): Poisson graphical model with Bernoulli random variables and Bernoulli graphical model with Gaussian latent variables.

In the first experiment, we consider the setting where the latent variables are Bernoulli variables and the conditional distribution of $p = 60$ observed variables conditioned on the latent variables is given by a Poisson graphical model, which is a cycle with edge weights -0.2. In other words, $K^\star \in \mathbb{S}^{60}$ with $K_{i,i+1} = -0.4$. We vary the number of latent variables $k = \{1, 2, 3\}$ and generate the latent effect matrix $B^\star \in \mathbb{R}^{60\times k}$ as follows: sample the row-space of $B^\star$ uniformly from the Steifel manifold, and the column-space of $B^\star$ is chosen so that the incoherence $\text{inc}(\text{col-space}(B^\star)) = k/30$. The singular values of $B^\star$ are sampled uniformly from the interval $[3, 3.2]$. The latent variable Bernoulli variables are generated iid with mean 0.5. We further set $\alpha^\star \in \mathbb{R}^{60\times 1}$ to have entry-wise zero elements. These parameters specify the a distribution from the model class (4.1) with node function $f(x_s) = -\log(x_s)$ and link function $D(\eta_s) = \exp(\eta_s)$, and we generate $n = [1000, 2500]$ with interval 100 via Gibbs sampling to obtain observations $X \in \mathbb{R}^{60\times n}$. We then fit the observations to the estimator (4.5) with the link function D to identify parameters $(\hat{K}, \hat{L}, \hat{\alpha})$. To evaluate the performance of the estimator, in Figure 4.1(a) we represent over 10 trials the success probability based on correct graph recovery and rank recovery, e.g. $\text{support}(\hat{K}) = \text{support}(K^\star)$ and $\text{rank}(\hat{L}) = \text{rank}(B^\star)$. Evidently, for large enough samples, the estimator is consistent.

In the second experiment, we consider the setting where the latent variables are Gaussian variables and the conditional distribution of $p = 100$ observed variables conditioned on the latent variables is given by a Ising model which a random Erdös-Rényi graph with probability 0.05 and edge weights 0.2. This yields the matrix $K^\star \in \mathbb{S}^{100}$. We vary the number of latent variables $k = \{1, 2, 3\}$ and generate the latent effect matrix $B^\star \in \mathbb{R}^{100\times k}$ as follows: sample the row-space of $B^\star$ uniformly from the Steifel manifold, and the column-space of $B^\star$ is chosen so that the incoherence $\text{inc}(\text{col-space}(B^\star)) = k/50$. The singular values of $B^\star$ are sampled uniformly from the interval $[1.2, 1.5]$. The latent variable Bernoulli variables are generated iid with mean 0.5. We further set $\alpha^\star \in \mathbb{R}^{100\times 1}$ to have entry-wise zero elements. These parameters specify the a distribution from the model class (4.1)

with node function $f(x_s) = 0$ and link function $D(\eta_s) = \log(\exp(\eta_s) + \exp(-\eta_s))$, and we generate $n = [1000, 2500]$ with interval 100 via Gibbs sampling to obtain observations $X \in \mathbb{R}^{100\times n}$. We then fit the observations to the estimator (4.5) with the link function D to identify parameters $(\hat{K}, \hat{L}, \hat{\alpha})$. Figure 4.1(b) shows the success probability based on correct graph recovery and rank recovery, e.g. support$(\hat{K})$ = support$(K^\star)$ and rank$(\hat{L})$ = rank$(B^\star)$. Evidently, for large enough samples, the estimator is consistent.

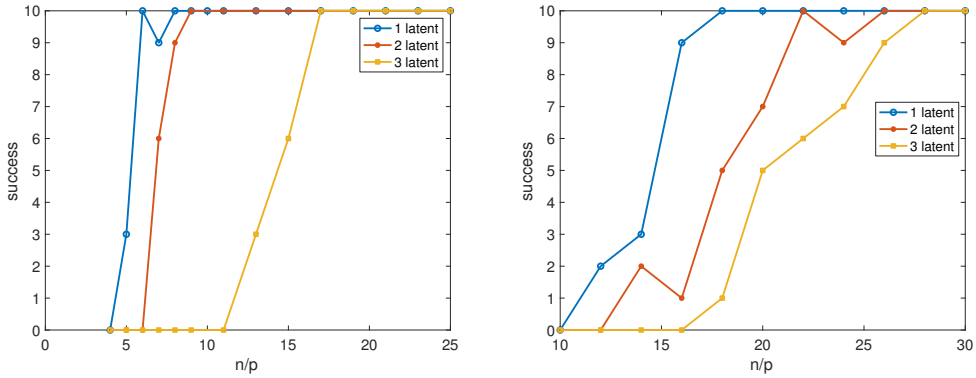


Figure 4.1: left: structure and rank consistency of Poisson (observed) and Bernoulli (latent) model with cyclic graph; right: structure and rank consistency of Bernoulli (observed) and Gaussian (latent) model with random graph with 5% sparsity.

The experimental results of Figure 4.1 check for rank and graph structure consistency. Naturally, one may wonder how close the estimated latent subspace is to the population latent subspace. Focusing on the Poisson-Bernoulli setup, Table 4.4 demonstrates that these two subspaces are close given sufficiently enough obsersations.

# of latent variables	$n = 20p$	$n = 10p$
1	3.4°	4.2°
2	7.5°	8.0°

Table 4.1: Finding the largest principal angle between the estimated latent space (e.g. col-space$(B^\star)$) and the population latent space (e.g. col-space$(\hat{L})$) for the Poisson-Bernoulli cycle with different number of latent variables and number of observations.

Benefits of Tailored Regularizers

In Section 4.3, we described tailored regularizers that exploit the fact that the latent effect $L = B^\star Z$ may have additional structure on top of low-rankness. In this

section, we explore the benefits of applying a more tailored regularizer via a stylized experiment. In particular, we consider the model $x = \mathcal{B}^\star z + \epsilon$, where $z \in \mathbb{R}^2$ is a Poisson random vector,$B^\star \in \mathbb{R}^{30\times 3}$ is a latent effect matrix, and $\epsilon \in \mathbb{R}^{30}$ is a Gaussian random vector with independent entries. Notice that $x|z$ is an independent Gaussian graphical model and thus is from the model class (4.1). We let the conditional variance of $\epsilon = 0.2$ and generate entries of $B^\star$ iid from a normal distribution. We generate $n = \{30, 40\}$ samples $X \in \mathbb{R}^{30\times n}$ from this distribution and input the data into the estimator (4.5) where K is set to 0. We compare the regularizers $\|\cdot\|_\star$ (e.g. nuclear norm) to the tailored regularizer $|\cdot\|_{\tilde{\mathcal{A}}}$ that is expressed by the following semidefinite program:

$$\begin{aligned}\|L\|_{\tilde{\mathcal{A}}} \;=\; \min \Bigg\{ & \frac{1}{2}\left(\mathrm{trace}(W_1) + \mathrm{trace}(W_2)\right) \\ & : \quad \begin{pmatrix} W_1 & L \\ L' & W_2 \end{pmatrix} \geq 0 \;:\; \text{for } W_1 \geq 0, W_2 \geq 0 \Bigg\}.\end{aligned}$$

Table 4.4 compares the performance of the nuclear norm estimator and the tailored norm estimator. Specifically, we choose λ in each estimator to find a rank 1 estimate and find FDR and Power (from 10 trials) associated with column-spaces estimates of each approach. Recall from Chapter 3 that FD and PW in subspace estimation is given by:

$$\mathrm{FD} = \mathbb{E}\left[\mathrm{trace}\left(\mathcal{P}_{\hat{C}}\mathcal{P}_{C^{\star\perp}}\right)\right] \;\;;\;\; \mathrm{PWR} = \mathbb{E}\left[\mathrm{trace}\left(\mathcal{P}_{\hat{C}}\mathcal{P}_{C^\star}\right)\right].$$

Examining Table 4.4, we observe that the tailored regularizer produces smaller FD and larger PW.

Regularizer	n = 30	n = 40
Nuclear norm FD; PW	1.17 ; 0.82	1.10; 0.89
Tailored norm FD; PW	0.56; 0.86	0.54; 0.91

Table 4.2: Comparing the FD and PW of the column-space estimate obtained from employing a nuclear norm vs a tailored regularizer that exploits the structure of the latent variables.

Experiments with Real Dataset

Solving the estimator (4.5) requires choosing the regularizers λ, γ as input. A typical approach for model selection is cross-validation that identifies a model from training data (e.g. a subset of the columns of X) and validates the model on a separate

test data (e.g. an alternative subset of the columns of X). It is straightforward to see that this approach for selecting a model will yield a full-rank $\hat{L}$ and a completely disconnected graph $\hat{K}$. Hence, we develop an alternate model selection approach. Specifically, inspired by the model selection technique developed in [LRW10], our approach is to choose a *low-complexity* and *stable* model. We measure complexity by $|\hat{K}| + 2p \operatorname{rank}(\hat{L}) - \operatorname{rank}(\hat{L})^2$. The stability metric is measured via the tangent space concepts in Chapter 3. In particular, for a given λ, γ, we obtain $n/2$ bags of data and compute for each bag the estimate $\hat{K}, \hat{L}$. From these estimates, we find the tangent spaces $T(\hat{K})$ and $T(\hat{L})$, where the first tangent space is with respect to the sparse matrix variety and the second is with respect to the quotient manifold of the determinantal variety (see Chapter 3 for more details). Across B bags, we compute the variability of the tangent space estimates via the following terms:

$$\tau_{\text{low-rk}} = \frac{\operatorname{trace}\left(\left(\frac{1}{B}\sum_{i=1}^{B}\mathcal{P}_{T_i(\hat{L})}\right)^2\right) - \operatorname{trace}\left(\left(\frac{1}{B}\sum_{i=1}^{B}\mathcal{P}_{T_i(\hat{L})}\right)\right)}{p^2}$$

$$\tau_{\text{sparse}} = \frac{\operatorname{trace}\left(\left(\frac{1}{B}\sum_{i=1}^{B}\mathcal{P}_{T_i(\hat{K})}\right)^2\right) - \operatorname{trace}\left(\left(\frac{1}{B}\sum_{i=1}^{B}\mathcal{P}_{T_i(\hat{K})}\right)\right)}{p^2}$$

and subsequently $\tau = \tau_{\text{low-rk}} + \tau_{\text{sparse}}$. To better understand the quantities $\tau_{\text{low-rk}}$ and τ_{sparse}, note that using the idempotence of projection operators, the numerators of these quantities compute a variance metric, and the normalization by p^2 ensures that they are between $[0, 1]$. If the tangent spaces are stable across bags, then both τ_{sparse} and $\tau_{\text{low-rk}}$ will be small. We note that τ_{sparse} is precisely that variability metric introduced in [LRW10]. Our procedure for selecting a model is then as follows: given a variability threshold $(0, 1)$, we sweep over the regularization parameters and choose the lowest complexity model that has τ approximately equal to this threshold.

Ising Model for Senate Voter Record Dataset We apply our latent-variable modeling framework to the 109th Senate voting record dataset. The dataset was obtained from the website of the US Congress (`http://www.senate.gov`). It contains the voting records of the 100 senators of the 109th congress (January 3, 2005 — January 3, 2007) on 645 bills that the senate voted on. The votes are recorded as one for "yes" and zero for "no". The data contains missing votes as some senators abstained on certain bills. The missing values (missed votes) for each senator were imputed with the majority vote of that senator's party on that particular bill and the missing votes of the Independent Senator Jeffords were imputed with the Democratic majority vote. Finally, we exclude bills where the "yes/no" proportion fell outside the

interval $[0.3, 0.7]$. This results in $n = 343$ votes across $p = 100$ senators with a data matrix $X \in \mathbb{R}^{100\times 343}$.

Since the data consists of binary variables, we consider a model where the conditional distribution of senators conditioned on latent variables is modeled by an Ising graphical model. Specifically, we supply the data to the estimator (4.5) with $D(x) = \log(\exp(x) + \exp(-x))$ where the regularization parameters λ, γ are selected based on the model selection technique described earlier. Figure 4.2 displays the learned graphical model without latent variables that has 5% sparsity and one with 4 latent variables and 2% sparsity. As expected, for both models, the republicans and democrats cluster. Furthermore, by introducing latent variables, we tease away the first order dependencies that occur due to similar party affiliation and obtain residual dependencies.

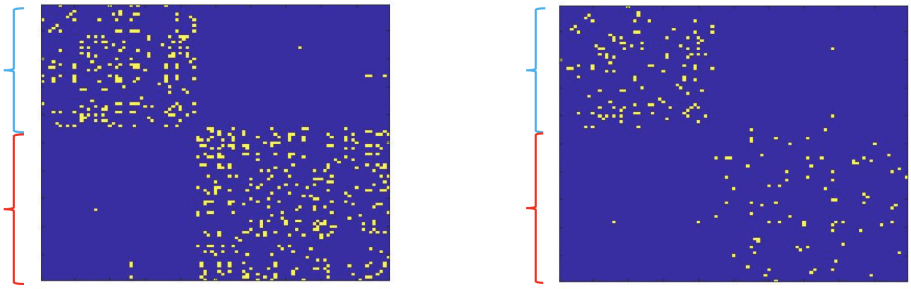


Figure 4.2: (left) Graphical model without latent variables having 5% sparsity and (right) graphical model with rank 4 and 2% sparsity. Here senators are clustered together according to their party affiliation with Democrats labeled by blue bracket and Republicans by red bracket.

Poisson Model for RNA-Seq Count Data

We next demonstrate the applicability of our approach by estimating miRNA inhibitory network for Level III breast cancer miRNA expression (downloaded from `http://tcga-data.nci.nih.gov/tcga/`). The data consists of 262 miRNA's and 544 subjects. The data was approximately made Poisson by following the steps described in [AL13]. Specifically, the data was quantile adjusted to correct for sequencing depth, the miRNAs with little variation across the samples (the bottom 50%), were filtered out, and the data was adjusted for possible over-dispersion using a power transform and a goodness of fit test. Further, since our model class only allows for negative dependencies, we group strongly positively correlated miRNA's using hierarchical clustering with average linkage and one minus the correlation as the distance. This resulted in 40 clusters of tightly positively correlated miRNAs

and our miRNA network was taken as miRNA closest in Euclidean distance to the cluster centroid, in each group. As a result of these processing techniques, we obtain a data matrix $X \in \mathbb{R}^{40\times 544}$ that is well-modeled by a Poisson distribution.

Supplying X as input to (4.5) with $D(x) = \exp(x)$ and selecting the regularization parameters λ, γ via the model selection technique described earlier, we find a latent variable graphical model with 5 latent variables and 32.5% sparsity level. Figure 4.4 shows the graphical structure with the strongest edge highlighted in the network.

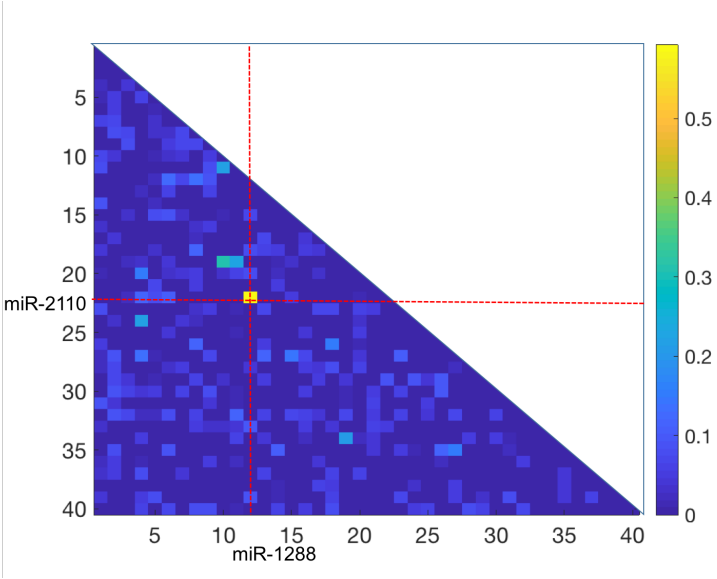


Figure 4.3: The graphical model with latent variables underlying 40 miRNA. The strongest edge is between miR-1288 and miR-2110.

To validate this model, we perform the following test. Recall that the modeling framework (4.1) implies that the node-wise conditional is distributed according to an exponential family. If the neighborhood and the latent variable observations are known, goodness-of-fit can be assessed to a null model by a likelihood ratio test that follows a chi-squared distribution [MN83]. However the addition of ℓ_1 penalty and the nuclear norm penalty makes this test inexact and likely to be conservative. In the absence of exact tests, we propose the heuristic: estimate the overall graphical structure and the latent subspace. Then, restricted to the graph structure and latent subspace, solve (4.5) without the regularizers, compare the fit of this model to that of a null model (only an intercept term) via the likelihood ratio test. Using this approximate test, we find that 34 of the 40 nodes have p-values less than 0.05. Furthermore 32 of the 40 nodes have p-values less than the Bonferroni-corrected value of $0.05/40$. Evidently, our model is a good approximation of the data.

4.5 Discussions

In this chapter, we proposed a framework to identify a latent variable graphical model from data for conditional GLM's. Our approach provides a flexible setup to enable modeling of Gaussian, Bernoulli, and Poisson variables. The proposed algorithm was based on a psuedo-likelihood approach that led to a convex optimization problem over all the parameters of interest. A particularly novel aspect of our formulation is that it incorporates regularizers that are tailored to the type of latent variables: nuclear norm for Gaussian latent variables, max-2 norm for Bernoulli variables, and complete positive norm for Poisson variables. For each case, we provide a semidefinite relaxation and demonstrate that the associated norm yields a better sample complexity (than the nuclear norm) for similar computational cost. There are several interesting avenues for further investigation that arise from our work. First, we employed an approximate goodness of fit test to validate the model we obtained on real data. It is of interest to develop exact tests. Along this direction, there may be observed data that are not heavily influenced by the impact of latent phenomena. Naturally, one would wish to perform a statistical test for the presence of latent variables. Finally, the modeling class (3.12) does not account for mixed observed variables (e.g. Gaussian and Bernoulli variables). It is of interest to appeal to ideas developed in [CWS15] to extend our framework to enable heterogeneity in data types.

Chapter 5

LATENT VARIABLE MODEL SELECTION WITH NON-IID DATA & APPLICATION TO HYPERSPECTRAL IMAGING

The data we observe and process is typically both non-iid and high-dimensional. As an example, the reservoir volumes that were analyzed in Chapter 2 exhibit significant temporal correlations so that the data is non-iid, and the reservoir network is large so that the data is high-dimensional. An application that will be explored in this chapter is hyperspectal imaging, where the data is highly spatially correlated to the continuity of a natural image (non-iid) and the number of spectral channels is large (high-dimensional). In this chapter, we will address this modeling challenge via a novel framework that is amenable to efficient algorithms.

The results in this chapter will be submitted in a paper that is in preparation. This work was joint with Andrew Stuart, David Thompson, Michael Turmon, and Venkat Chandrasekaran. The author contributed by helping develop the modeling framework and associated parameter estimation algorithm, as well as producing numerical experiments. The description of the work contained in this chapter was written by the author.

5.1 Introduction

In Chapters 1, 2, and 3, we described a number of latent-variable modeling techniques and estimators that account for the presence of unobserved phenomena when modeling the behavior of a large collection of observed variables (i.e. high-dimensional data). These procedures are typically derived from assuming independence of the data points and finding parameters that maximize the likelihood of observing the given data (i.e. maximum-likelihood estimator). However, much of data we observe and process is non-iid. As an example, reservoir volumes (analyzed in Chapter 2) exhibit significant temporal correlations so that the data is non-iid [Tae+17]. As another example, the data underlying a hyperspectral image (discussed in detail in this chapter) contains strong spatial dependencies due to the natural continuity of the scene and thus the data is far from independent. To overcome the strong dependencies of observations, one typically performs some preprocessing to make the data look independent. As an example, in the reservoir setting, we applied standard seasonal adjustment techniques to substantially reduce

the dependencies among reservoirs observations across months and years.

In this chapter, we blend ideas from Stochastic PDE and high-dimensional latent variable modeling to develop a rigorous technique to account for the non-iid structure in high-dimensional data. For the sake of concreteness, we consider modeling a collection of jointly Gaussian variables. We model this data as arising from a *factorizable precision operator* that leads to a rich model class that is amenable to an efficient estimator to identify the parameters of this model.

Factorizable Precision Operators

We consider observations from a random field with an underlying precision operator $\mathcal{P}$. Using the Kronecker product notation, we assume $\mathcal{P}$ takes the form:

$$\mathcal{P} = \mathcal{G}^{(1)}(\theta_1; x_1) \otimes \mathcal{G}^{(2)}(\theta_2; x_2) \ldots \otimes \mathcal{G}^{(m)}(\theta_m; x_m) \otimes \mathcal{A}. \tag{5.1}$$

Here $\mathcal{A} \in \mathbb{S}^p$ characterizes the covariance structure across the p variables and the operators $\{\mathcal{G}^{(i)}\}_{i=1}^m$ (which may be discrete or continuous) encode different components of the observation dependencies across dimensions x_i in the boundary $\mathcal{D}_i$. Each operator is specified by parameters $\theta_i \in \mathbb{R}^{q_i}$. Letting $u(x_1, x_2, \ldots, x_m) \in \mathbb{R}^p$ be an observation from the field, the Kronecker product representation means that the action of the operator $\mathcal{P}$ on a function $u(x_1, x_2, \ldots, x_k)$ may be viewed as each of the factorized operators acting on the specific coordinate while keeping the other fixed. The factorized precision form (5.1) provides a rich framework to model data arising from many real-world applications including measurements from hyperspectral imaging and spatiotemporal data.

Application to Imaging Spectroscopy: Imaging spectroscopy is the process of collecting and analyzing information across the electromagnetic spectrum. The objective of imagine spectroscopy is to use spectral properties of a scene —- obtained typically by instruments with sensors that fly over air —- to identify materials or objects that may be present. As demonstrated in Figure 5.1, the data that is obtained from an imaging instruments comes in the form of a collection of images across sensors and flight line, with each image associated to a particular wavelength. Statistical modeling is a crucial piece of the imaging spectroscopy estimation and detection pipeline. Specifically, recovery of surface reflectance by optimal estimation techniques relies on an accurate assessment of the noise covariance for each sensor. Furthermore, the accuracy of existing detection techniques (e.g., matched filters) largely depends on a faithful reconstruction of the statistics of the scene such as the underlying covariance matrix. For push-broom imaging instruments (e.g.,

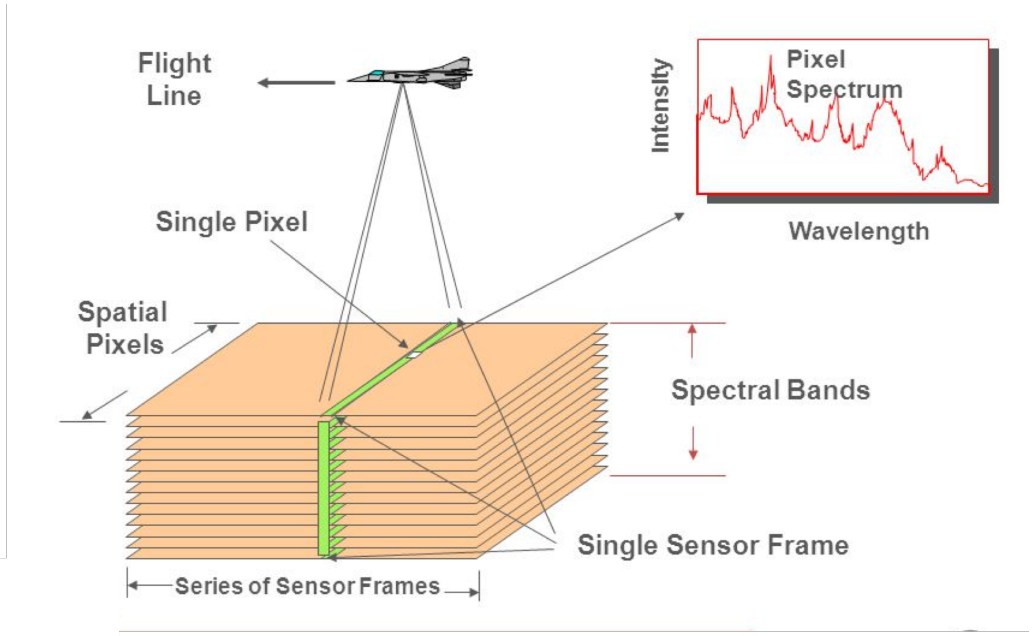


Figure 5.1: Imaging spectroscopy data collection: instrument with multiple sensors hovering over space in a flight line collecting spectral profile for each pixel location.

Aviris NG), a challenges that arise in estimating a representative covariance matrix is that the spatial continuity of scenes implies that observed samples are highly correlated and should not be treated as independent. Existing approaches address these challenges by applying an ad-hoc preprocessing step to de-correlate the samples. To overcome this challenge, we model the data via the factorizable Gaussian process (5.1). In particular, the matrix $\mathcal{A}$ encodes the dependencies among the spectral channels that combine to form the data. Furthermore, a single precision operator $\mathcal{G}$ may be used to encode the spatial dependencies (the relationship among nearby locations in the image). In a latter part of this section, we describe the Whittle Matèrn covariance model that is a natural candidate for the precision operator $\mathcal{G}$.

Application to Spatiotemporal Data: Spatiotemporal models are ubiquitous in applications ranging from climate science, transportation, and social media. Here, there are two types of dependencies, those across time, and those across space. The modeling framework (5.1) is appropriate in this setting. Specifically, suppose we are observing a collection of p variables $z(t, x) \in \mathbb{R}^p$ indexed by time t and location x. Letting the operator $\mathcal{G}^{(1)}(\theta_1; x_1)$ account for time, $\mathcal{G}^{(2)}(\theta_2; x_2)$ account for location, and the matrix $\mathcal{A}$ account for dependencies among the variables, the data can be

modeled as coming from a Gaussian process with the precision operator:

$$\mathcal{P} = \mathcal{A} \otimes \mathcal{G}^{(1)}(\theta_1; x_1) \otimes \mathcal{G}^{(2)}(\theta_2; x_2).$$

Whittle Matérn Process

A conceptually appealing process to encode the spatial dependencies among observations in a d-dimensional field is the Whittle-Matérn distribution. This class of stationary Gaussian distributions allow control over smoothness, amplitude, and length scale with a covariance function:

$$\Sigma_{\nu,\ell}(x, y) = \sigma^2 \frac{2^{1-nu}}{\Gamma(\nu)} \left(\frac{|x - y|}{\ell} \right)^{\nu} K_\nu \left(\frac{|x - y|}{\ell} \right), \tag{5.2}$$

where K_ν is the modified Bessel function of the second kind of order ν and $x, y \in \mathbb{R}^d$. The smoothness parameter ν controls the regularity of the process generated by the covariance (5.2) with larger values indicating more regularity; in particular $\nu = 1/2$ yields the exponential covariance and $\nu = \infty$ yields the Gaussian covariance. Furthermore, the parameter $\ell > 0$ acts as a characteristic length scale, and $\sigma > 0$ is the amplitude of the Gaussian process. Due to the flexibility of these distributions, they have been widely employed in applications ranging from spatial statistics, geostatistics, machine learning, and image analysis.

The process $u(x)$ generated by the covariance model (5.2) obeys a certain type of Stochastic Partial Differential Equation (SPDE), derived by [LHR14]:

$$\frac{1}{\sqrt{\beta \ell^d}} (\mathcal{I} - \ell^2 \Delta)^{(\nu+d/2)/2} u = W, \tag{5.3}$$

where W is white noise on $\mathbb{R}^d$, and $\beta = \frac{2^d \pi^{d/2} \Gamma(\nu+d/2)}{\Gamma(\nu)}$. From (5.3), it follows that the precision operator associate with the covariance function is given by:

$$\mathcal{G}(\ell, \nu) = \frac{1}{\ell^d} \left(\mathcal{I} - \ell^2 \Delta \right)^{\nu+d/2}. \tag{5.4}$$

Here, we have removed the dependence of $\mathcal{G}(\ell, \nu)$ on β as it will be absorbed by the matrix $\mathcal{A}$ in (5.1). Throughout, we assume Δ is the Laplacian with Dirichlet boundary conditions on domain D; it is hence invertible, self-adjoint, positive-definite on $L^2(D)$. This property implies that precision operator restricted to the interactions within the boundary and outside the boundary is zero; the statistical interpretation is that variables defined with respect to the precision $\mathcal{P}$ inside the boundary are conditionally independent of variables with respect to the precision $\mathcal{P}$

outside the boundary. The parameter ℓ characterizes the length scale of the random field and ν the smoothness. The formulation of $\mathcal{G}(\ell, \nu)$ has the property that the variance of the random field at any point is independent of ℓ (up to boundary effects which can be ameliorated by defining the covariance operator on a larger domain than that where observations are made, giving approximate stationarity — stationarity holds on the unbounded domain). See the formula for the marginal variance on p. 427 of [LRL11], rescale the covariance operator to make this independent of κ and then set $\kappa = \ell^{-1}$. Alternatively see Theorem 1 of [DIS17].

Our Contributions

As our first contribution, we develop in Section 4.2, a regularized maximum likelihood estimator to solve for the parameters $\{\theta_i\}_{i=1}^m$ in (5.1). Using n gridded observations from the random field as input, this estimator resembles the standard log-det estimator for identifying a precision matrix underlying the data, where the sufficient statistic is a modified sample covariance matrix (the modification is dependent on the parameters $\{\theta_i\}_{i=1}^m$). An appealing attribute of the proposed estimator is that a graphical model, latent-variable graphical model structure, or a factor modeling structure, can be readily imposed via a plugin regularization function of $\mathcal{A}$. Furthermore, under the assumptions that the eigenfunctions of $\mathcal{G}^{(i)}(\theta_i)$ are known and their eigenvalues are a known function of the parameters θ_i (i.e. $\mathcal{G}^{(i)}(\theta_i)$ has a small degree of freedom), we demonstrate that the maximum-likelihood estimator can be solved to optimality with complexity $\mathcal{O}(\text{card}(\theta)(n \log n + p^2 n + p^6))$, where $\theta = \begin{pmatrix} \theta_1' & \theta_2' & \dots & \theta_m' \end{pmatrix}'$. We demonstrate that the assumption on the eigenvalues and eigenvectors are satisfied by the Whittle Matèrn process, as well as other sensible operators to encode observation dependency structure.

As a second contribution, we provide extensive demonstrations that corroborate the utility of the modeling framework (5.1) and the proposed estimator. Specifically, we apply the procedure to model hyperspectral imaging datasets obtained from AVIRIS-NG pushbroom instruments to obtain a statistical model that is more faithful model as compared to previous techniques, and lead to improved signature detection capabilities.

5.2 Maximum-Likelihood Estimator for Parameter Identification

In this section, we provide an efficient estimator to identify the parameters in the model 5.1. To motivate the need for an estimator that explicitly accounts for the dependency structure among the observations (e.g. the operators $\{\mathcal{G}^{(i)}\}_{i=1}^m$), we

prove that the naive approach of treating the data as iid produces a biased estimate of parameters of interest for a simple 1-D Markov chain with a joint precision matrix from factorizable model 5.1.

Proposition 4. *Consider n observations of a scalar normal Gaussian variable obeying an order-1 Markov chain with same conditional variances θ and partial correlation ρ. The joint precision obeys the factorizable model 5.1 with $\mathcal{A} = \frac{1}{\theta}$, $m = 1$ and $\mathcal{G}$ a tri-diagonal toeplitz matrix with parameters 1, and $\rho > 0$. Let $u \in \mathbb{R}^n$ be an observation from the chain and $\hat{\theta}_n$ be the sample variance from the n observations (i.e. treating the observations in the chain as independent). Then,* $\lim_{n\to\infty} \mathbb{E}[\hat{\theta}_n] > \frac{(\rho+1)\theta}{1+\rho-2\rho^2} > \theta$.

The proof of Proposition 4 is presented in Appendix B.1. This simple thought experiment demonstrates the necessity for an estimator that takes into account the rich dependency structure of observations in (5.1).

Our development of a maximum-likelihood estimator is based on the spectrum of the operators $\mathcal{G}^{(i)}$. Specifically, let $\{\phi_k^{(i)}(\theta_i)\}_{k\in\mathbb{Z}_+}$ and $\{\lambda_k^{(i)}(\theta_i)\}_{k\in\mathbb{Z}_+}$ be the orthonormal eigenfunction/eigenvalue pairs for the operator $\mathcal{G}^{(i)}(\theta_i; x_i)$ (we assume that the operators $\mathcal{G}^{(i)}$ are continuous although the discussion can be easily specialized to the discrete setting). For notational convenience, we let

$$\mathcal{G}(\theta; \mathbf{x}) = \otimes_{i=1}^m \mathcal{G}^{(i)}(\theta_i; x_i),$$

where $\theta = \begin{pmatrix}\theta_1' & \theta_2' & \dots & \theta_m'\end{pmatrix}' \in \mathbb{R}^q$,$q = \text{card}(\theta)$, and $\mathbf{x} = (x_1, x_2, \dots, x_m)$. The orthonormal eigenfunctions and eigenvalues of $\mathcal{G}(\theta; \mathbf{x})$ are denoted by $\{\phi_k\}_{k\in\mathbb{Z}_+}$ and $\{\lambda_k(\theta)\}_{k\in\mathbb{Z}_+}$ respectively. Given data function $u(\mathbf{x}) \in \mathbb{R}^p$ in the field $\otimes_{i=1}^m \mathcal{D}_i$, the maximum likelihood estimator with respect to the precision operator $\mathcal{G}(\theta; \mathbf{x})$ is:

$$\begin{aligned} \min_{\mathcal{A}\in\mathbb{R}^{p\times p},\theta\in\mathbb{R}^q} \quad & \mathsf{J}((\mathcal{A}, \theta); \tilde{\Sigma}(\theta)) + \lambda\, \mathcal{R}(\mathcal{A}), \\ \text{subject-to} \quad & \mathcal{A} \succ 0; \mathcal{G}(\theta) \succ 0 \end{aligned} \tag{5.5}$$

where $\mathsf{J}((\mathcal{A}, \theta); \tilde{\Sigma}(\theta))$ is the negative log-likelihood

$$\begin{aligned} \mathsf{J}((\mathcal{A}, \theta); \tilde{\Sigma}(\theta)) &:= \langle u, \mathcal{P}u\rangle - \log\det\mathcal{P} \\ &= \text{trace}\left(\mathcal{A}\tilde{\Sigma}(\theta)\right) - \log\det\mathcal{P} \end{aligned}$$

and $\tilde{\Sigma}(\theta) \in \mathbb{S}^p$ is the *sufficient statistic* with (s, t) entries

$$[\tilde{\Sigma}(\theta)]_{s,t} = \int_{\otimes_{i=1}^m \mathcal{D}_i} u_s(\mathbf{x})\mathcal{G}(u_t(\mathbf{x}))\mathrm{d}\mathbf{x}.$$

Further, $\mathcal{R}(\mathcal{A})$ represents a regularization function provides a flexible approach to encode the interactions between the variables as a latent-variable model (e.g., nuclear norm penalty). The constraints $\mathcal{A} \succ 0, \mathcal{G}(\theta) \succ 0$ in (5.6) enforce positive-definiteness condition on the precision operator $\mathcal{P}$. To evaluate and optimize (5.6), we consider finite sample approximation of the continuous operator $\mathcal{P}$ given by $\mathcal{P}_n$ that leads to a sample sufficient statistic $\Sigma_n(\theta)$ and a corresponding loss function $\mathsf{J}_n((\mathcal{A};\theta);\Sigma_n(\theta))$. In particular, we assume that we have n_i equidistant observations (spaced apart by distance h_i) of a random vector $u(\mathbf{x}) \in \mathbb{R}^p$ for each dimension $i = 1, 2, \ldots, m$ within a boundary $\mathcal{D}_i$. Letting $n := \prod_i n_i$, the observations can be stacked together as $\{\bar{u}^{(i)}\}_{i=1}^n \subseteq \mathbb{R}^p$. Furthermore, let $\{\phi_{n,k}(\theta)\}_{k=1}^n \subset \mathbb{R}^n$ be n-sample point-mass approximation of $\{\phi_k\}_{k\in\mathbb{Z}^+}$ normalized by $\prod_{i=1}^m h_i$ so that $\{\phi_{n,k}(\theta)\}_{k=1}^n \subset \mathbb{R}^n$ is an orthonormal basis set. Then, $\mathcal{P}_n = \mathcal{G}_n(\theta) \otimes \mathcal{A}$, where $\mathcal{G}_n(\theta) \in \mathbb{S}^n$ is the matrix

$$\mathcal{G}_n(\theta) := \sum_{k=1}^{n} \lambda_k(\theta)[\phi_{n,k}(\theta)][\phi_{n,k}(\theta)]'. \tag{5.6}$$

By construction, $\frac{1}{\prod_{i=1}^m h_i}\{\phi_{n,k}\}_{k=1}^n$ and $\{\lambda_k(\theta)\}_{k=1}^n$ are eigenvector/eigenvalue pair for the matrix $\mathcal{G}_n(\theta)$. Due to the Kronecker product form of $\mathcal{P}_n$, the two quantities of $\mathsf{J}_n(\mathcal{A};\theta)$ can be simplified to

$$\langle u, \mathcal{P}_n u\rangle = \text{trace}\left(\mathcal{A}\tilde{\Sigma}_n(\theta)\right) \;;\; \log\det(\mathcal{P}_n) = n\log\det(\mathcal{A}) + p\sum_{k=1}^{n}\log(\lambda_k(\theta)).$$

where $\tilde{\Sigma}_n(\theta)$ is the *sufficient statistic* $\tilde{\Sigma}_n(\theta) := \sum_{i,j=1}^n [\mathcal{G}_n(\theta)]_{i,j}[u^{(i)}][u^{(j)}]'$. As a result, the loss function $\mathsf{J}_n((\mathcal{A};\theta);\tilde{\Sigma}_n(\theta))$ takes the form:

$$\mathsf{J}_n((\mathcal{A};\theta);\tilde{\Sigma}_n(\theta)) = \text{trace}\left(\mathcal{A}\tilde{\Sigma}_n(\theta)\right) - n\log\det(\mathcal{A}) - p\sum_{k=1}^{n}\log(\lambda_k(\theta). \tag{5.7}$$

Putting everything together, we obtain the following discrete approximation to the continuum maximum-likelihood estimator (5.6)

$$\begin{aligned}(\hat{\mathcal{A}},\hat{\theta}) = \underset{\mathcal{A}\in\mathbb{R}^{p\times p},\theta\in\mathbb{R}^q}{\text{argmin}} \quad & \mathsf{J}_n((\mathcal{A},\theta);\tilde{\Sigma}_n(\theta)) + \lambda\,\mathcal{R}(\mathcal{A}). \\ \text{subject-to} \quad & \mathcal{A} \succ 0; \lambda_k(\theta) > 0 \text{ for } k = 1, 2, \ldots, n\end{aligned} \tag{5.8}$$

For any feasible θ of (5.9), $\mathcal{G}_n(\theta)$ is positive definite and subsequently $\tilde{\Sigma}_n \succeq 0$. For small q, (5.9) can be efficiently solved to optimality. Specifically, consider the following convex optimization problem for a fixed $\theta \in \mathbb{R}^q$ that satisfy the constraint

(5.9):

$$\begin{aligned} \hat{\mathcal{A}}(\theta) = \underset{\mathcal{A}\in\mathbb{R}^{p\times p}}{\text{argmin}} \quad & \text{trace}\left(\mathcal{A}\tilde{\Sigma}_n(\theta)\right) - \log\det(\mathcal{A}) + \lambda\,\mathcal{R}(\mathcal{A}) \\ \text{subject-to} \quad & \mathcal{A} \succ 0. \end{aligned} \tag{5.9}$$

Then, the optimal $\hat{\theta}$ is given by

$$\begin{aligned} \hat{\theta} = \underset{\theta\in\mathbb{R}^q}{\text{argmin}} \quad & \mathsf{J}_n((\mathcal{A}(\theta),\theta);\Sigma_n(\theta)) + \lambda\,\mathcal{R}(\mathcal{A}(\theta)) \\ \text{subject-to} \quad & \lambda_k(\theta) \succ 0 \text{ for } k = 1,2,\ldots,n, \end{aligned}$$

and subsequently $\hat{\mathcal{A}} = \mathcal{A}(\hat{\theta})$. Solving (5.9) requires sweeping over all possible values of θ, which is tractable when the number of free parameters q in θ is small. Given a particular choice of θ, $\mathcal{G}_n(\theta)$ requires $O(n^3)$ computations and thus computing the sufficient statistic $\tilde{\Sigma}_n(\theta)$ requires $O(n^3)$ computations. Assuming coarse gridding over the values of θ, the overall complexity of the estimator is $O(\#\text{ grid points}(n^3 + n^2p^2 + p^6))$.

The computational complexity of computing the sufficient statistic $\tilde{\Sigma}_n(\theta)$ is restrictive in the large sample settings. We consider the following two assumptions that would imply $\tilde{\Sigma}_n(\theta)$ can be computed in $O(\#\text{grid points}p^2n + pn^2)$.

> *Assumption 1:* The eigenfunctions $\mathcal{G}(\theta)$ are known and independent of θ
>
> *Assumption 2:* The eigenvalues of $\mathcal{G}(\theta)$ are given by $f_k(\theta)$ where the functions $f_k : \mathbb{R}^q \to \mathbb{R}$ are known.

Under Assumptions 1 & 2, the entries of $\tilde{\Sigma}_n(\theta)$ can be computed in the following manner. Consider the s-th coordinate of the data $\bar{u}_s = \begin{pmatrix} u_s^{(1)} & u_s^{(2)} & \ldots & u_s^{(n)} \end{pmatrix}'$. Since $\{\phi_{n,k}\}_{k=1}^n$ form an orthonormal basis for $\mathbb{R}^n$, $\bar{u}_s = \sum_{k=1}^n \phi_{n,k}u_{s,k}$ where the coefficients $u_{s,k}$ can be computed in $O(n^2)$ computations. Then, expanding the data points with respect to the basis set $\{\phi_{n,k}\}_{k=1}^n$ yields $\tilde{\Sigma}_{s,t} = \sum_{k=1}^n f_k(\theta)u_{s,k}u_{t,k}$. Evidently, the sufficient statistic $\tilde{\Sigma}_n(\theta)$ can be computed in $O(\#\text{grid points}p^2n + pn^2)$ computations for overall complexity $O(\#\text{ grid points}(np^2 + p^6) + pn^2)$. We present the algorithm in Algorithm 2.

Algorithm 2 Learning Parameters of 5.1

1: **Input**: n Gridded observations of a random vector $u \in \mathbb{R}^p$ across a random field with equidistant values h_i; orthonormal eigenfunctions $\{\phi_k\}_{k\in\mathbb{Z}_+}$; collection of functions $f_k : \mathbb{R}^q \to \mathbb{R}$ for $k \in \mathbb{Z}_+$; a regularization function $\mathcal{R}_\lambda(\cdot)$

2: **Obtaining the Eigenvectors of $\mathcal{G}_n$**: Compute n sample point mass approximation $\{\phi_{n,k}\}_{k=1}^n \mathbb{R}^n$ for $\{\phi_k\}_{k\in\mathbb{Z}_+}$ normalized by $\prod_{i=1}^m h_i$ to obtain an orhonormal basis set. Concatenate data to the form $\{u^{(i)}\}_{i=1}^n \subseteq \mathbb{R}^p$.

3: **Computing Basis Coefficients**: For every $s = 1, 2, \ldots, p$, let $\bar{u}_s = \begin{pmatrix} u_s^{(1)} & u_s^{(2)} & \ldots & u_s^{(n)} \end{pmatrix}$. Compute the coefficients $u_{s,k} = \langle \bar{u}_s, \phi_{n,k} \rangle$ for $k = 1, 2, \ldots, n$.

4: **Sweep Over** θ: Consider a grid for range of values of θ that ensure $f_k(\theta) > 0$ for all k. For every element θ in the gridded set, perform steps

   (a) Compute $\tilde{\Sigma}_n(\theta)$ with coefficients $[\tilde{\Sigma}_n(\theta)]_{s,t} = \sum_{k=1}^n f_k(\theta) u_{s,k} u_{t,k}$

   (b) Solve the convex optimization problem (5.10) for optimal $\mathcal{A}(\theta)$

   (c) Compute the objective function $\mathsf{J}_n((\mathcal{A}(\theta), \theta), \Sigma_n(\theta))$

5: **Output**: Set $\hat{\theta} = \theta$ for θ that achieves the smallest $\mathsf{J}_n((\mathcal{A}(\theta), \theta); \Sigma_n(\theta))$ and subsequently set $\hat{\mathcal{A}} = \mathcal{A}(\hat{\theta})$

A few remarks in order:

Remark 1: Sufficiency The sufficient statistic of the maximum likelihood estimator (5.9) is $\tilde{\Sigma}_n(\theta)$. When $\mathcal{G}(\theta; \mathbf{x})$ and consequently $\mathcal{G}_n(\theta)$ are restricted to be the identity operator, this sufficient statistic equals the sample covariance matrix. In other words, in our settings where the data may not be iid, the sample covariance operator is modified in a suitable manner to account for the dependencies among observations.

Remark 2: When are Assumptions 1 & 2 Satisfied? Below we describe three types of operators $\mathcal{G}_i(\theta_i)$ (which form the operator $\mathcal{G}(\theta)$) that are sensible to employ in practice and satisfy Assumptions 1 & 2:

$$\mathcal{G}_{\mathtt{tri}}(\cdot) = \begin{pmatrix} a & b & 0 & \ldots \\ b & a & b & 0 \\ & \ddots & \ddots & \ddots \\ 0 & \ldots & b & a \end{pmatrix} \quad ; \quad \mathcal{G}_{\mathtt{cyc}}(\cdot) = \begin{pmatrix} a & b & 0 & \ldots & b \\ b & a & b & 0 & \ldots \\ \vdots & \ddots & \ddots & \ddots & \\ 0 & \ldots & b & a & b \\ b & 0 & \ldots & b & a \end{pmatrix}$$

$$\mathcal{G}_{\mathtt{WM}}(\cdot) \quad = \quad \ell^{-1}(\mathcal{I} - \Delta)^{\nu+1/2},$$

where the tri-diagonal matrix $\mathcal{G}_{\texttt{tri}}(\cdot) \in \mathbb{S}^n$ and cyclic matrix $\mathcal{G}_{\texttt{cyc}}(\cdot) \in \mathbb{S}^n$ are a function of scalars (a, b) with eigenvalues and eigenvectors:

$$\mathcal{G}_{\texttt{tri}} \;:\; \phi_k = \sqrt{\frac{2}{n}}\left(\sin(\pi k[1:n]/(n+1))\right) \;;\; \lambda_k = a - 2b\cos(\pi k/(n+1))$$
$$\mathcal{G}_{\texttt{cyc}} \;:\; \phi_k = \frac{1}{\sqrt{n}}(1, w_k, w_k^2, \ldots, w_k^{n-1}) \text{ for } w_k = e^{2i\pi k/n} \;;\; \lambda_k = a + 2b\cos(2\pi k/n).$$

Evidently, the eigenvectors of both matrices are known and independent of (a, b) and the eigenvalues are known functions of (a, b). The Whittle Matérn precision operator $\mathcal{G}_{\texttt{WM}}(\cdot)$ is a function of the length scale ℓ and smoothing parameter ν. The eigenvalues and eigenfunctions of this operator on a domain $[0, L]$ with Dirichlet boundary conditions on the Laplacian are

$$\mathcal{G}_{\texttt{WM}} \;:\; \phi_k = \sqrt{\frac{2}{L}}\sin\left(\frac{k\pi x}{L}\right) \;;\; \lambda_k = \frac{1}{\ell}\left(1 + \ell^2 k^2 \frac{\pi^2}{L^2}\right)^{\nu+1/2}.$$

Once again, the eigenfunctions ϕ_k are known and independent of (ℓ, ν) and the eigenvalues are known functions of (ℓ, ν).

Remark 3: Speedup Using FFT The eigenfunctions of certain classes of operators $\mathcal{G}^{(i)}$ are sinusoidal functions (e.g. the operators $\mathcal{G}_{\texttt{wm}}, \mathcal{G}_{\texttt{tri}}, \mathcal{G}_{\texttt{cyc}}$). In such settings, the basis coefficients $\{u_{s,k}\}_{k=1}^n$ in step 3 of Algorithm 2 can be computed in $O(pn\log(n))$ for on overall complexity of $O(\#\text{ grid points } (np^2 + p^6) + n\log(n)p^2)$.

Remark 4: Alleviating Boundary Condition Effects Some natural choices of operators $\mathcal{G}^{(i)}$ place boundary conditions on the field $\mathcal{D}_i$. For example, we impose Dirichlet boundary conditions on the Whittle Matèrn operator which implies that the observations sampled from the distribution with precision operator $\mathcal{G}^{(i)}$ inside the boundary $\mathcal{D}_i$ are conditionally independent of observations obtained outside the boundary. To ameliorate the boundary effect, one can define the the covariance operator on a larger domain than that where observations are made, and find the parameters of the model (5.1) with this configuration as follows. Compute $\mathcal{G}_{n,\text{restrict}} = \left(\mathcal{G}_n(\theta)^{-1}|_{\text{restrict}}\right)^{-1}$ where the restriction is the based on the domain of the data in relation to the entire field. Then the sufficient statistic becomes $\tilde{\Sigma}_n(\theta) := \sum_{i,j}^n [\mathcal{G}_{n,\text{restrict}}]_{i,j}[\bar{u}^{(i)}][\bar{u}^{(j)}]'$. Evidently, the computational cost of solving (5.9) in this setting increases to $O(\#\text{ grid points } (n^3 + n^2p^2 + p^6)$.

5.3 Real experiments with Hyperspectral Imaging

An outstanding challenge in hyperspectral imaging is accurate detection of signature materials from spectral data. The accuracy of existing detection techniques

(e.g., matched filters) largely depends on a faithful reconstruction of the statistics of the scene such as the underlying covariance matrix. As described in Section 5.1, a modeling challenge that arises is that the data exhibits strong dependencies due to the natural continuity of the scene. Existing techniques (see [Tho+15] and the references therein) ignore these dependencies and treat the data as i.i.d. to find a covariance model underlying data. In this section, we explore the utility of the factorizable precision modeling class (5.1) in signature detection for hyperspectral imaging problems. Specifically, since the noise characteristics of each sensor is different, we consider a separate model from the class (5.1) for each sensor. Furthermore, since the dependency structure is 1-dimensional (direction of travel), we model the underlying precision $\mathcal{P}$ as:

$$\mathcal{P} = \mathcal{A} \otimes \mathcal{G}(\theta; x), \tag{5.10}$$

where $G(\theta; x)$ is the 1-dimensional Whittle-Matèrn precision operator $\frac{1}{\ell}\left(\mathcal{I} - \ell^2\Delta\right)^{\nu+1/2}$. As a result, $\theta = (\ell, \nu)$ consists of the length scale of the process and a smoothing parameter. The matrix $\mathcal{A} \in \mathbb{R}^{p\times p}$ encodes the spectral dependencies across the p channels. The model employed by [Tho+15] is a specialization of (5.10) where $\mathcal{G}(\theta; x)$ is the identity operator. Before proceeding to the experiments, we will briefly describe standard single-pixel matched filter analysis as well as our approach to perform multi-pixel matched filter analysis.

Standard & multi-pixel matched-filter: Let $s \in \mathbb{R}^p$ be signature profile, Σ be the underlying covariance of the scene, and $x \in \mathbb{R}^p$ is a data point for which we would like to test its significance. Standard matched filter has the form [Tur60]:

$$\mathrm{mf}(x) = \frac{x'\Sigma^{-1}s}{\sqrt{s'\Sigma^{-1}s}}.$$

Assuming that the data x is centered, with underlying covariance Σ, i.e. $x \sim \mathcal{N}(0, \Sigma)$, then the matched filter quantity has the statistic $\mathrm{mf} \sim \mathcal{N}(0, 1)$. If, however if the data consists of signature material, i.e. $x \sim \mathcal{N}(\alpha s, \Sigma)$, then $\mathrm{mf} \sim \mathcal{N}(\alpha\sqrt{s'\Sigma^{-1}s}, \Sigma)$. Hence, standard techniques compute matched-filter across the pixels and those with matched-filter that is sufficiently large are declared as containing the signature material. A drawback of this standard approach for matched-filter analysis is that it examines each pixel separately, and as a result, the matched filter value may vary substantially for nearby pixels. This is of course undesirable due to the continuity of a scene. Hence, we develop a multiple-pixel matched-filter statistic. In particular, suppose we consider n nearby pixels with observations $\bar{x} \in \mathbb{R}^{pn}$ and let $\bar{s} \in \mathbb{R}^{pn}$ be

n copies of s, and $\bar{\Sigma} \in \mathbb{S}^{pn}$ be the joint covariance over the multiple pixels. Since the strength of the signature may be varied across the signals, we wish to devise a matched filter that tests for different weights of the signature. This is naturally given by the following optimization problem with respect to unknown coefficients

$$\begin{aligned}
\mathrm{mf}_{\mathrm{multi}}(\bar{x}) &= \max_{d \in \mathbb{R}^n_+ \;;\; D = \begin{pmatrix} d_1 \mathcal{I}_p & 0 & \dots & 0 \\ 0 & d_2 \mathcal{I}_p & \dots & 0 \\ & & \ddots & d_n \mathcal{I}_p \end{pmatrix}} \frac{\bar{x}' \bar{\Sigma}^{-1} \bar{D} s}{(\sqrt{D\bar{s}})' \bar{\Sigma}^{-1} D\bar{s}} \\
&= \max_{d \in \mathbb{R}^n_+ \;;\; D = \begin{pmatrix} d_1 \mathcal{I}_p & 0 & \dots & 0 \\ 0 & d_2 \mathcal{I}_p & \dots & 0 \\ & & \ddots & d_n \mathcal{I}_p \end{pmatrix} \atop \|\bar{\Sigma}^{-1/2} D\bar{s}\|_F \le 1} \langle \bar{\Sigma}^{-1/2} \bar{x}, \bar{\Sigma}^{-1/2} D\bar{s} \rangle, \quad (5.11)
\end{aligned}$$

where the second equality follows from some straightforward manipulations. Solving (5.11) is evidently a convex optimization problem that can be computed efficiently. Furthermore, the optimal value of d in (5.11) computes locations in the window where the signature material is enhanced.

Hyperspectral Imaging dataset In this experiment, we consider reflectance data from the flight 'ang20150420t160719' in a desert area of southern Colorado. The AVIRIS-NG instrument acquires 598 cross-track spectra at 100 Hz. Frames are captured with a custom field programmable gate array (FPGA) frame grabber under dedicated Camera Link interface at 500 Mbs^{-1}. The sensors obtain reflectance data over the spectrum 355.59nm – 2577.08nm with resolution of approximately $9nm$ for a total number of 496 spectral channels. We restrict the channels to the range 900nm – 2400nm for a total number of $p = 291$ channels where methane has large refractivity coefficients. During this flight, 17165 lines (observations across the flight-line) were taken. A snapshot of the scene is shown in Figure 5.2(a). For our subsequent numerical analysis, we consider sensor 202 and take lines $1 : 500$ as training set so that $\mathcal{D}_{\mathrm{train}} = \{x^{(i)}\}_{i=1}^{500} \subseteq \mathbb{R}^p$ and lines $1001 : 1500$ as testing set so that $\mathcal{D}_{\mathrm{test}} = \{x^{(i)}\}_{i=501}^{1000} \subseteq \mathbb{R}^p$. In the experiments with this dataset, we compare the method in [Tho+15] vs the factorizable model (5.10). The method in [Tho+15] treats the data across the lines as i.i.d. without taking into account natural dependencies of the scene, whereas the model (5.10) explicitly accounts for these dependencies (via the operator $\mathcal{G}(\theta; x)$) in a computationally efficient and tractable manner. We evaluate the performance of each approach in methane signature gas detection.

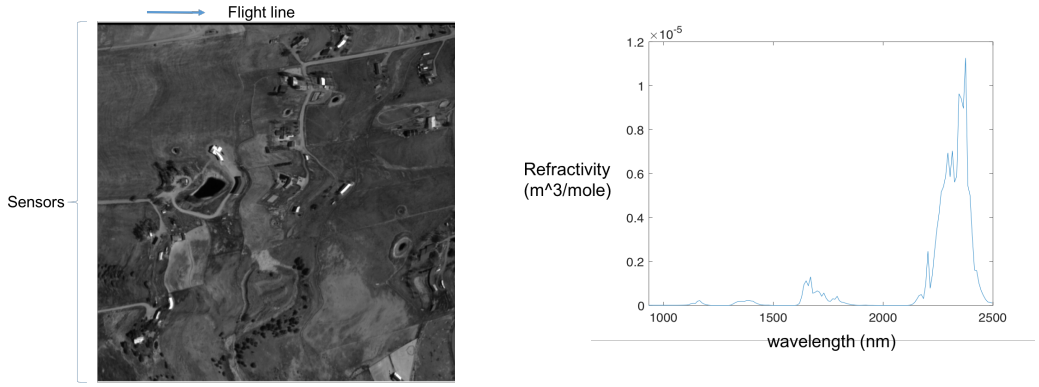


Figure 5.2: left: Snapshot of a scene from the 'ang20150420t160719' flight; right: spectral refractivity co efficient of methane gas across infrared and visible spectrum.

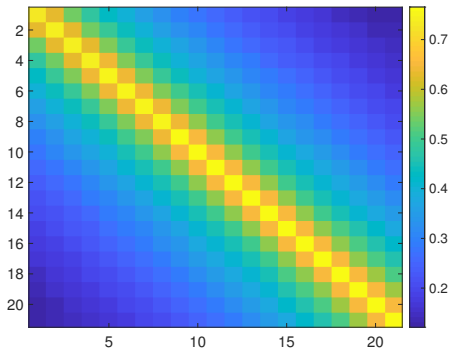


Figure 5.3: Dependency structure of the factorizable precision approach (5.10). Notice a strong correlation across multiple pixels away.

Calibration We begin with obtaining a calibration model, i.e. a covariance model of the scene without a methane signature. The covariance model will then be used on the test dataset to check for the presence of methane via the matched filter analysis (5.11). We posit that the data comes from a Gaussian distribution and use the training data $\mathcal{D}_{\text{train}}$ to obtain two models: one from treating the data i.i.d (e.g. setting $G(\theta) = \mathcal{I}$), and the other by optimizing over the parameters ℓ, ν in (5.10) via the estimator (5.9). Both models yield a precision operator $\mathcal{P}$ and the parameters $\theta = (\ell, \nu)$ of the model (5.10) are identified to be $\ell = 50.03$ and $\nu = 0.73$. The resulting dependency structures encoded by $\mathcal{G}(\theta; x)$ for the model (5.10) is shown in Figure 5.3. The i.i.d model sets $\mathcal{G}(\theta) = \mathcal{I}$ and thus assumes nearby pixels to be statistically independent. On the other hand, the model (5.10) identifies strong dependencies between nearby pixels, e.g. for pixels 6 apart (physical distance 24 meters), the correlation is estimated to be ≈ 0.4.

We further evaluate the training & testing performance of each model based on the

negative log-likelihood metric. In particular, the negative log-likelihood is computed by $\text{trace}(\mathcal{A}\tilde{\Sigma}_n(\theta)) - \log\det(\mathcal{P})$, where $\tilde{\Sigma}_n(\theta)$ is the sufficient statistic of the n data points. For the i.i.d model, the quantity $\tilde{\Sigma}_n(\theta)$ is the sample covariance matrix and for (5.10), this is the modified sample covariance $\tilde{\Sigma}_n(\theta) := \sum_{i,j=1}^{n}[\mathcal{G}_n(\theta)]_{i,j}[x^{(i)}][x^{(j)}]'$ with $\mathcal{G}_n(\theta)$ given in (5.6). Table 5.3 summarizes the training and testing performance of both models, demonstrating that our model outperforms the i.i.d. model employed in [Tho+15]. Evidently, accounting for the spatial dependencies leads to a superior model of the hyperspectral imaging data.

Model	Training performance	Validation performance	# parameters
i.i.d model	$-9.24x10^3$	$-5.04x10^3$	42486
Our approach	$-1.34x10^4$	$-1.48x10^3$	42488

Table 5.1: Training and validation performances of i.i.d. model employed in [Tho+15] and our proposed model (5.10).

Methane Detection In this section, we demonstrated that our approach (5.10) yields a more representative model of the data. In this analysis, we explore how this improvement translates to better detection of methane gas. To that end, we add an artificial methane signature to a pre-specified location in the testing data. Specifically, for the pixel locations $1201-1250$ in the test data $\mathcal{D}_{\text{test}}$, we add a random multiple of the methane signature, e.g. $\tilde{x}^{(i)} = x^{(i)} + \alpha_i t$ for $i = 1201, \ldots, 1250$, where the collection $\{\alpha_i\} > 0$ control the magnitude of methane at different locations and t is the methane signature. To add the methane signature in a manner that is continuous over neighboring pixels, the collection of $\{\alpha\}_{i=1}^{50}$ is sampled from a joint Gaussian distribution with a Toeplitz tridiagonal precision matrix with diagonal element equaling to 10^4, and nonzero off-diagonal elements equaling $5x10^4$. We apply the multi-pixel matched filter analysis (5.11) on the test data $\mathcal{D}_{\text{test}}$ with window size $w = 6$. The matched-filter output of the i.i.d. approach is shown in blue in Figure 5.4(a) and the matched-filter output of our approach is shown in red. As expected, both models produce a high matched-filter result in the region $1201-1250$. To test the significance of these high values, we also produce matched-filter results of both models on the methane-free training data in Figure 5.4(b), which show substantially smaller values. An interesting observation however is that, on the testing data that contains methane, our approach shows consistently high and rather uniform matched-filter values. On the other hand, the i.i.d approach fluctuates

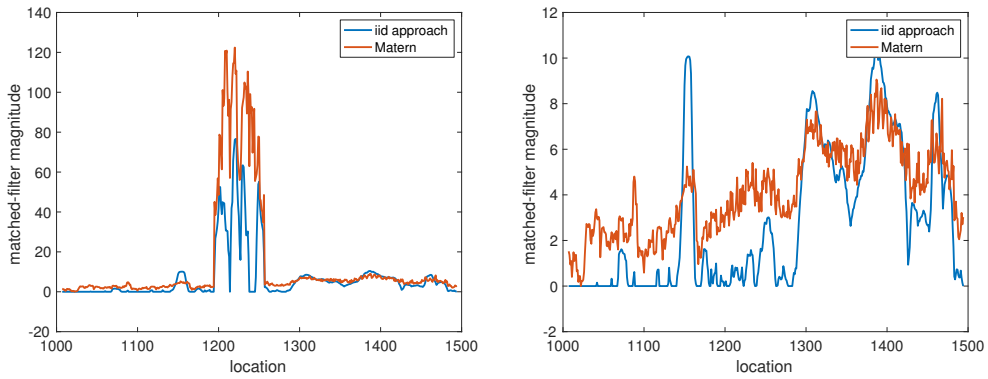


Figure 5.4: left: matched-filter outputs of our modeling approach (5.10) and i.i.d approach on the test dataset; right: matched-filter outputs of our modeling approach (5.10) and i.i.d approach on the training dataset.

significantly, with nearby pixel locations having significantly different matched-filter values. Evidently, the proposed model (5.1) yields more accurate methane detection that the iid approach.

5.4 Discussion

In this chapter, we introduced *factorizable precision operators* that provide a conceptually and computationally appealing framework to account for the complex dependencies among a large number of variables in the system as well as the dependencies among observations of these variables (e.g. non-iid and high-dimensional). The parameters of this model can be efficiently solved with complexity $O(pn^2 + p^2n + p^6)$ via a maximum-likelihood estimator. We apply our proposed approach to hysperspectral imaging datasets to obtain more representative covariance models than previous techniques, which also lead to more accurate signature detection. There are several interesting avenues for further investigation that arise from our work. First, the factorizable precision operator is relevant beyond the hyperspectral imaging setting. For example, our framework is naturally amenable to spatiotemporal data where the time and space components can be decoupled. Second, the Matérn operator that was employed as one of the components in the factorizable precision operator, has a single length scale corresponding to all the variables. In hyperspectral imaging, this may be too stringent of a modeling restriction since the channels are likely to have a different natural length scale. Finally, the circularity of time resolutions (month and days as an example) adds a modeling challenge that our framework is not able to address. It is of interest to build on this framework to incorporate cyclo-stationary processes (e.g. statistical processes that

repeat), so that the methodology could be useful for reservoir modeling as well as other applications that involve seasonal patterns.

Chapter 6

INTERPRETING LATENT VARIABLES VIA CONVEX OPTIMIZATION

In Chapters 2, 4, and 5, we developed latent-variable models to encode the effect of unmodeled phenomena on the observed system. Naturally, the following question arises: how do we attribute semantics to these latent variables? This chapter addresses this question with a rigorous methodological contribution that is supported by theory and experiments.

The results in this chapter were published in [TC18] and was developed jointly with Venkat Chandrasekaran. The author contributed by developing the modeling framework and the associated algorithm to estimate model parameters, providing theoretical support of the algorithm, and producing numerical experiments. The description of the work contained in this chapter was written by the author and Venkat Chandrasekaran.

6.1 Introduction

A central goal in data analysis is to identify concisely described models that characterize the statistical dependencies among a collection of variables. Such concisely parametrized models avoid problems associated with overfitting, and they are often useful in providing meaningful interpretations of the relationships inherent in the underlying variables. Latent or unobserved phenomena complicate the task of determining concisely parametrized models as they induce confounding dependencies among the observed variables that are not easily or succinctly described. Consequently, significant efforts over many decades have been directed towards the problem of accounting for the effects of latent phenomena in statistical modeling. A common shortcoming of approaches to latent-variable modeling is that the latent variables are typically mathematical constructs that are derived from the originally observed data, and these variables do not directly have semantic information linked to them. Discovering interpretable meaning underlying latent variables would clearly impact a range of contemporary problem domains throughout science and technology. For example, in data-driven approaches to scientific discovery, the association of semantics to latent variables would lead to the identification of new phenomena that are relevant to a scientific process, or would guide data-gathering

exercises by providing choices of variables for which to obtain new measurements.

In this paper, we focus for the sake of concreteness on the challenge of interpreting the latent variables in a factor model [Spe04]. *Factor analysis* is perhaps the most widely used latent-variable modeling technique in practice. The objective with this method is to fit observations of a collection of random variables $y \in \mathbb{R}^p$ to the following linear model:

$$y = \mathcal{B}\zeta + \epsilon, \tag{6.1}$$

where $\mathcal{B} \in \mathbb{R}^{p\times k}, k \ll p$. The random vectors $\zeta \in \mathbb{R}^k, \epsilon \in \mathbb{R}^p$ are independent of each other, and they are normally distributed as[1] $\zeta \sim \mathcal{N}(0, \Sigma_\zeta), \epsilon \sim \mathcal{N}(0, \Sigma_\epsilon)$, with $\Sigma_\zeta \succ 0, \Sigma_\epsilon \succ 0$ and Σ_ϵ being diagonal. Here the random vector ζ represents a small number of unobserved, latent variables that impact all the observed variables y, and the matrix $\mathcal{B}$ specifies the effect that the latent variables have on the observed variables. However, the latent variables ζ themselves do not have any interpretable meaning, and they are essentially a mathematical abstraction employed to fit a concisely parameterized model to the conditional distribution of $y|\zeta$ (which represents the remaining uncertainty in y after accounting for the effects of the latent variables ζ); this conditional distribution is succinctly described as it is specified by a model consisting of independent variables (as the covariance of the Gaussian random vector ϵ is diagonal).

A natural approach to attributing semantic information to the latent variables ζ in a factor model is to obtain measurements of some additional plausibly useful covariates $x \in \mathbb{R}^q$ (the choice of these variables is domain-specific), and to link these to the variables ζ. However, defining and specifying such a link in a precise manner is challenging. Indeed, a fundamental difficulty that arises in establishing this association is that the variables ζ in the factor model (6.1) are not identifiable. In particular, for any non-singular matrix $\mathcal{W} \in \mathbb{R}^{k\times k}$, we have that $\mathcal{B}\zeta = (\mathcal{B}\mathcal{W}^{-1})(\mathcal{W}\zeta)$. In this paper, we describe a systematic and computationally tractable methodology based on convex optimization that integrates factor analysis and the task of interpreting the latent variables. Our convex relaxation approach generalizes the *minimum-trace factor analysis* technique, which has received much attention in the mathematical programming community over the years [Led40; Sha82a; Sha82b; Sha04; Sau+12].

[1] The mean vector does not play a significant role in our development, and therefore we consider zero-mean random variables throughout this paper.

A Composite Factor Model

We begin by making the observation that the column space of $\mathcal{B}$ — which specifies the k-dimensional component of y that is influenced by the latent variables ζ — is invariant under transformations of the form $\mathcal{B} \to \mathcal{B}\mathcal{W}^{-1}$ for non-singular matrices $\mathcal{W} \in \mathbb{R}^{k\times k}$. Consequently, we approach the problem of associating the covariates x to the latent variables ζ by linking the effects of x on y to the column space of $\mathcal{B}$. Conceptually, we seek a decomposition of the column space of $\mathcal{B}$ into transverse subspaces $\mathfrak{H}_x, \mathfrak{H}_u \subset \mathbb{R}^p$, $\mathfrak{H}_x \cap \mathfrak{H}_u = \{0\}$ so that column-space$(\mathcal{B}) \approx \mathfrak{H}_x \oplus \mathfrak{H}_u$ — the subspace $\mathfrak{H}_x$ specifies those components of y that are influenced by the latent variables ζ and are also affected by the covariates x, and the subspace $\mathfrak{H}_u$ represents any unobserved residual effects on y due to ζ that are not captured by x. To identify such a decomposition of the column space of $\mathcal{B}$, our objective is to split the term $\mathcal{B}\zeta$ in the factor model (6.1) as

$$\mathcal{B}\zeta \approx \mathcal{A}x + \mathcal{B}_u\zeta_u, \tag{6.2}$$

where the column space of $\mathcal{A} \in \mathbb{R}^{p\times q}$ is the subspace $\mathfrak{H}_x$ and the column space of $\mathcal{B}_u \in \mathbb{R}^{p\times \dim(\mathfrak{H}_u)}$ is the subspace $\mathfrak{H}_u$, i.e., dim(column-space($\mathcal{A}$))+dim(column-space($\mathcal{B}_u$)) = dim(column-space($\mathcal{B}$)) and column-space($\mathcal{A}$) $\cap$ column-space($\mathcal{B}_u$) = $\{0\}$. Since the number of latent variables ζ in the factor model (6.1) is typically much smaller than p, the dimension of the column space of $\mathcal{A}$ is also much smaller than p; as a result, if the dimension q of the additional covariates x is large, the matrix $\mathcal{A}$ has small rank. Hence, the matrix $\mathcal{A}$ plays two important roles: its column space (in $\mathbb{R}^p$) identifies those components of the subspace $\mathcal{B}$ that are influenced by the covariates x, and its rowspace (in $\mathbb{R}^q$) specifies those components of (a potentially large number of) the covariates x that influence y. Thus, *the projection of the covariates x onto the rowspace of $\mathcal{A}$ represents the interpretable component of the latent variables ζ*. The term $\mathcal{B}_u\zeta_u$ in (6.2) represents, in some sense, the effects of those phenomena that continue to remain unobserved despite the incorporation of the covariates x.

Motivated by this discussion, we fit observations of $(y, x) \in \mathbb{R}^p \times \mathbb{R}^q$ to the following *composite factor model* that incorporates the effects of the covariates x as well as of additional unobserved latent phenomena on y:

$$y = \mathcal{A}x + \mathcal{B}_u\zeta_u + \bar{\epsilon}, \tag{6.3}$$

where $\mathcal{A} \in \mathbb{R}^{p\times q}$ with rank$(\mathcal{A}) \ll \min\{p, q\}$, $\mathcal{B}_u \in \mathbb{R}^{p\times k_u}$ with $k_u \ll p$, and the variables $\zeta_u, \bar{\epsilon}$ are independent of each other (and of x) and normally distributed as $\zeta_u \sim \mathcal{N}(0, \Sigma_{\zeta_u}), \bar{\epsilon} \sim \mathcal{N}(0, \Sigma_{\bar{\epsilon}})$, with $\Sigma_{\zeta_u} \succ 0, \Sigma_{\bar{\epsilon}} \succ 0$ and $\Sigma_{\bar{\epsilon}}$ being a diagonal matrix.

The matrix $\mathcal{A}$ may also be viewed as the map specifying the best linear estimate of y based on x. In other words, the goal is to identify a low-rank matrix $\mathcal{A}$ such that the conditional distribution of $y|x$ (and equivalently of $y|\mathcal{A}x$) is specified by a standard factor model of the form (6.1).

Composite Factor Modeling via Convex Optimization

Next we describe techniques to fit observations of $y \in \mathbb{R}^p$ to the model (6.3). This method is a key subroutine in our algorithmic approach for associating semantics to the latent variables in a factor model (see Section 1.3 for a high-level discussion of our approach and Section 6.3 for a more detailed experimental demonstration). Fitting observations of $(y, x) \in \mathbb{R}^p \times \mathbb{R}^q$ to the composite factor model (6.3) is accomplished by identifying a Gaussian model over (y, x) with the covariance matrix of the model satisfying certain algebraic properties. For background information on multivariate Gaussian statistical models, we refer the reader to [Kay98].

Examining the factor model in (6.1), the covariance matrix of y is decomposable as the sum of a low-rank matrix $\mathcal{B}\Sigma_\zeta\mathcal{B}'$ (corresponding to the $k \ll p$ latent variables ζ) and a diagonal matrix Σ_ϵ. Based on this algebraic structure, a natural approach to factor modeling is to find the smallest rank (positive semidefinite) matrix such that the difference between this matrix and the empirical covariance of the observations of y is close to being a diagonal matrix (according to some measure of closeness, such as in the Frobenius norm). This problem is computationally intractable to solve in general due to the rank minimization objective [Nat95]. As a result, a common heuristic is to replace the matrix rank by the trace functional, which results in the minimum trace factor analysis problem [Led40; Sha82a; Sha82b; Sha04]; this problem is convex and it can be solved efficiently. The use of the trace of a positive semidefinite matrix as a surrogate for the matrix rank goes back many decades, and this topic has received much renewed interest over the past several years [MP97; Faz02; RFP10; CR09].

In attempting to generalize the minimum-trace factor analysis approach to the composite factor model, one encounters a difficulty that arises due to the parametrization of the underlying Gaussian model in terms of covariance matrices. Specifically, with the additional covariates $x \in \mathbb{R}^q$ in the composite model (6.3), our objective is to identify a Gaussian model over $(y, x) \in \mathbb{R}^p \times \mathbb{R}^q$ with the joint covariance $\Sigma = \begin{pmatrix} \Sigma_y & \Sigma_{yx} \\ \Sigma'_{yx} & \Sigma_x \end{pmatrix} \in \mathbb{S}^{p+q}$ satisfying certain structural properties. One of these properties is that the conditional distribution of $y|x$ is specified by a factor model, which

implies that the conditional covariance of $y|x$ must be decomposable as the sum of a low-rank matrix and a diagonal matrix. However, this conditional covariance is given by the Schur complement $\Sigma_y - \Sigma_{yx}\Sigma_x^{-1}\Sigma_{yx}'$, and specifying a constraint on the conditional covariance matrix in terms of the joint covariance matrix Σ presents an obstacle to obtaining computationally tractable optimization formulations.

A more convenient approach to parameterizing conditional distributions in Gaussian models is to consider models specified in terms of inverse covariance matrices, which are also called *precision matrices*. Specifically, the algebraic properties that we desire in the joint covariance matrix Σ of (y, x) in a composite factor model can also be stated in terms of the joint precision matrix $\Theta = \Sigma^{-1}$ via conditions on the submatrices of $\Theta = \begin{pmatrix} \Theta_y & \Theta_{yx} \\ \Theta_{yx}' & \Theta_x \end{pmatrix}$. First, the precision matrix of the conditional distribution of $y|x$ is specified by the submatrix Θ_y; as the covariance matrix of the conditional distribution of $y|x$ is the sum of a diagonal matrix and a low-rank matrix, the Woodbury matrix identity [2] implies that the submatrix Θ_y is the difference of a diagonal matrix and a low-rank matrix. Second, the rank of the submatrix $\Theta_{yx} \in \mathbb{R}^{p\times q}$ is equal to the rank of $\mathcal{A} \in \mathbb{R}^{p\times q}$ in non-degenerate models (i.e., if $\Sigma \succ 0$) because the relation between $\mathcal{A}$ and Θ is given by $\mathcal{A} = -[\Theta_y]^{-1}\Theta_{yx}$. Based on this algebraic structure desired in Θ, we propose the following natural convex relaxation for fitting a collection of observations $\mathcal{D}_n^+ = \{(y^{(i)}, x^{(i)})\}_{i=1}^n \subset \mathbb{R}^{p+q}$ to the composite model (6.3):

$$
\begin{aligned}
(\hat{\Theta}, \hat{D}_y, \hat{L}_y) = \arg \min_{\substack{\Theta\in\mathbb{S}^{p+q},\ \Theta\succ 0 \\ D_y, L_y \in \mathbb{S}^p}} \quad & -\ell(\Theta; \mathcal{D}_+^n) + \lambda_n[\gamma\|\Theta_{yx}\|_\star + \operatorname{trace}(L_y)]. \\
\text{s.t.} \quad & \Theta_y = D_y - L_y,\ L_y \succeq 0, D_y \text{ is diagonal} \qquad (6.4)
\end{aligned}
$$

The term $\ell(\Theta; \mathcal{D}_+^n)$ is the Gaussian log-likelihood function that enforces fidelity to the data, and it is given as follows (up to some additive and multiplicative terms):

$$
\ell(\Theta; \mathcal{D}_+^n) = \log\det(\Theta) - \operatorname{trace}\left[\Theta \cdot \tfrac{1}{n}\sum_{i=1}^n \begin{pmatrix} y^{(i)} \\ x^{(i)} \end{pmatrix}\begin{pmatrix} y^{(i)} \\ x^{(i)} \end{pmatrix}'\right]. \qquad (6.5)
$$

This function is concave as a function of the joint precision matrix [3] Θ. The matrices D_y, L_y represent the diagonal and low-rank components of Θ_y. As with the idea

[2] Recall that the woodbury identity states that $(A+UCV')^{-1} = A^{-1} - A^{-1}U(C^{-1}+VA^{-1}U)^{-1}VA^{-1}$ for matrices A, U, V, C of appropriate dimensions.

[3] An additional virtue of parameterizing our problem in terms of precision matrices rather than in terms of covariance matrices is that the log-likelihood function in Gaussian models is not concave over the cone of positive semidefinite matrices when viewed as a function of the covariance matrix.

behind minimum-trace factor analysis, the role of the trace norm penalty on L_y is to induce low-rank structure in this matrix. Based on a more recent line of work originating with the thesis of Fazel [Faz02; RFP10; CR09], the nuclear norm penalty $\|\Theta_{yx}\|_\star$ on the submatrix Θ_{yx} (which is in general a non-square matrix) is useful for promoting low-rank structure in that submatrix of Θ. The parameter γ provides a tradeoff between the observed/interpretable and the unobserved parts of the composite factor model (6.3), and the parameter λ_n provides a tradeoff between the fidelity of the model to the data and the overall complexity of the model (the total number of observed and unobserved components in the composite model (6.3)). In summary, for $\lambda_n, \gamma \geq 0$ the regularized maximum-likelihood problem (6.4) is a convex program. From the optimal solution $(\hat{\Theta}, \hat{D}_y, \hat{L}_y)$ of (6.4), we can obtain estimates for the parameters of the composite factor model (6.3) as follows:

$$\begin{aligned} \hat{\mathcal{A}} &= -[\hat{\Theta}_y]^{-1}\hat{\Theta}_{yx} \\ \hat{\mathcal{B}}_u &= \text{any squareroot of } (\hat{D}_y - \hat{L}_y)^{-1} - \hat{D}_y^{-1} \text{ such that } \hat{\mathcal{B}}_u \in \mathbb{R}^{p \times \text{rank}(\hat{L}_y)}, \end{aligned} \tag{6.6}$$

with the covariance of ζ_u being the identity matrix of appropriate dimensions and the covariance of $\bar{\epsilon}$ being $\hat{D}_y^{-1}$. The convex program (6.4) is log-determinant semidefinite programs that can be solved efficiently using existing numerical solvers such as the LogDetPPA package [TTT16].

Algorithmic Approach for Interpreting Latent Variables in a Factor Model

Our discussion has led us to a natural (meta-) procedure for interpreting latent variables in a factor model. Suppose that we are given a factor model underlying $y \in \mathbb{R}^p$. The analyst proceeds by obtaining simultaneous measurements of the variables y as well as some additional covariates $x \in \mathbb{R}^q$ of plausibly relevant phenomena. Based on these joint observations, we identify a suitable composite factor model (6.3) via the convex program (6.4). In particular, we sweep over the parameters λ_n, γ in (6.4) to identify composite models that achieve a suitable decomposition — in terms of effects attributable to the additional covariates x and of effects corresponding to remaining unobserved phenomena — of the effects of the latent variables in the factor model given as input.

To make this approach more formal, consider a composite factor model (6.3) $y = \mathcal{A}x + \mathcal{B}_u\zeta_u + \epsilon$ underlying a pair of random vectors $(y, x) \in \mathbb{R}^p \times \mathbb{R}^q$, with $\text{rank}(\mathcal{A}) = k_x$, $\mathcal{B}_u \in \mathbb{R}^{p \times k_u}$, and column-space$(\mathcal{A}) \cap$ column-space$(\mathcal{B}_u) = \{0\}$. As described in Section 6.1, the algebraic aspects of the underlying composite factor

model translate to algebraic properties of submatrices of $\Theta \in \mathbb{S}^{p+q}$. In particular, the submatrix Θ_{yx} has rank equal to k_x and the submatrix Θ_y is decomposable as $D_y - L_y$ with D_y being diagonal and $L_y \succeq 0$ having rank equal to k_u. Finally, the transversality of column-space($\mathcal{A}$) and column-space($\mathcal{B}_u$) translates to the fact that column-space(Θ_{yx}) $\cap$ column-space(L_y) = $\{0\}$ have a transverse intersection. One can simply check that the factor model underlying the random vector $y \in \mathbb{R}^p$ that is induced upon marginalization of x is specified by the precision matrix of y given by $\tilde{\Theta}_y = D_y - [L_y + \Theta_{yx}(\Theta_x)^{-1}\Theta_{xy}]$. Here, the matrix $L_y + \Theta_{yx}(\Theta_x)^{-1}\Theta_{xy}$ is a rank $k_x + k_u$ matrix that captures the effect of latent variables in the factor model. This effect is decomposed into $\Theta_{yx}(\Theta_x)^{-1}\Theta_{xy}$ — a rank k_x matrix representing the component of this effect attributed to x, and L_y — a matrix of rank k_u representing the effect attributed to residual latent variables.

These observations motivate the following algorithmic procedure. Suppose we are given a factor model that specifies the precision matrix of y as the difference $\hat{\tilde{D}}_y - \hat{\tilde{L}}_y$, where $\hat{\tilde{D}}_y$ is diagonal and $\hat{\tilde{L}}_y$ is low rank. Then the composite factor model of (y, x) with estimates $(\hat{\Theta}, \hat{D}_y, \hat{L}_y)$ offers an interpretation of the latent variables of the given factor model if (i) rank($\hat{\tilde{L}}_y$) = rank($\hat{L}_y + \hat{\Theta}_{yx}\hat{\Theta}_x^{-1}\hat{\Theta}_{xy}$), (ii) column-space($\hat{\Theta}_{yx}$) $\cap$ column-space($\hat{L}_y$) = $\{0\}$, and
(iii) $\max\{\|\hat{\tilde{D}}_y - \hat{D}_y\|_2/\|\hat{\tilde{D}}_y\|_2, \|\hat{\tilde{L}}_y - [\hat{L}_y + \hat{\Theta}_{yx}\hat{\Theta}_x^{-1}\hat{\Theta}_{xy}]\|_2/\|\hat{\tilde{L}}_y\|_2\}$ is small. The full algorithmic procedure for attributing meaning to latent variables of a factor model is outlined below:

Algorithm 3 Interpreting Latent Variables in a Factor Model

1: **I**nput: A collection of observations $\mathcal{D}_n^+ = \{(y^{(i)}, x^{(i)})\}_{i=1}^n \subset \mathbb{R}^p \times \mathbb{R}^q$ of the variables y and of some auxiliary covariates x; Factor model with parameters $(\hat{\tilde{D}}_y, \hat{\tilde{L}}_y)$.

2: **C**omposite Factor Modeling: For each $d = 1, \ldots, q$, sweep over parameters (λ_n, γ) in the convex program (6.4) (with $\mathcal{D}_n^+$ as input) to identify composite models with estimates $(\hat{\Theta}, \hat{D}_y, \hat{L}_y)$ that satisfy the following three properties: (i) rank($\hat{\Theta}_{yx}$) = d, (ii) rank($\hat{\tilde{L}}_y$) = rank($\hat{L}_y$) + rank($\hat{\Theta}_{yx}$), and (iii) rank($\hat{\tilde{L}}_y$) = rank($\hat{L}_y$ + rank($\hat{\Theta}_{yx}\hat{\Theta}_x^{-1}\hat{\Theta}_{xy}$)).

3: **I**dentifying Subspace: For each $d = 1, \ldots, q$ and among the candidate composite models (from the previous step), choose the composite factor model that minimizes the quantity $\max\{\|\hat{\tilde{D}}_y - \hat{D}_y\|_2/\|\hat{\tilde{D}}_y\|_2, \|\hat{\tilde{L}}_y - [\hat{L}_y + \hat{\Theta}_{yx}\hat{\Theta}_x^{-1}\hat{\Theta}_{xy}]\|_2/\|\hat{\tilde{L}}_y\|_2\}$.

4: **O**utput: For each $d = 1, \ldots q$, the d-dimensional projection of x into the row-space of $\hat{\Theta}_{yx}$ represents the interpretable component of the latent variables in the factor model.

The effectiveness of Algorithm 1 is dependent on the size of the quantity $\max\{\|\hat{\tilde{D}}_y - \hat{D}_y\|_2/\|\hat{\tilde{D}}_y\|_2, \|\hat{\tilde{L}}_y - \hat{L}_y - \hat{\Theta}_{yx}\hat{\Theta}_x^{-1}\hat{\Theta}_{xy}]\|_2/\|\hat{\tilde{L}}_y\|_2\}$. The smaller this quantity, the better the composite factor model fits to the given factor model. Finally, recall from Section 6.1 that the projection of covariates x onto to the row-space of $\mathcal{A}$ (from the composite model (6.3)) represents the interpretable component of the latent variables of the factor model. Because of the relation $\mathcal{A} = -[\Theta_y]^{-1}\Theta_{yx}$, this interpretable component is obtained by projecting the covariates x onto the row-space of Θ_{yx}. This observation explains the final step of Algorithm 1.

The input to Algorithm 1 is a factor model underlying a collection of variables $y \in \mathbb{R}^p$, and the algorithm proceeds to obtain semantic interpretation of the latent variables of the factor model. However, in many situations, a factor model underlying $y \in \mathbb{R}^p$ may not be available in advance, and must be learned in a data-driven fashion based on observations of $y \in \mathbb{R}^p$. In our experiments (see Section 6.3), we learn a factor model using a specialization of the convex program (6.4). It is reasonable to ask whether one might directly fit to a composite model to the covariates and responses jointly without reference to the underlying factor model based on the responses. However, in our experience with applications, it is often the case that observations of the responses y are much more plentiful than of joint observations of responses y and covariates x. As an example, consider a setting in which the responses are a collection of financial asset prices (such as stock return values); observations of these variables are available at a very fine time-resolution on the order of seconds. On the other hand, some potentially useful covariates such as GDP, government expenditures, federal debt, and consumer rate are available at a much coarser scale (usually on the order of months or quarters). As another example, consider a setting in which the responses are reservoir volumes of California; observations of these variables are available at a daily scale. On the other hand, reasonable covariates that one may wish to associate to the latent variables underlying California reservoir volumes such as agricultural production, crop yield rate, average income, and population growth rate are available at a much coarser time scale (e.g. monthly). In such settings, the analyst can utilize the more abundant set of observations of the responses y to learn an accurate factor model first. Subsequently, one can employ our approach to associate semantics to the latent variables in this factor model based on the potentially limited number of observations of the responses y and the covariates x.

Our Results

In Section 6.2 we carry out a theoretical analysis to investigate whether the framework outlined in Algorithm 1 can succeed. We discuss a model problem setup, which serves as the basis for the main theoretical result in Section 6.2. Suppose we have Gaussian random vectors $(y, x) \in \mathbb{R}^p \times \mathbb{R}^q$ that are related to each other via a composite factor model (6.3). Note that this composite factor model induces a factor model underlying the variables $y \in \mathbb{R}^p$ upon marginalization of the covariates x. In the subsequent discussion, we assume that the factor model that is supplied as input to Algorithm 1 is the factor model underlying the responses y.
Now we consider the following question: Given observations jointly of $(y, x) \in \mathbb{R}^{p+q}$, does the convex relaxation (6.4) (for suitable choices of regularization parameters λ_n, γ) estimate the composite factor model underlying these two random vectors accurately? An affirmative answer to this question demonstrates the success of Algorithm 1. In particular, a positive answer to this question implies that we can decompose the effects of the latent variables in the factor model underlying y using the convex relaxation (6.4), as the accurate estimation of the composite model underlying (y, x) implies a successful decomposition of the effects of the latent variables in the factor model underlying y. That is, steps 2-3 in the Algorithm are successful. In Section 6.2, we show that under suitable identifiability conditions on the population model of the joint random vector (y, x), the convex program (6.4) succeeds in solving this question. Our analysis is carried out in a high-dimensional asymptotic scaling regime in which the dimensions p, q, the number of observations n, and other model parameters may all grow simultaneously [BG11; Wai14].

We give concrete demonstration of Algorithm 1 with experiments on synthetic data and real-world financial data. For the financial asset problem, we consider as our variables y the monthly averaged stock prices of 45 companies from the Standard and Poor index over the period March 1982 to March 2016, and we identify a factor model (6.1) over y with 10 latent variables (the approach we use to fit a factor model is described in Section 6.3). We then obtain observations of $q = 13$ covariates on quantities related to oil trade, GDP, government expenditures, etc. (See Section 6.3 for the full list), as these plausibly influence stock returns. Following the steps outlined in Algorithm 1, we use the convex program (6.4) to identify a two-dimensional projection of these 13 covariates that represent an interpretable component of the 10 latent variables in the factor model, as well as a remaining set of 8 latent variables that constitute phenomena not observed via the covariates x. In

further analyzing the characteristics of the two-dimensional projection, we find that EUR to USD exchange rate and government expenditures are the most relevant of the 13 covariates considered in our experiment, while mortgage rate and oil imports are less useful. See Section 6.3 for complete details.

Related Work

Elements of our approach bear some similarity with *canonical correlations analysis* [Hot02], which is a classical technique for identifying relationships between two sets of variables. In particular, for a pair of jointly Gaussian random vectors $(y, x) \in \mathbb{R}^{p\times q}$, canonical correlations analysis may be used as a technique for identifying the most relevant component(s) of x that influence y. However, the composite factor model (6.3) allows for the effect of further unobserved phenomena not captured via observations of the covariates x. Consequently, our approach in some sense incorporates elements of both canonical correlations analysis and factor analysis. Furthermore, a body of work has considered *factor regression models* [Cav+08] that blend regression analysis and factor analysis similiar in spirit to the composite factor model (6.4). A key distinction is that we model the matrix $\mathcal{A}$ to have low rank. As discussed earlier, this modeling constraint is motivated by the goal of associating semantics to latent variables. It is also important to note that algorithms for factor analysis and for canonical correlations analysis usually operate on covariance and cross-covariance matrices. However, we parametrize our regularized maximum-likelihood problem (6.4) in terms of precision matrices, which is a crucial ingredient in leading to a computationally tractable convex program.

The nuclear-norm heuristic has been employed widely over the past several years in a range of statistical modeling tasks involving rank minimization problems; see [Wai14] and the references therein. The proof of our main result in Section 6.2 incorporates some elements from the theoretical analyses in these previous papers, along with the introduction of some new ingredients. We give specific pointers to the relevant literature in Section 6.4.

Notation

Given a matrix $U \in \mathbb{R}^{p_1\times p_2}$, and the norm $\|U\|_2$ denotes the spectral norm (the largest singular value of U). We define the linear operators $\mathcal{F} : \mathbb{S}^p\times\mathbb{S}^p\times\mathbb{R}^{p\times q}\times\mathbb{S}^q \to \mathbb{S}^{(p+q)}$ and its adjoint $\mathcal{F}^\dagger : \mathbb{S}^{(p+q)} \to \mathbb{S}^p \times \mathbb{S}^p \times \mathbb{R}^{p\times q} \times \mathbb{S}^q$ as follows:

$$\mathcal{F}(M, N, K, O) \triangleq \begin{pmatrix} M - N & K \\ K^T & O \end{pmatrix}, \qquad \mathcal{F}^\dagger \begin{pmatrix} Q & K \\ K^T & O \end{pmatrix} \triangleq (Q, Q, K, O). \tag{6.7}$$

Similarly, we define the linear operators $\mathcal{G} : \mathbb{S}^p \times \mathbb{R}^{p\times q} \to \mathbb{S}^{(p+q)}$ and its adjoint $\mathcal{G}^\dagger : \mathbb{S}^{(p+q)} \to \mathbb{S}^p \times \mathbb{R}^{p\times q}$ as follows:

$$\mathcal{G}(M, K) \triangleq \begin{pmatrix} M & K \\ K^T & 0 \end{pmatrix}, \qquad \mathcal{G}^\dagger \begin{pmatrix} Q & K \\ K^T & O \end{pmatrix} \triangleq (Q, K). \tag{6.8}$$

Finally, for any subspace $\mathfrak{H}$, the projection onto the subspace is denoted by $\mathcal{P}_{\mathfrak{H}}$.

6.2 Theoretical Results

In this section, we state a theorem to prove the consistency of convex program (6.4). This theorem requires assumptions on the population precision matrix, which are discussed in Section 6.2. We provide examples of population composite factor models (6.4) that satisfy these conditions. The theorem statement is given in Section 6.2 and the proof of the theorem is given in Section 6.4 with some details deferred to the supplementary material.

Technical Setup

As discussed in Section 6.1, our theorems are premised on the existence of a population composite factor model (6.3) $y = \mathcal{A}^\star x + \mathcal{B}_u^\star \zeta_u + \epsilon$ underlying a pair of random vectors $(y, x) \in \mathbb{R}^p \times \mathbb{R}^q$, with $\text{rank}(\mathcal{A}^\star) = k_x$, $\mathcal{B}_u^\star \in \mathbb{R}^{p\times k_u}$, and column-space$(\mathcal{A}^\star) \cap$ column-space$(\mathcal{B}_u{}^\star) = \{0\}$. As the convex relaxation (6.4) is solved in the precision matrix parametrization, the conditions for our theorems are more naturally stated in terms of the joint precision matrix $\Theta^\star \in \mathbb{S}^{p+q}$, $\Theta^\star \succ 0$ of (y, x). The algebraic aspects of the parameters underlying the factor model translate to algebraic properties of submatrices of $\Theta^\star$. In particular, the submatrix $\Theta_{yx}^\star$ has rank equal to k_x, and the submatrix $\Theta_y^\star$ is decomposable as $D_y^\star - L_y^\star$ with $D_y^\star$ being diagonal and $L_y^\star \succeq 0$ having rank equal to k_u. Finally, the transversality of column-space$(\mathcal{A}^\star)$ and column-space$(\mathcal{B}_u{}^\star)$ translates to the fact that column-space$(\Theta_{yx}^\star) \cap$ column-space$(L_y^\star) = \{0\}$ have a transverse intersection.

To address the requirements raised in Section 6.1, we seek an estimate $(\hat{\Theta}, \hat{D}_y, \hat{L}_y)$ from the convex relaxation (6.4) such that $\text{rank}(\hat{\Theta}_{yx}) = \text{rank}(\Theta_{yx}^\star)$, $\text{rank}(\hat{L}_y) = \text{rank}(L_y^\star)$, and that $\|\hat{\Theta} - \Theta^\star\|_2$ is small. Building on both classical statistical estimation theory [BD07] as well as the recent literature on high-dimensional statistical inference [BG11; Wai14], a natural set of conditions for obtaining accurate parameter estimates is to assume that the curvature of the likelihood function at $\Theta^\star$ is bounded in certain directions. This curvature is governed by the Fisher information

at $\Theta^\star$:

$$\mathbb{I}^\star \triangleq \Theta^{\star -1} \otimes \Theta^{\star -1} = \Sigma^\star \otimes \Sigma^\star.$$

Here $\otimes$ denotes a tensor product between matrices and $\mathbb{I}^\star$ may be viewed as a map from $\mathbb{S}^{(p+q)}$ to $\mathbb{S}^{(p+q)}$. We impose conditions requiring that $\mathbb{I}^\star$ is well-behaved when applied to matrices of the form:

$$\Theta - \Theta^\star = \begin{pmatrix} (D_y - D_y^\star) - (L_y - L_y^\star) & \Theta_{yx} - \Theta_{yx}^\star \\ \Theta_{yx}{}' - \Theta_{yx}^\star{}' & \Theta_x - \Theta_x^\star, \end{pmatrix}$$

where (L_y, Θ_{yx}) are in a neighborhood of $(L_y^\star, \Theta_{yx}^\star)$ restricted to sets of low-rank matrices. These local properties of $\mathbb{I}^\star$ around $\Theta^\star$ are conveniently stated in terms of *tangent spaces* to the algebraic varieties of low-rank matrices. In particular, the tangent space at a rank-r matrix $N \in \mathbb{R}^{p_1 \times p_2}$ with respect to the algebraic variety of $p_1 \times p_2$ matrices with rank less than or equal to r is given by[4]:

$$\begin{aligned} T(N) \quad \triangleq \quad & \{N_R + N_C | N_R, N_C \in \mathbb{R}^{p_1 \times p_2}, \\ & \text{row-space } (N_R) \subseteq \text{row-space } (N), \\ & \text{column-space } (N_C) \subseteq \text{column-space } (N)\}. \end{aligned}$$

In the next section, we describe conditions on the population Fisher information $\mathbb{I}^\star$ in terms of the tangent spaces $T(L_y^\star)$, and $T(\Theta_{yx}^\star)$; under these conditions, we present a theorem in Section 6.2 showing that the convex program (6.4) obtains accurate estimates.

Fisher Information Conditions

Given a norm $\|\cdot\|_\Upsilon$ on $\mathbb{S}^p \times \mathbb{S}^p \times \mathbb{R}^{p \times q} \times \mathbb{S}^q$, we first consider a classical condition in statistical estimation literature, which is to control the minimum gain of the Fisher information $\mathbb{I}^\star$ restricted to a subspace $\mathbb{H} \subset \mathbb{S}^p \times \mathbb{S}^p \times \mathbb{R}^{p \times q} \times \mathbb{S}^q$ as follows:

$$\chi(\mathbb{H}, \|\cdot\|_\Upsilon) \triangleq \min_{\substack{Z \in \mathbb{H} \\ \|Z\|_\Upsilon = 1}} \|\mathcal{P}_\mathbb{H} \mathcal{F}^\dagger \mathbb{I}^\star \mathcal{F} \mathcal{P}_\mathbb{H}(Z)\|_\Upsilon, \tag{6.9}$$

where $\mathcal{P}_\mathbb{H}$ denotes the projection operator onto the subspace $\mathbb{H}$ and the linear maps $\mathcal{F}$ and $\mathcal{F}^\dagger$ are defined in (6.7). The quantity $\chi(\mathbb{H}, \|\cdot\|_\Upsilon)$ being large ensures that

[4]We also consider the tangent space at a symmetric low-rank matrix with respect to the algebraic variety of symmetric low-rank matrices. We use the same notation 'T' to denote tangent spaces in both the symmetric and non-symmetric cases, and the appropriate tangent space is clear from the context.

the Fisher information $\mathbb{I}^\star$ is well-conditioned restricted to image $\mathcal{F}\mathbb{H} \subseteq \mathbb{S}^{p+q}$. The remaining conditions that we impose on $\mathbb{I}^\star$ are in the spirit of irrepresentibility-type conditions [MB06; ZY06; Wai09; Rav+11a; CPW12] that are frequently employed in high-dimensional estimation. In the subsequent discussion, we employ the following notation to denote restrictions of a subspace $\mathbb{H} = H_1 \times H_2 \times H_3 \times H_4 \subset \mathbb{S}^p \times \mathbb{S}^p \times \mathbb{R}^{p\times q} \times \mathbb{S}^q$ (here H_1, H_2, H_3, H_4 are subspaces in $\mathbb{S}^p$, $\mathbb{S}^p$, $\mathbb{R}^{p\times q}$, $\mathbb{S}^q$, respectively) to its individual components. The restriction to the second components of $\mathbb{H}$ is given by $\mathbb{H}[2] = H_2$. The restriction to the second and third component of $\mathbb{H}$ is given by $\mathbb{H}[2,3] = H_2 \times H_3 \subset \mathbb{S}^p \times \mathbb{R}^{p\times q}$. Given a norm $\|.\|_\Pi$ on $\mathbb{S}^p \times \mathbb{R}^{p\times q}$, we control the gain of $\mathbb{I}^\star$ restricted to $\mathbb{H}[2,3]$

$$\Xi(\mathbb{H}, \|\cdot\|_\Pi) \triangleq \min_{\substack{Z\in\mathbb{H}[2,3]\\ \|Z\|_\Pi=1}} \|\mathcal{P}_{\mathbb{H}[2,3]}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{G}\mathcal{P}_{\mathbb{H}[2,3]}(Z)\|_\Pi. \tag{6.10}$$

Here, the linear maps $\mathcal{G}$ and $\mathcal{G}^\dagger$ are defined in (6.8). In the spirit of irrepresentability conditions, we control the inner-product between elements in $\mathcal{G}\mathbb{H}[2,3]$ and $\mathcal{G}\mathbb{H}[2,3]^\perp$, as quantified by the metric induced by $\mathbb{I}^\star$ via the following quantity:

$$\varphi(\mathbb{H}, \|\cdot\|_\Pi) \triangleq \max_{\substack{Z\in\mathbb{H}[2,3]\\ \|Z\|_\Pi=1}} \|\mathcal{P}_{\mathbb{H}[2,3]^\perp}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{G}\mathcal{P}_{\mathbb{H}[2,3]}(\mathcal{P}_{\mathbb{H}[2,3]}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{G}\mathcal{P}_{\mathbb{H}[2,3]})^{-1}(Z)\|_\Pi. \tag{6.11}$$

The operator $(\mathcal{P}_{\mathbb{H}[2,3]}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{G}\mathcal{P}_{\mathbb{H}[2,3]})^{-1}$ in (6.11) is well-defined if $\Xi(\mathbb{H}) > 0$, since this latter condition implies that $\mathbb{I}^\star$ is injective restricted to $\mathcal{G}\mathbb{H}[2,3]$. The quantity $\varphi(\mathbb{H}, \|\cdot\|_\Upsilon)$ being small implies that any element of $\mathcal{G}\mathbb{H}[2,3]$ and any element of $\mathcal{G}\mathbb{H}[2,3]^\perp$ have a small inner-product (in the metric induced by $\mathbb{I}^\star$). The reason that we restrict this inner product to the second and third components of $\mathbb{H}$ in the quantity $\varphi(\mathbb{H}, \|.\|_\Upsilon)$ is that the regularization terms in the convex program (6.4) are only applied to the matrices L_y and Θ_{yx}.

A natural approach to controlling the conditioning of the Fisher information around $\Theta^\star$ is to bound the quantities $\chi(\mathbb{H}^\star, \|\cdot\|_\Upsilon)$, $\Xi(\mathbb{H}^\star, \|\cdot\|_\Pi)$, and $\varphi(\mathbb{H}^\star, \|\cdot\|_\Upsilon)$ for $\mathbb{H}^\star = \mathcal{W} \times T(L_y^\star) \times T(\Theta_{yx}^\star) \times \mathbb{S}^q$ where $\mathcal{W} \in \mathbb{S}^p$ is the set of diagonal matrices. However, a complication that arises with this approach is that the varieties of low-rank matrices are locally curved around $L_y^\star$ and around $\Theta_{yx}^\star$. Consequently, the tangent spaces at points in neighborhoods around $L_y^\star$ and around $\Theta_{yx}^\star$ are not the same as $T(L_y^\star)$ and $T(\Theta_{yx}^\star)$. In order to account for this curvature underlying the varieties of low-rank matrices, we bound the distance between nearby tangent spaces

via the following induced norm:

$$\rho(T_1, T_2) \triangleq \max_{\|N\|_2 \leq 1} \|(\mathcal{P}_{T_1} - \mathcal{P}_{T_2})(N)\|_2.$$

The quantity $\rho(T_1, T_2)$ measures the largest angle between T_1 and T_2. Using this approach for bounding nearby tangent spaces, we consider subspaces $\mathbb{H}' = \mathcal{W} \times T'_y \times T'_{yx} \times \mathbb{S}^q$ for all T'_y close to $T(L^\star_y)$ and for all T'_{yx} close to $T(\Theta^\star_{yx})$, as measured by ρ [CPW12]. For $\omega_y \in (0, 1)$ and $\omega_{yx} \in (0, 1)$, we bound $\chi(\mathbb{H}', \|\cdot\|_\Upsilon)$, $\Xi(\mathbb{H}', \|\cdot\|_\Pi)$, and $\varphi(\mathbb{H}', \|\cdot\|_\Pi)$ in the sequel for all subspaces $\mathbb{H}'$ in the following set:

$$\begin{aligned} U(\omega_y, \omega_{yx}) \triangleq \Big\{ \mathcal{W} \times T'_y \times T'_{yx} \times \mathbb{S}^q \mid \rho(T'_y, T(L^\star_y)) \leq \omega_y \\ \rho(T'_{yx}, T(\Theta^\star_{yx})) \leq \omega_{yx} \Big\}. \end{aligned} \tag{6.12}$$

We control the quantities $\Xi(\mathbb{H}', \|\cdot\|_\Pi)$ and $\varphi(\mathbb{H}', \|\cdot\|_\Pi)$ using the dual norm of the regularizer $\text{trace}(L_y) + \gamma\|\Theta_{yx}\|_\star$ in (6.4):

$$\Gamma_\gamma(L_y, \Theta_{yx}) \triangleq \max\left\{ \|L_y\|_2, \frac{\|\Theta_{yx}\|_2}{\gamma} \right\}. \tag{6.13}$$

Furthermore, we control the quantity $\chi(\mathbb{H}', \|\cdot\|_\Upsilon)$ using a slight variant of the dual norm:

$$\Phi_\gamma(D_y, L_y, \Theta_{yx}, \Theta_x) \triangleq \max\left\{ \|D_y\|_2, \|L_y\|_2, \frac{\|\Theta_{yx}\|_2}{\gamma}, \|\Theta_x\|_2 \right\}. \tag{6.14}$$

As the dual norm $\max\left\{ \|L_y\|_2, \frac{\|\Theta_{yx}\|_2}{\gamma} \right\}$ of the regularizer in (6.4) plays a central role in the optimality conditions of (6.4), controlling the quantities $\chi(\mathbb{H}', \Phi_\gamma)$, $\Xi(\mathbb{H}', \Gamma_\gamma)$, and $\varphi(\mathbb{H}', \Gamma_\gamma)$ leads to a natural set of conditions that guarantee the consistency of the estimates produced by (6.4). In summary, given a fixed set of parameters $(\gamma, \omega_y, \omega_{yx}) \in \mathbb{R}_+ \times (0, 1) \times (0, 1)$, we assume that $\mathbb{I}^\star$ satisfies the following conditions:

$$\text{Assumption 1} \quad : \quad \inf_{\mathbb{H}' \in U(\omega_y, \omega_{yx})} \chi(\mathbb{H}', \Phi_\gamma) \geq \alpha, \quad \text{for some } \alpha > 0 \tag{6.15}$$

$$\text{Assumption 2} \quad : \quad \inf_{\mathbb{H}' \in U(\omega_y, \omega_{yx})} \Xi(\mathbb{H}', \Gamma_\gamma) > 0 \tag{6.16}$$

$$\text{Assumption 3} \quad : \quad \max_{\mathbb{H}' \in U(\omega_{yx}\beta, \omega_{yx})} \varphi(\mathbb{H}', \Gamma_\gamma) \leq 1 - \frac{2}{\beta + 1} \quad \text{for some } \beta \geq 2 \tag{6.17}$$

For fixed $(\gamma, \omega_y, \omega_{yx})$, larger value of α and smaller value of β in these assumptions lead to a better conditioned $\mathbb{I}^\star$.

Assumptions 1, 2, and 3 are analogous to conditions that play an important role in the analysis of the Lasso for sparse linear regression, graphical model selection via the Graphical Lasso [Rav+11a], and in several other approaches for high-dimensional estimation. As a point of comparison with respect to analyses of the Lasso, the role of the Fisher information $\mathbb{I}^\star$ is played by $A^T A$, where A is the underlying design matrix. In analyses of both the Lasso and the Graphical Lasso in the papers referenced above, the analog of the subspace $\mathbb{H}$ is the set of models with support contained inside the support of the underlying sparse population model. Assumptions 1, 2, and 3 are also similar in spirit to conditions employed in the analysis of convex relaxation methods for latent-variable graphical model selection [CPW12].

When Do the Fisher Information Assumptions Hold?

In this section, we provide examples of composite models (6.3) that satisfy Assumptions 1, 2 and 3 in (6.15) (6.16), and (6.17) for some choices of $\alpha > 0$, $\beta \geq 2$, $\omega_y \in (0,1)$, $\omega_{yx} \in (0,1)$ and $\gamma > 0$. Specifically, consider a population composite factor model $y = \mathcal{A}^\star x + \mathcal{B}_u^\star \zeta_u + \bar{\epsilon}$, where $\mathcal{A}^\star \in \mathbb{R}^{p \times q}$ with $\text{rank}(\mathcal{A}^\star) = k_x$, $\mathcal{B}_u^\star \in \mathbb{R}^{p,k_u}$, column-space$(\mathcal{A}^\star) \cap$ column-space$(\mathcal{B}_u^\star) = \{0\}$, and the random variables $\zeta_u, \bar{\epsilon}, x$ are independent of each other and normally distributed as $\zeta_u \sim \mathcal{N}(0, \Sigma_{\zeta_u}), \bar{\epsilon} \sim \mathcal{N}(0, \Sigma_{\bar{\epsilon}})$. As described in Section 6.1, the properties of the composite factor model translate to algebraic properties on the underlying precision matrix $\Theta^\star \in \mathbb{S}^{p+q}$. Namely, the submatrix $\Theta_{yx}^\star$ has rank equal to k_x and the submatrix $\Theta_y^\star$ is decomposable as $D_y^\star - L_y^\star$ with $D_y^\star$ being diagonal and $L_y^\star \geq 0$ having rank equal to k_u. Recall that the factor model underlying the random vector $y \in \mathbb{R}^p$ that is induced upon marginalization of x is specified by the precision matrix of y given by $\tilde{\Theta}_y^\star = D_y^\star - \left[L_y^\star + \Theta_{yx}^\star (\Theta_x^\star)^{-1} \Theta_{xy}^\star \right]$. Here, $L_y^\star + \Theta_{yx}^\star (\Theta_x^\star)^{-1} \Theta_{xy}^\star$ represents the effect of the latent variables in the underlying factor model. When learning a composite factor model, this effect is decomposed into: $\Theta_{yx}^\star (\Theta_x^\star)^{-1} \Theta_{xy}^\star$ – a rank k_x matrix representing the component of this affect attributed to x – and $L_y^\star$ – a matrix of rank k_u representing the effect of residual latent variables. There are two identifiability concerns that arise when learning a composite factor model. First, the low rank matrices $L_y^\star$ and $\Theta_{yx}^\star (\Theta_x^\star)^{-1} \Theta_{xy}^\star$ must be distinguishable from the diagonal matrix $D_y^\star$. Following previous literature in diagonal and low rank matrix decompositions [Sau+12; CPW12], this task can be achieved by ensuring that the column/row spaces of $L_y^\star$ and $\Theta_{yx}^\star (\Theta_x^\star)^{-1} \Theta_{xy}^\star$ are *incoherent* with respect to the standard basis. Specifically, given a subspace $U \subset \mathbb{R}^p$, the coherence of the

subspace U is defined as:

$$\mu(U) = \max_{i=1,2\ldots p} \|\mathcal{P}_U(e_i)\|_{\ell_2}^2,$$

where $\mathcal{P}$ denotes a projection operation and $e_i \in \mathbb{R}^p$ denotes the i'th standard basis vector. It is not difficult to show that this incoherence parameter satisfies the following inequality:

$$\frac{\dim(U)}{p} \leq \mu(U) \leq 1.$$

A subspace U with small coherence is necessarily of small dimension and far from containing standard basis elements. As such, a symmetric matrix with incoherent row and column spaces is low-rank and quite different from being a diagonal matrix. Consequently, we require that the quantities $\mu(\text{column-space}(L_y^\star))$ and $\mu(\text{column-space}(\Theta_{yx}^\star \Theta_x^{\star -1} \Theta_{xy}^\star))$ are small [5]. The second identifiability issue that arises is distinguishing the low rank matrices $L_y^\star$ and $\Theta_{yx}^\star (\Theta_x^\star)^{-1} \Theta_{xy}^\star$ from one another. This task is made difficult when the row/column spaces of these matrices are nearly aligned. Thus, we must ensure that the row/column spaces of $L_y^\star$ and $\Theta_{yx}^\star (\Theta_x^\star)^{-1} \Theta_{xy}^\star$ are sufficiently transverse (i.e. have large angles).

These identifiability issues directly translate to conditions on the population composite factor model. Specifically, $\mu(\text{column-space}(L_y^\star))$ and $\mu(\text{column-space}(\Theta_{yx}^\star (\Theta_x^\star)^{-1} \Theta_{xy}^\star))$ being small translates to $\mu(\text{column-space}(\mathcal{A}^\star))$ and $\mu(\text{column-space}(\mathcal{B}_u^\star))$ being small. Such a condition has another interpretation. It states that the effect of x and ζ_u must not concentrate on any one variable of y; otherwise, this effect can be absorbed by the random variable $\bar{\epsilon}$ in (6.3). The second identifiability assumption that the row/column spaces of $L_y^\star$ and $\Theta_{yx}^\star (\Theta_x^\star)^{-1} \Theta_{xy}^\star$ have a large angle translates to the angle between column spaces of $\mathcal{A}^\star$ and $\mathcal{B}_u^\star$ being large. This assumption ensures that the effect of x and ζ_u on y can be distinguished.

Having these identifiability concerns in mind, we give a stylized composite factor model (6.3) and numerically check that the Fisher Information Assumptions 1,2, and 3 in (6.15), (6.16), and (6.17) are satisfied for appropriate choices of parameters. Specifically, we let $p = 30$, $q = 2$, $k_x = 1$, and $k_u = 1$. We let the random variables $x \in \mathbb{R}^q$, $\zeta_u \in \mathbb{R}^{k_u}$, $\bar{\epsilon} \in \mathbb{R}^p$ be distributed according to $x \sim \mathcal{N}(0, \mathcal{I}_{q\times q})$,

[5] We only need to control the coherence of the column spaces since these matrices are symmetric.

$\zeta_u \sim \mathcal{N}(0, \mathcal{I}_{k_u \times k_u})$, and $\bar{\epsilon} \sim \mathcal{N}(0, \mathcal{I}_{p \times p})$. We generate matrices $J \in \mathbb{R}^{p \times k_x}, K \in \mathbb{R}^{q \times k_x}$ with i.i.d Gaussian entries, and let $\mathcal{A}^\star = JK^T$. Similarly, we generate $\mathcal{B}_u^\star \in \mathbb{R}^{p \times k_u}$ with i.i.d Gaussian entries. We scale the matrices $\mathcal{A}^\star$ and $\mathcal{B}_u^\star$ to have spectral norm equal to 0.1. Taking an instantiation of these matrices, the smallest angle between the column spaces of $\mathcal{A}^\star$ and $\mathcal{B}_u^\star$ is 87 degrees. Furthermore, the quantities $\mu(\text{column-space}(\mathcal{A}^\star))$ and $\mu(\text{column-space}(\mathcal{B}_u^\star))$ are 0.009 and 0.01 respectively. Thus, our stylized model satisfies the identifiability assumptions discussed earlier in this section. Under this stylized setting, we numerically evaluate Assumptions 1, 2, and 3 in (6.15), (6.16), and (6.17) with a Fisher information $\mathbb{I}^\star$ that takes the form:

$$\mathbb{I}^\star = \begin{pmatrix} \mathcal{I} + \mathcal{A}^\star \mathcal{A}^{\star T} + \mathcal{B}_u^\star \mathcal{B}_u^{\star T} & \mathcal{A}^\star \\ \mathcal{A}^{\star T} & \mathcal{I} \end{pmatrix} \otimes \begin{pmatrix} \mathcal{I} + \mathcal{A}^\star \mathcal{A}^{\star T} + \mathcal{B}_u^\star \mathcal{B}_u^{\star T} & \mathcal{A}^\star \\ \mathcal{A}^{\star T} & \mathcal{I} \end{pmatrix}.$$

We let $\omega_y = 0.003, \omega_{yx} = 0.003$ so that the largest angle between the pair of tangent spaces $T'_y, T(L_y^\star)$ and tangent spaces $T'_{yx}, T(\Theta_{yx}^\star)$ is less than 0.1 degrees. Employing a numerical procedure described in Section 1 of the supplementary material, we obtain a range of values of γ, $\alpha > 0$, and $\beta \geq 2$ that satisfy Assumptions 1,2, and 3 in (6.15), (6.16), and (6.17). The values of α and β that are computed using this procedure serve as a lower and upper bound for the optimal α and β, respectively. Indeed, an exciting direction for future research is to develop sharper numerical or analytical techniques to precisely characterize the optimal values of α and β. Table 1 illustrates ranges of γ and the corresponding values of α and β that satisfy Fisher information assumptions 1 and 3. We note that for all the ranges of γ shown in this table, $\inf_{\mathbb{H}' \in U(\omega_y, \omega_{yx})} \Xi(\mathbb{H}') > 0.32$ so that Assumption 2 is also satisfied. Examining Table 1, we observe that a larger range of γ results in a smaller value of α and a larger value of β.

γ	$\alpha \geq$	$\beta \leq$
(0.87,1.04)	0.058	49
(0.89,1.04)	0.060	24
(0.91,1.03)	0.061	15
(0.95,1.02)	0.065	9

Table 6.1: Ranges of γ and the corresponding values of α and β that satisfy Assumptions 1,2, and 3.

Theorem Statement

We now describe the performance of the regularized maximum-likelihood programs (6.4) under suitable conditions on the quantities introduced in the previous

section. Before formally stating our main result, we introduce some notation. Let σ_y denote the minimum nonzero singular value of $L_y^\star$ and let σ_{yx} denote the minimum nonzero singular value of $\Theta_{yx}^\star$. We state the theorem based on essential aspects of the conditions required for the success of our convex relaxation (i.e. the Fisher information conditions) and omit complicated constants. We specify these constants in Section 4.

Theorem 5. *Suppose that there exists $\alpha > 0$, $\beta \geq 2$, $\omega_y \in (0,1)$, $\omega_{yx} \in (0,1)$, and the choice of parameter γ so that the population Fisher information $\mathbb{I}^\star$ satisfies Assumptions 1, 2, and 3 in* (6.15), (6.16) *and* (6.17). *Let $m \triangleq \max\{1, \frac{1}{\gamma}\}$, and $\bar{m} \triangleq \max\{1, \gamma\}$. Furthermore, suppose that the following conditions hold:*

1. $n \gtrsim \left[\frac{\beta^2}{\alpha^2} m^6\right](p+q)$
2. $\lambda_n \sim \left[\frac{\beta}{\alpha} m^2\right]\sqrt{\frac{p+q}{n}}$
3. $\sigma_y \gtrsim \left[\frac{\beta}{\alpha^5 \omega_y} m^4\right]\lambda_n$
4. $\sigma_{yx} \gtrsim \left[\frac{\beta}{\alpha^5 \omega_{yx}} m^5 \bar{m}^2\right]\lambda_n$

Then with probability greater than $1 - 2\exp\left\{-\tilde{C}_{prob}\frac{\alpha^2}{\beta^2 m^4} n\lambda_n^2\right\}$, the optimal solution $(\hat{\Theta}, \hat{D}_y, \hat{L}_y)$ of (6.4) *with i.i.d. observations $\mathcal{D}_n^+ = \{y^{(i)}, x^{(i)}\}_{i=1}^n$ of (y,x) satisfies the following properties:*

1. *$rank(\hat{L}_y) = rank(L_y^\star)$, $rank(\hat{\Theta}_{yx}) = rank(\Theta_{yx}^\star)$*
2. $\|\hat{D}_y - D_y^\star\|_2 \lesssim \frac{m\bar{m}}{\alpha^2}\lambda_n$, $\|\hat{L}_y - L_y^\star\|_2 \lesssim \frac{m\bar{m}}{\alpha^2}\lambda_n$, $\|\hat{\Theta}_{yx} - \Theta_{yx}^\star\|_2 \lesssim \frac{m\bar{m}}{\alpha^2}\lambda_n$, $\|\hat{\Theta}_x - \Theta_x^\star\|_2 \lesssim \frac{m\bar{m}}{\alpha^2}\lambda_n$

We outline the proof of Theorem 5 in Section 6.4. The quantities $\alpha, \beta, \omega_y, \omega_{yx}$ as well as the choices of parameters γ play a prominent role in the results of Theorem 5. Indeed larger values of $\alpha, \omega_y, \omega_{yx}$ and smaller values of β (leading to a better conditioned Fisher information even for large distortions around the tangent space $T(L_y^\star)$ and $T(\Theta_{yx}^\star)$ lead to less stringent requirements on the sample complexity, on the minimum nonzero singular value of σ_y of $L_y^\star$, and on the minimum nonzero singular value σ_{yx} of $\Theta_{yx}^\star$.

Identifying an Accurate Factor Model

Our objective is to learn a composite factor model that is close to a factor model underlying y. As such a factor model is often not available in advance, we present an approach for learning a factor model (6.1) using observations of y. In particular, we fit observations $\mathcal{D}_n = \{y^{(i)}\}_{i=1}^n$ to the factor model (6.1) using the following convex relaxation:

$$(\hat{\tilde{D}}_y, \hat{\tilde{L}}_y) = \arg \min_{\substack{\tilde{D}_y, \tilde{L}_y \in \mathbb{S}^p \\ \tilde{D}_y - \tilde{L}_y \succ 0}} \quad -\ell(\tilde{D}_y - \tilde{L}_y; \mathcal{D}_n) + \tilde{\lambda}_n \mathrm{trace}(\tilde{L}_y)$$
$$\text{s.t.} \qquad \tilde{L}_y \succeq 0, \tilde{D}_y \text{ is diagonal.} \tag{6.18}$$

We note that the convex program (6.18) is a specialization of the convex program (6.4) for learning a composite factor model. The parameter $\tilde{\lambda}_n$ in (6.18) provides a tradeoff between fidelity of the model to the observations and the complexity of the model (i.e., the number of latent variables). In contrast to minimum-trace factor analysis – in which the objective is to decompose a covariance matrix as the sum of a diagonal matrix and a low-rank matrix [Led40; Sha82a; Sha82b; Sha04]– the regularized maximum-likelihood convex program (6.18) fits factor models by decomposing a precision matrix as the difference between a diagonal matrix and a low-rank matrix. Although the focus of this paper is not about learning a factor model accurately, we characterize the consistency of the convex relaxation (6.18) under Assumptions on the population Fisher information with respect to y. Specifically, let $\tilde{\alpha}$ and $\tilde{\beta}$ denote analogous quantities to α and β in Fisher information assumptions 1 and 3. Let σ denote the minimum nonzero singular value of $L_y^\star + \Theta_{yx}^\star (\Theta_x^\star)^{-1} \Theta_{xy}^\star$. Then, the convex program (6.18) succeeds under appropriate Fisher information conditions and $n \gtrsim \left\lceil \frac{\tilde{\beta}^2}{\tilde{\alpha}^2} \right\rceil p$, $\tilde{\lambda}_n \sim \frac{\tilde{\beta}}{\tilde{\alpha}} \sqrt{\frac{p}{n}}$, and $\sigma \gtrsim \frac{\tilde{\beta}}{\tilde{\alpha}^5 \tilde{\omega}_y} \tilde{\lambda}_n$. We present the complete technical discussion in Section 6.6 of the supplementary material.

6.3 Experimental Results

In this section, we demonstrate the utility of Algorithm 1 for interpreting latent variables in factor models both with synthetic and real financial asset data.

Synthetic Simulations

We give experimental evidence for the utility of Algorithm 1 on synthetic examples. Specifically, we generate a composite factor model (6.3) $y = \mathcal{A}^\star x + \mathcal{B}_u^\star \zeta_u + \bar{\epsilon}$ as follows: we fix $p = 60$ and $q = 10$. We let the random variables $x \in \mathbb{R}^q$, $\zeta_u \in \mathbb{R}^{k_u}$, $\bar{\epsilon} \in \mathbb{R}^p$ be distributed according to $x \sim \mathcal{N}(0, \mathcal{I}_{q \times q})$, $\zeta_u \sim \mathcal{N}(0, \mathcal{I}_{k_u \times k_u})$,

and $\bar{\epsilon} \sim \mathcal{N}(0, \mathcal{I}_{p\times p})$. We generate matrices $J \in \mathbb{R}^{p\times k_x}, K \in \mathbb{R}^{q\times k_x}$ with iid Gaussian entries, and let $\mathcal{A}^\star = JK^T$. Similarly, we generate $\mathcal{B}_u^\star \in \mathbb{R}^{p\times k_u}$ with i.i.d Gaussian entries. This approach generates a factor model (6.1) with $k = k_x + k_u$. The composite factor model translates to a joint precision matrix $\Theta^\star$, with the submatrix $\Theta_y^\star = D_y^\star - L_y^\star$ where $D_y^\star$ is diagonal, $\text{rank}(L_y^\star) = k_u$, and $\text{rank}(\Theta_{yx}^\star) = k_x$. We scale matrices $\mathcal{A}^\star$ and $\mathcal{B}_u^\star$ to have spectral norm equal to τ. The value τ is chosen to be as large as possible without the condition number of $\Theta^\star$ exceeding 7 (this is imposed for the purposes of numerical conditioning). We obtain four models with $(k_x, k_u) = (1, 1), (k_x, k_u) = (2, 2)$, and $(k_x, k_u) = (3, 3)$, and $(k_x, k_u) = (4, 4)$.

For the purposes of this experiment, we assume that the input to Algorithm 1 is the oracle factor model specified by the parameters $(D_y^\star, L_y^\star + \Theta_{yx}^\star(\Theta_x)^{-1}\Theta_{xy}^\star)$, and demonstrate the success of steps 2-3 of Algorithm 1. In particular, for each model, we generate n samples of responses y and covariates x, and use these observations as input to the convex program (6.4). The regularization parameters λ_n, γ are chosen so that the estimates $(\hat{\Theta}, \hat{L}_y, \hat{D}_y)$ satisfy (i) $\text{rank}(L_y^\star + \Theta_{yx}^\star(\Theta_x^\star)^{-1}\Theta_{xy}^\star) = \text{rank}(\hat{L}_y) + \text{rank}(\hat{\Theta}_{yx}\hat{\Theta}_x^{-1}\hat{\Theta}_{xy})$,
(ii) $\text{column-space}(\hat{\Theta}_{yx}) \cap \text{column-space}(\hat{L}_y) = \{0\}$, and the deviation from the underlying factor model
$\max\{\|D_y^\star - \hat{D}_y\|_2/\|D_y^\star\|_2, \|L_y^\star - [\hat{L}_y + \hat{\Theta}_{yx}\hat{\Theta}_x^{-1}\hat{\Theta}_{xy}]\|_2/\|L_y^\star\|_2\}$ is minimized. Figure 1(a) shows the magnitude of the deviation for different values of n. Furthermore, for each fixed n, we use the choice of regularization parameters (λ_n, γ) to compute the probability of obtaining structurally correct estimates of the composite model (i.e. $\text{rank}(\hat{L}_y) = \text{rank}(L_y^\star)$ and $\text{rank}(\Theta_{yx}^\star) = \text{rank}(\hat{\Theta}_y)$). These probabilities are evaluated over 10 experiments and are shown in Figure 1(b). These results support Theorem 1 that given (sufficiently many) samples of responses/covariates, the convex program (6.4) provides accurate estimates of the composite factor model (6.3).

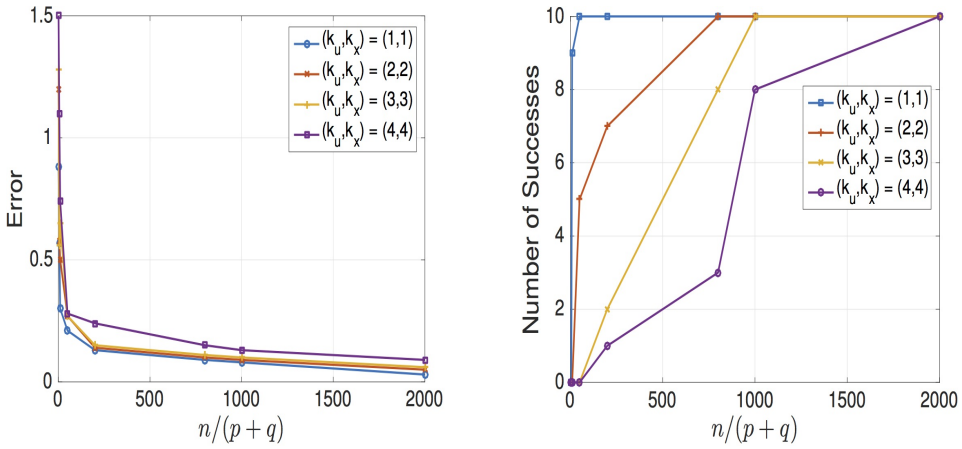


Figure 6.1: Synthetic data: plot shows the error (defined in the main text) and probability of correct structure recovery in composite factor models. The four models studied are (i) $(k_x, k_u) = (1, 1)$, (ii) $(k_x, k_u) = (2, 2)$, and (iii) $(k_x, k_u) = (3, 3)$, and (iv) $(k_x, k_u) = (4, 4)$. For each plotted point in (b), the probability of structurally correct estimation is obtained over 10 trials.

Experimental Results on Financial Asset Data

We consider as our responses y the monthly stock returns of $p = 45$ companies from the Standard and Poor index over the period March 1982 to March 2016, which leads to a total of $n = 408$ observations. We then obtain measurements of 13 covariates that can plausibly influence the values of stock prices: consumer price index, producer price index, EUR to USD exchange rate, federal debt (normalized by GDP), federal reserve rate, GDP growth rate, government spending (normalized by GDP), home ownership rate, industrial production index, inflation rate, mortgage rate, oil import, and saving rate. Of these 13 covariates, the covariates federal debt, government spending, GDP growth rate, and home ownership rate are only available at a quarterly scale. Monthly observations are available for the remaining covariates. Evidently, many more observations of y are available than of (y, x) jointly. As described in Section 6.1, this scenario motivates us to first learn a factor model using the monthly observations of y. We then associate semantics to the latent variables of this factor model by fitting a composite factor model to the more limited joint observations of (y, x).

For the purpose of learning a factor model, we set aside a random subset of $n_{\text{train}} = 308$ of the total $n = 408$ observations as a training set and the remaining subset of $n_{\text{test}} = 100$ as the test set. We let $\mathcal{D}_{\text{train}} = \{y^{(i)}\}_{i=1}^{n_{\text{train}}}$ and $\mathcal{D}_{\text{test}} = \{y^{(i)}\}_{i=1}^{n_{\text{test}}}$ be the corresponding training and testing data sets respectively. We use the observations $\mathcal{D}_{\text{train}}$ as input to the convex program (6.18) where the regularization parameter $\tilde{\lambda}_n$

is chosen via cross-validation. Concretely, for a particular choice of $\tilde{\lambda}_n$, we supply $\mathcal{D}_{\text{train}}$ as input to the convex program (6.18), and solve (6.18) to obtain a factor model specified by $(\hat{\tilde{D}}_y, \hat{\tilde{L}}_y)$. We then compute the average log-likelihood over the testing set $\mathcal{D}_{\text{test}}$ using the distribution specified by the precision matrix $\hat{\tilde{D}}_y - \hat{\tilde{L}}_y$. We perform this procedure as we vary $\tilde{\lambda}_n$ from 0.04 to 4 in increments of 0.004. Figure 2 shows a plot of $\text{rank}(\hat{\tilde{L}}_y))$ (i.e. number of latent factors) vs. average log-likelihood performance on the testing set. Notice that fixing the number of latent factors does not lead to a unique factor model as varying the regularization parameter $\tilde{\lambda}_n$ may lead to a change in the estimated model, but no change in its structure (i.e. $\text{rank}(\hat{\tilde{L}}_y)$ remains the same). As larger values of average log-likelihood are indicative of a better fit to test samples, these results suggest that 10 latent factors influence stock prices. We thus focus on associating semantics to the factor model with the largest average log-likelihood performance that consists of 10 latent factors.

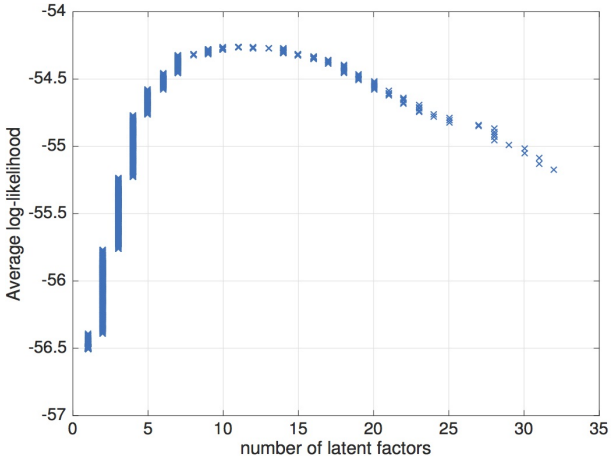


Figure 6.2: Number of latent factors vs. average log-likelihood over testing set. These results are obtained by sweeping over parameters $\tilde{\lambda}_n \in [0.04, 4]$ in increments of 0.004 and solving the convex program (6.18).

We now proceed with the steps 2-3 of Algorithm 1. To obtain a consistent set of joint observations (y, x) to employ as input to the convex program (6.4), we apply a 3-month averaging for each variable that is available at a monthly scale (i.e. the responses y and the covariates x with the exception of the four specified earlier) to obtain quarterly measurements. This leads to $n = 137$ quarterly measurements. We denote the quarterly responses and covariates by $\tilde{y}$ and $\tilde{x}$, respectively. We let $\mathcal{D}_n^+ = \{(\tilde{y}^{(i)}, \tilde{x}^{(i)})\}_{i=1}^n$ be the set of joint quarterly observations of response $\tilde{y}$ and covariates $\tilde{x}$. Using observations $\mathcal{D}_n^+$ as input to the convex program (6.4) , we perform an exhaustive sweep over parameter space (λ_n, γ) to learn composite

models with estimates $(\hat{\Theta}, \hat{D}_y, \hat{L}_y)$ such that $\text{rank}(\hat{\Theta}) = 0, 1, 2, \ldots 10$, and $\text{rank}(\hat{L}_y) = 0, 1, 2, \ldots 10$. As we are interested comparing these composite models to the factor model with 10 latent variables, we finely grid the parameter space (λ_n, γ) so that there are a large number of models for which $\text{rank}(\hat{\Theta}) + \text{rank}(\hat{L}_y)$ is equal to 10. Among these models, we restrict to those that satisfy the conditions of step 3 of Algorithm 1. Table 2 shows the number of models that satisfy these conditions for $\text{rank}(\hat{\Theta}_{yx}) = 1, \ldots, 5$. For each $d = 1, \ldots, 5$, we then identify the composite factor model which minimizes the quantity $\max\{\|\hat{\tilde{D}}_y - \hat{D}_y\|_2/\|\hat{\tilde{D}}_y\|_2, \|\hat{\tilde{L}}_y - \hat{L}_y - \hat{\Theta}_{yx}\hat{\Theta}_x^{-1}\hat{\Theta}_{xy}]\|_2/\|\hat{\tilde{L}}_y\|_2\}$. Table 3 shows the values of this quantity for $\text{rank}(\hat{\Theta}_{yx}) = 1, \ldots, 5$ with respect to the factor model with 10 latent variables.

$(\text{rank}(\hat{\Theta}_{yx}), \text{rank}(\hat{L}_y))$	# models satisfying conditions of step 2.
(1,9)	167
(2,8)	196
(3,7)	218
(4,6)	110
(5,5)	98

Table 6.2: Number of composite factor models with $\text{rank}(\hat{\Theta}_{yx}) = 1, \ldots, 5$ that satisfy the requirements of step 2 in Algorithm 1(for the factor model with 10 latent variables).

$(\text{rank}(\hat{\Theta}_{yx}), \text{rank}(\hat{L}_y))$	$\max\{\|\hat{\tilde{D}}_y - \hat{D}_y\|_2/\|\hat{\tilde{D}}_y\|_2, \|\hat{\tilde{L}}_y - \hat{L}_y - \hat{\Theta}_{yx}\hat{\Theta}_x^{-1}\hat{\Theta}_{xy}]\|_2/\|\hat{\tilde{L}}_y\|_2\}$
(1,9)	0.39
(2,8)	0.40
(3,7)	0.47
(4,6)	0.51
(5,5)	0.55

Table 6.3: Deviation of the candidate composite factor model from the factor model consisting of 10 latent variables.

Examining Table 3, we note that there is large increase in deviation as $\text{rank}(\hat{\Theta}_{yx})$ is increased above 2. Thus, we consider the composite factor model with $\text{rank}(\hat{\Theta}_{yx}) = 2$ to be an acceptable approximation of the underlying factor model. As a final step of the algorithm, we investigate the properties of the two-dimensional row-space of $\hat{\Theta}_{yx}$ to shed some light on those covariates that appear to play a significant role in capturing some of the latent phenomena in the 10-factor model. In particular, for the composite factor model with $(\text{rank}(\hat{\Theta}_{yx}), \text{rank}(\hat{L}_y)) = (2, 8)$ (second row in Table 3), we let $V \in \mathbb{R}^{13\times 2}$ denote a matrix with orthogonal, unit-norm columns such

that the columns of V form a basis for the row space of $\hat{\Theta}_{yx}$ (such a matrix may be computed, for example, via the singular value decomposition). Thus, the projection of x onto the row-space of $\hat{\Theta}_{yx}$ — given by $V^T x$ — represents the interpretable component of the latent variables. We then consider the Euclidean-squared-norm of the i-th row of V, as this specifies the relative strength of the i-th covariate. As shown in Table 6.4, all covariates have some contribution (as we allow general linear combinations of the covariates x in the composite factor model (6.3)). However, the covariates exchange rate, government expenditures, and GDP growth rate seem to be the most relevant, and the covariates mortgage rate and oil import seem to be the least relevant.

covariate	strength
Exchange rate	0.18
Government expenditures	0.14
GDP growth rate	0.11
Home ownership rate	0.09
Industrial production rate	0.08
PPI	0.08
CPI	0.07
Federal debt	0.06
Saving rate	0.04
Inflation rate	0.04
Federal reserve rate	0.03
Oil import	0.03
Mortgage rate	0.01

Table 6.4: Strength of each covariate in the composite factor model with 2-dimensional projection of covariates and 8 latent variables.

6.4 Proof Strategy of Theorem 5

We first begin by specifying the constants in Theorem 5. Let $\psi \triangleq \|\Theta^{\star -1}\|_2$, $\tilde{C} = 112\psi^3$, $\tilde{C}_0 = \max\{\frac{1}{196\psi}, \frac{7}{12}\psi, \frac{1}{2\psi^4}\}$, $\tilde{C}_{samp} = \tilde{C}\tilde{C}_0$, $\tilde{C}_1 = 148\psi^2 + 24\psi^4$, $\tilde{C}_\sigma = 84\psi^4(24\psi^4 + 148\psi^2)^2$, and $\tilde{C}_{prob} = \frac{1}{25088\psi^6}$. The precise conditions on the number of observations, the regularization parameter λ_n, minimum nonzero singular value of $L_y^\star$ and minimum nonzero singular value of $\Theta_{yx}^\star$ for Theorem 5 are given by:

1. $n \geq \tilde{C}_{samp}^2 \left[\frac{\beta^4}{\alpha^2} m^6 (p+q)\right]$
2. $\lambda_n \in \left[\tilde{C}\left\{\frac{\beta}{\alpha} m^2 \sqrt{\frac{p+q}{n}}\right\}, \frac{1}{\beta m \tilde{C}_0}\right]$

3. $\sigma_y \geq \tilde{C}_\sigma \left[\frac{\beta}{\alpha^5 \omega_y} m^4 \bar{m} \lambda_n \right]$

4. $\sigma_{yx} \geq \tilde{C}_\sigma \left[\frac{\beta}{\alpha^5 \omega_{yx}} m^5 \bar{m}^3 \lambda_n \right]$

Moreover, under these conditions, with probability greater than $1 - 2\exp\left(- \tilde{C}_{prob} \frac{\alpha^2}{m^4 \beta^2} n \lambda_n^2 \right)$, the optimal solution of the convex program (6.4) with estimates $(\hat{\Theta}, \hat{L}_y, \hat{D}_y)$ satisfies the following properties:

1. $\text{rank}(\hat{L}_y) = \text{rank}(L_y^\star)$, $\text{rank}(\hat{\Theta}_{yx}) = \text{rank}(\Theta_{yx}^\star)$

2. $\|\hat{D}_y - D_y^\star\|_2 \leq \tilde{C}_1 \frac{m\bar{m}}{\alpha^2} \lambda_n$, $\|\hat{L}_y - L_y^\star\|_2 \leq \tilde{C}_1 \frac{m\bar{m}}{\alpha^2} \lambda_n$, $\|\hat{\Theta}_{yx} - \Theta_{yx}^\star\|_2 \leq \tilde{C}_1 \frac{m\bar{m}^2}{\alpha^2} \lambda_n$, $\|\hat{\Theta}_x - \Theta_x^\star\|_2 \leq \tilde{C}_1 \frac{m\bar{m}}{\alpha^2} \lambda_n$.

Now under assumptions of Theorem 5, we construct appropriate primal feasible variables $(\hat{\Theta}, \hat{D}_y, \hat{L}_y)$ that satisfy the conclusions of the theorem - i.e., $\hat{\Theta}_{yx}$, $\hat{L}_y$ are low-rank (with the same ranks as the underlying population quantities $\Theta_{yx}^\star$ and $L_y^\star$) - and for which there exists a corresponding dual variable certifying optimality. This proof technique is sometimes also referred to as a primal-dual witness or certificate approach [Wai09]. The high-level proof strategy is similar in spirit to the proofs of consistency results for sparse graphical model recovery [Rav+11a] and latent variable graphical model recovery [CPW12], although our convex program and the conditions required for its success are different from these previous results. Consider the following convex program:

$$(\hat{\Theta}, \hat{D}_y, \hat{L}_y) = \arg \min_{\substack{\Theta \in \mathbb{S}^{p+q},\ \Theta \succ 0 \\ D_y, L_y \in \mathbb{S}^p}} -\ell(\Theta; \mathcal{D}_n^+) + \lambda_n [\gamma \|\Theta_{yx}\|_\star + \|L_y\|_\star]$$
$$\text{s.t.} \qquad \Theta_y = D_y - L_y, D_y \text{ is diagonal} \tag{6.19}$$

Comparing (6.19) with the convex program (6.4), the difference is that we no longer constrain L_y to be a positive semidefinite matrix. In particular, if $L_y \succeq 0$, then the nuclear norm of the matrix L_y in the objective function of (6.19) reduces to the trace of L_y. We show in the supplementary material that with high probability, the matrix $\hat{L}_y$ is positive semidefinite. Standard convex analysis states that $(\hat{\Theta}, \hat{D}_y, \hat{L}_y)$ is the solution of the convex program (6.19) if there exists a dual variable $\Lambda \in \mathbb{S}^p$

with the following optimality conditions being satisfied:

$$\begin{aligned} [\Sigma_n - \hat{\Theta}^{-1}]_y + \Lambda = 0 \quad ; \quad & [\Sigma_n - \hat{\Theta}^{-1}]_y \in \lambda_n \partial \|\hat{L}_y\|_\star \\ [\Sigma_n - \hat{\Theta}^{-1}]_{yx} \in -\lambda_n \gamma \partial \|\hat{\Theta}_{yx}\|_\star \quad ; \quad & [\Sigma_n - \hat{\Theta}^{-1}]_x = 0 \\ \hat{\Theta}_y = \hat{D}_y - \hat{L}_y; \quad \hat{D}_y \text{ is diagonal} \quad ; \quad & \Lambda_{i,i} = 0 \text{ for } i = 1, 2, \ldots p \end{aligned}$$

Recall that elements of the subdifferential with respect to nuclear norm at a matrix M have the key property that they decompose with respect to the tangent space $T(M)$. Specifically, the subdifferential with respect to the nuclear norm at a matrix M with (reduced) SVD given by $M = UQV^T$ is as follows:

$$N \in \partial \|M\|_\star \Leftrightarrow \mathcal{P}_{T(M)}(N) = UV^T \,,\ \|\mathcal{P}_{T(M)^\perp}(N)\|_2 \leq 1,$$

where $\mathcal{P}$ denote a projection operator. Let us denote the subspace $\mathcal{W} \in \mathbb{S}^p$ as the set of diagonal matrices with nonnegative entries. Let SVD of $\hat{L}_y$ and $\hat{\Theta}_{yx}$ be given by $\hat{L}_y = \bar{U}\bar{Q}\bar{V}'$ and $\hat{\Theta}_{yx} = \breve{U}\breve{Q}\breve{V}'$ respectively, and $Z \triangleq (0, \quad \lambda_n \bar{U}\bar{V}', \quad -\lambda_n \gamma \breve{U}\breve{V}', \quad 0)$. Setting $\Lambda = [\Sigma_n - \hat{\Theta}^{-1}]_{Y,\text{off diagonal}}$, and letting $\mathbb{H} = \mathcal{W} \times T(\hat{L}_y) \times T(\hat{\Theta}_{yx}) \times \mathbb{S}^q$, the optimality conditions of (6.19) can be reduced to:

1. $\mathcal{P}_{\mathbb{H}} \mathcal{F}^\dagger (\Sigma_n - \hat{\Theta}^{-1}) = Z$
2. $\|\mathcal{P}_{T(\hat{L}_y)^\perp}(\Sigma_n - \hat{\Theta}^{-1})_y\|_2 < \lambda_n$; $\|\mathcal{P}_{T(\hat{\Theta}_{yx})^\perp}(\Sigma_n - \hat{\Theta}^{-1})_{yx}\|_2 < \lambda_n \gamma$

Our analysis proceeds by constructing variables $(\hat{\Theta}, \hat{D}_y, \hat{L}_y)$ that satisfy the optimality conditions specified above. Consider the optimization program (6.19) with additional (non-convex) constraints that L_y and Θ_{yx} belong to algebraic variety of low rank matrices specified by $L_y^\star$ and $\Theta_{yx}^\star$. While this new program is non-convex, it has a very interesting property that at the global optimal solution (and indeed at any locally optimal solution) $\hat{L}_y$ and $\hat{\Theta}_{yx}$ are smooth points of their respective algebraic varieties. This observation suggests that the Lagrange multipliers corresponding to the additional variety constraints belongs to $T(\hat{L}_y)^\perp$ and $T(\hat{\Theta}_{yx})^\perp$ respectively. We show under suitable conditions that $(\hat{\Theta}, \hat{D}_y, \hat{L}_y)$ also satisfy the second optimality condition of (6.19) corresponding to the tangent spaces $T(\hat{L}_y)^\perp$ and $T(\hat{\Theta}_{yx})^\perp$. Thus $(\hat{\Theta}, \hat{D}_y, \hat{L}_y)$ is a unique solution of (6.4) and as constructed, is algebraically consistent (i.e. $\text{rank}(\hat{L}_y) = \text{rank}(L_y^\star)$ and $\text{rank}(\hat{\Theta}_{yx}) = \text{rank}(\Theta_{yx}^\star)$)

Results Proved in the Supplementary Material

To ensure that the estimate $\hat{\Theta}$ is close to the population quantity $\Theta^\star$, the quantity $E = \hat{\Theta} - \Theta^\star$ must be small. Since the optimality conditions of (6.19) are stated

in terms of $\hat{\Theta}^{-1}$, we bound the deviation between $\hat{\Theta}^{-1}$ and $\Theta^{\star^{-1}}$. Specifically, the Taylor series expansion of $\hat{\Theta}^{-1}$ around $\Theta^\star$ is given by:

$$\hat{\Theta}^{-1} = (\Theta^\star + E)^{-1} = \Theta^{\star^{-1}} + \Theta^{\star^{-1}} E \Theta^{\star^{-1}} + R_{\Sigma^\star}(E),$$

where $R_{\Sigma^\star}(E) = \Sigma^\star \left[\sum_{k=2}^{\infty} (-E\Theta^\star)^k \right]$. Recalling that $\mathbb{I}^\star = \Theta^{\star^{-1}} \otimes \Theta^{\star^{-1}}$, we note that $\hat{\Theta}^{-1} - \Theta^{\star^{-1}} = \mathbb{I}^\star(E) + R_{\Sigma^\star}(E)$. In Section 6.2, we imposed assumptions 1, 2, and 3 in (6.15), (6.16), and (6.17) on $\mathbb{I}^\star$. These assumptions allow us to control $\mathbb{I}^\star(E)$ when E is restricted to certain directions. We bound the remainder term $R_{\Sigma^\star}(E)$ in Proposition 5 where E is restricted to live in a certain space. Specifically, consider the following constrained optimization program:

$$\begin{aligned}(\tilde{\Theta}, \tilde{D}_y, \tilde{L}_y) = \underset{\substack{\Theta \in \mathbb{S}^{q+p},\, \Theta \succ 0 \\ D_y, L_y \in \mathbb{S}^p}}{\operatorname{argmin}} \quad & -\ell(\Theta; \mathcal{D}_+^n) + \lambda_n [\|L_y\|_\star + \gamma \|\Theta_{yx}\|_\star]. \\ \text{s.t.} \quad & \Theta_y = D_y - L_y,\ (D_y, L_y, \Theta_{yx}, \Theta_x) \in \mathbb{H}' \end{aligned} \tag{6.20}$$

Here $\mathbb{H}' = \mathcal{W} \times T_y' \times T_{yx}' \times \mathbb{S}^q$, where T_y' is a subspace in $\mathbb{S}^p$, and T_{yx}' is a subspace in $\mathbb{R}^{p \times q}$. Let $\Delta = (\tilde{D}_y - D_y^\star, \tilde{L}_y - L_y^\star, \tilde{\Theta}_{yx} - \Theta_{yx}^\star, \tilde{\Theta}_x - \Theta_x^\star)$ denote the error in the estimated variables. Furthermore, let $\Delta_1 = \tilde{D}_y - D_y^\star$, $\Delta_2 = \tilde{L}_y - L_y^\star$ and so forth. In the following proposition, we bound the remainder term $R_{\Sigma^\star}(\mathcal{F}(\Delta))$ defined earlier.

Proposition 5. *Let* $\psi \triangleq \|\Theta^{\star^{-1}}\|_2$ *and* $C' = (3+\gamma)\psi$. *If* $\Phi_\gamma[\Delta] \leq \frac{1}{2C'}$, *then* $\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\mathcal{F}(\Delta))] \leq 2m\psi C'^2 \Phi_\gamma[\Delta]^2$.

Notice the bound on $R_{\Sigma^\star}(\mathcal{F}(\Delta))$ is dependent on the error term $\Phi_\gamma[\Delta]$. In the following proposition, we bound this error so that we can control the remainder term. Suppose that for $\alpha > 0$, $\beta \geq 2$, $\omega_y \in (0,1)$, and $\omega_{yx} \in (0,1)$, the Fisher information conditions (6.15), (6.16), and (6.17) are satisfied. Suppose we let T_y' and T_{yx}' be tangent spaces to the low-rank matrix varieties and $\rho(T_y', T(L_y^\star)) \leq \omega_y$ and $\rho(T_{yx}', T(\Theta_{yx}^\star)) \leq \omega_{yx}$. Let $E_n = \Sigma^\star - \Sigma_n$ denote the difference between the true joint covariance and the sample covariance and let $C_T = (\mathcal{P}_{T_y'^\perp}(L_y^\star), \mathcal{P}_{T_{yx}'^\perp}(\Theta_{yx}^\star))$. The proof of the following result uses Brouwer's fixed-point theorem, and is inspired by the proof of a similar result in [Rav+11a; CPW12].

Proposition 6. *Let* $\kappa \triangleq \beta(3 + \frac{16}{\alpha}\psi^2 m)$. *Consider the following two quantities:*

$$r_1 \triangleq \max\left\{ \frac{4}{\alpha}\Big(\Phi_\gamma[\mathcal{F}^\dagger E_n] + \Phi_\gamma[\mathcal{F}^\dagger \mathbb{I}^\star \mathcal{F} C_T] + \lambda_n \Big),\ \ \Phi_\gamma[C_T] \right\}, \tag{6.21}$$

$$r_2 \triangleq \frac{4}{\alpha}\Big(\Phi_\gamma[\mathcal{F}^\dagger E_n] + \Phi_\gamma[\mathcal{F}^\dagger \mathbb{I}^\star \mathcal{F} C_T] \Big) \tag{6.22}$$

Define $r_1^u \triangleq \max\left\{\frac{4}{\alpha}\left(\frac{2\lambda_n}{\kappa}+\lambda_n\right),\ \frac{\lambda_n}{\kappa}\right\}$ *and* $r_2^u \triangleq \frac{8\lambda_n}{\alpha\kappa}$. *Suppose that 1)* $r_1 \le r_1^u$, *2)* $r_2 \le r_2^u$, *and 3)* $r_1^u \le \min\{\frac{1}{4C'}, \frac{\alpha}{32\max\{1+\frac{\kappa}{2},\frac{\alpha}{8}\}^2 m\psi C'^2}\}$,
then $\max\{\|\Delta_2\|_2, \frac{1}{\gamma}\|\Delta_3\|_2\} \le 2r_1^u$ *and* $\max\{\|\Delta_1\|_2, \|\Delta_4\|_2\} \le r_2^u$. *Consequently,* $\Phi_\gamma(\Delta) \le 2r_1^u$.

In the following proposition, we prove algebraic correctness of program (6.20). The statement of this proposition requires us to define some constants. Let $C_1' = \frac{2\bar{m}m}{\kappa\alpha}\left(6\psi^2+\frac{5}{\alpha}\psi^2+\frac{46\psi^2\kappa}{\alpha}+\kappa\right)+\frac{1}{\psi^2}$, $C_2' = \frac{4}{\alpha}(\frac{1}{2\kappa}+1)$, $C'_{\sigma_y} = C_1'^2\psi^2\max\{2\kappa+1, \frac{2}{C_2'\psi^2}+1\}$, $C'_{\sigma_{yx}} = C_1'^2\psi^2\max\{2\kappa+\frac{\kappa}{\gamma}, \frac{2}{C_2'\psi^2}+\frac{\kappa}{\gamma}\}$, and
$C'_{samp} = \max\{\frac{1}{8m\psi\kappa}, \frac{\alpha}{16C'(\frac{2}{\kappa}+1)}, \frac{\alpha^2}{128(\frac{2}{\kappa}+1)\max\{1+\frac{\kappa}{2},\frac{\alpha}{8}\}^2 m\psi^2 C'^2}, \frac{1}{4C_1'C'}\}$.

Proposition 7. *Suppose that* $\sigma_y \ge \frac{m}{\omega}C'_{\sigma_y}\lambda_n$, $\sigma_{yx} \ge m\gamma^2 C'_{\sigma_{yx}}\lambda_n$. *Further, suppose that* λ_n *is chosen so that* $\lambda_n \le \frac{1}{C'_{samp}}$. *Then, there exists tangent space* $T_y' \subset \mathbb{S}^p$ *in the rank-*k_u *variety (*$k_u = rank(L_y^\star)$*) and tangent space* $T'_{yx} \subset \mathbb{R}^{p\times q}$ *in rank* k_x*-variety (*$k_x = rank(\Theta_{yx}^\star)$*) where* $\rho(T_y', T(L_y^\star)) \le \omega_y$, $\rho(T'_{yx}, T(\Theta_{yx}^\star)) \le \omega_{yx}$ *such that the corresponding solution* $(\tilde{\Theta}, \tilde{D}_y, \tilde{L}_y)$ *of* (6.20) *satisfies the following properties:*

1. $rank(\tilde{L}_y) = rank(L_y^\star)$ *and* $rank(\tilde{\Theta}_{yx}) = rank(\Theta_{yx}^\star)$
2. *Letting* $C_T = (0\,,\ \mathcal{P}_{T_y'^\perp}(L_y^\star)\,,\ \mathcal{P}_{T_{yx}'^\perp}(\Theta_{yx}^\star)\,,\ 0)$, *we have that* $\Phi_\gamma[\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}(C_T)] \le \frac{\lambda_n}{\kappa}$ *and* $\Phi_\gamma[C_T] \le \frac{4}{\alpha}(1+\frac{2}{\kappa})\lambda_n$
3. $\Phi_\gamma[\Delta] \le 2C_1'\lambda_n$
4. $\tilde{L}_y \succeq 0$

Furthermore, suppose that $\Phi_\gamma(\mathcal{F}^\dagger E_n) \le \frac{\lambda_n}{\kappa}$ *and* $\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\mathcal{F}(\Delta))] \le \frac{\lambda_n}{\kappa}$. *Then the tangent space constraint* $(D_y, L_y, \Theta_{yx}, \Theta_x) \in \mathbb{H}'$ *in* (6.20) *is inactive, so that* $(\tilde{\Theta}, \tilde{D}_y, \tilde{L}_y)$ *is the unique solution of the original convex program* (6.4).

Thus far, the analysis of the convex program so has been deterministic in nature. In the following proposition, we present the probabilistic component of our analysis by showing the rate at which the sample covariance matrix Σ_n converges to $\Sigma^\star$ in spectral norm. This result is well-known and is a specialization of a result proven by [DS01].

Proposition 8. *Suppose that the number of observed samples obeys* $n \ge 64\kappa^2 m^2\psi^2 C'^2_{samp}(p+q)$, *and the regularization parameter* λ_n *is chosen so that:*

$\lambda_n \in [8\psi\kappa m\sqrt{\frac{p+q}{n}}, \frac{1}{C'_{samp}}]$. *Then, with probability greater than* $1-2exp\Big\{-\frac{n\lambda_n^2}{128\kappa^2 m^2\psi^2}\Big\}$, $\Phi_\gamma[\mathcal{F}^\dagger E_n] \leq \frac{\lambda_n}{\kappa}$.

Proof of Theorem 5

We first relate the constants $\tilde{C}_{samp}$, $\tilde{C}$, $\tilde{C}_0$, $\tilde{C}_1$, and $\tilde{C}_\sigma$ of Theorem 5 to the constants C'_{samp}, C'_1 C'_{σ_y}, and $C'_{\sigma_{yx}}$. In particular, using the properties that $\beta \geq 2$ and $\frac{\psi^2}{\alpha} \geq \frac{1}{2}$ and $\bar{m}, m \geq 1$, one can check that: $\tilde{C}_0 \geq \frac{1}{\beta m}C'_{samp}$, $\tilde{C}_\sigma \geq \frac{\alpha^5}{\beta m^3\bar{m}}C'_{\sigma_y}$, $\tilde{C}_\sigma \geq \frac{\alpha^5}{\beta m^4\bar{m}}C'_{\sigma_{yx}}$, and $\tilde{C}_1 \geq \frac{\alpha^2}{m\bar{m}}C'_1$. Furthermore, we have that $\tilde{C} \geq \frac{\alpha}{\beta m}8\psi\kappa$. Using these relations, one can also check that the assumptions of Theorem 5 imply that the assumptions of Proposition 7 and Proposition 4 are satisfied. Thus we can conclude that the optimal solution $(\tilde{\Theta}, \tilde{D}_y, \tilde{L}_y)$ of (6.20) (with a particular choice of tangent spaces T'_y and T'_{yx}) satisfy results of Proposition 7. Further, by appealing to Proposition 8, we have that $\Phi_\gamma(\mathcal{F}^\dagger E_n) \leq \frac{\lambda_n}{\kappa}$. If we show that $\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\Delta)] \leq \frac{\lambda_n}{\kappa}$, then we conclude that the unique optimum $(\hat{\Theta}, \hat{D}_y, \hat{L}_y)$ of the original convex program (6.4) coincide with the optimum $(\tilde{\Theta}, \tilde{D}_y, \hat{L}_y)$ of the convex program (6.20). Thus, we conclude that the estimates of (6.4) have structurally correct structure (i.e. $\text{rank}(\hat{L}_y) = \text{rank}(L_y^\star)$ and $\text{rank}(\hat{\Theta}_{yx}) = \text{rank}(\Theta_{yx}^\star)$) and have their error bounded by $\Phi_\gamma(\Delta) \leq 2C'_1\lambda_n$. To show that $\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\Delta)] \leq \frac{\lambda_n}{\kappa}$, we note that:

$$\frac{4}{\alpha}\Big(\Phi_\gamma[\mathcal{F}^\dagger E_n] + \Phi_\gamma[\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}C_T] + \lambda_n\Big) \leq \frac{4}{\alpha}\Big(\frac{\lambda_n}{\kappa} + \frac{\lambda_n}{\kappa} + \lambda_n\Big) \leq \frac{4\lambda_n}{\alpha}\Big(\frac{2}{\kappa} + 1\Big)$$
$$\leq \min\Big\{\frac{1}{4C'}, \frac{\alpha}{32\max\{1+\frac{\kappa}{2}, \frac{\alpha}{8}\}^2 m\psi C'^2}\Big\}.$$

Here, we used the bound on $\Phi_\gamma[\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}C_T]$ provided by Proposition 7 and the bound on λ_n. Furthermore, appealing to Proposition 7 once again, we have $\Phi_\gamma[C_T] \leq \frac{4}{\alpha}(1+\frac{2}{\kappa})\lambda_n \leq \min\{\frac{1}{4C'}, \frac{\alpha}{16m\psi C'^2}\}$. Thus Proposition 6 provides us with the bound $\Phi_\gamma[\Delta] \leq 2C'_1\lambda_n \leq \frac{1}{2C'}$. We subsequently apply the results of Proposition 5 to obtain:

$$\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\mathcal{F}(\Delta))] \leq 2m\psi C'^2\Phi_{\delta,\gamma}[\Delta]^2 \leq \Big[2m\psi C'^2 C'^2_1\lambda_n\Big]\lambda_n \leq \frac{\lambda_n}{\kappa}.$$

The last inequality follows from the bound on λ_n.

Chapter 7

CONCLUSION

The central theme of this thesis is to provide solutions to address some of the challenges that arise in latent variable modeling and the utility of the proposed approaches on real applications. Here we describe the main contributions, and discuss some future research directions.

7.1 Summary of Contributions

Modeling California Reservoir Network: In Chapter 2, we developed a state-wide statistical graphical model to characterize the dependencies among a collection of 55 major California reservoirs across the state. We obtained and validated this model in a data-driven manner based on reservoir volumes over the period 2003˘2016. A key feature of our framework is a quantification of the effects of external phenomena that influence the entire reservoir network. We further characterized the degree to which physical factors (e.g., state-wide Palmer Drought Severity Index (PDSI), average temperature, snow pack, etc.) and economic factors (e.g., consumer price index, number of agricultural workers, etc.) explain these external influences. As a consequence of this analysis, we obtained a system-wide health diagnosis of the reservoir network as a function of PDSI.

This work subsequently motivated the following methodological advancements.

False Discovery Control in Low-rank Estimation: In many applications, such as modeling the reservoir volumes, the row/column space structure of a low-rank matrix carries information about some underlying phenomenon, and it is of interest in inferential settings to evaluate the extent to which the row/column spaces of an estimated low-rank matrix signify discoveries about the phenomenon. In Chapter 3, We developed a *geometric* reformulation of the concept of a discovery, which then enabled a natural definition in the low-rank case. We described and analyzed a generalization of the Stability Selection method of Meinshausen and Bühlmann to control for false discoveries in low-rank estimation, and we demonstrated its utility compared to previous approaches via numerical experiments.

Latent Variable Graphical Modeling: Beyond Gaussianity: The latent-variable

graphical modeling technique employed in Chapter 2 is relevant for Gaussian variables. In many applications of interest, e.g., in computational biology, the random variables of interest are fundamentally non-Gaussian. As such, we described in Chapter 4 a convex optimization framework for fitting latent-variable graphical models in the class of generalized linear models. The fitting algorithm was based on generalizing neighborhood selection technique of [MB06] by stitching separate node-wise optimization problems into one program that is parameterized by a sparse matrix (encoding variable dependencies) and a low-rank matrix (encoding latent effect). We demonstrated the utility of this algorithm on the number of real-world applications involving gnomic data and voter-records data.

Latent variable Model Selection with non-iid Data: A common challenge in latent-variable modeling is that the observed data is not i.i.d and consists of strong dependencies, e.g., seasonal and monthly dependencies in reservoir volumes of Chapter 2. To address this challenge, we developed the factorizable precision operator framework — that combined ideas from Stochastic PDE's and high-dimensional covariance selection — to efficiently obtain high dimensional models with rich dependencies across observations. We demonstrated the utility of our approach for signature detection in hyperspectral imaging.

Interpreting Latent Variables via Convex Optimization: The latent variables obtained from a latent-variable model are mathematical objects, without any physical interpretation. In Chapter 6, we proposed an approach to provide semantics to latent variables. Specifically, our approach is to measure auxiliary variables (motivated by domain knowledge), and solving a convex optimization program that links these additional variables to the latent variables. We further provided theoretical support for the utility of this algorithm.

7.2 Future Directions

Latent-Variable Modeling via Lift-and-Project: Lift-and-project is a powerful framework in mathematical optimization for designing relaxations of intractable combinatorial problems. It is based on the idea of describing a complicated set as a projection of a concisely parameterized convex set in a lifted (higher-dimensional) space. Latent-variable modeling techniques in data analysis can be viewed from this perspective, where the lifting step is the introduction of additional variables (latent variables) that leads to more concisely parameterized models in higher dimensions and the projection step is the marginalization of the latent variables. The lift-

and-project literature provides a machinery — via an appeal to ideas from real algebraic geometry such as sums-of-squares, moment theory, and functional analysis — to systematically and tractably generate a hierarchy of latent-variable model approximations. It will be exciting to explore this lift-and-project method to develop novel and computationally tractable techniques for latent-variable modeling.

Scientific Application of Low-rank False Discovery Control: In Chapter 3, we applied the proposed geometric false discovery framework on the application of hyperspectral imaging to control the size of false discoveries in signature detection. It would be of interest to apply this geometric perspective to other scientific applications where the row/column spaces of a low-rank matrix represent discovery of a physical phenomena. As an example, in biology, principal components analysis (PCA) is often used to extract a set of relevant directions in gene expression space for biological variation. These directions are claimed as discoveries, and thus it would be interesting to utilize our false discovery framework to assess the accuracy of these findings.

Optimal Regularization with Respect to False Discovery Control: Regularization techniques are ubiquitous in low-rank estimation. These depend on a regularization parameter that is typically tuned based on cross-validation to optimize for prediction performance. In scenarios where the objective is that of discovery, how do we prescribe a choice of regularization parameter to control for false discoveries while maximizing on power? Focusing on the matrix denoising setting, [HT14] prescribe an exact formula for the regularization parameter based on least squares error. In this setting, it would be interesting to determine a choice of regularization parameter that optimizes for false discovery control.

Statistical Confidence via Lattice Theory: Much of the literature on assessing and controlling false discoveries in statistics and signal processing has been limited to settings in which one has a (possibly large) collection of binary hypotheses and one wishes to identify a subset of these as discoveries from observations. Many modeling frameworks that are widely employed in contemporary data analysis do not fit this setting – e.g., ranking, causal discovery – and therefore new tools are needed to evaluate and control appropriate notions of false discovery in these cases. I believe that a fruitful approach to addressing this challenge is via lattice theory. Specifically, identifying a subset of discoveries in variable selection or multiple testing may be viewed as selecting a partial-ordered set of the Boolean lattice. Similarly, finding a row and column space pair in low-rank estimation selects an

element from the subspace lattice. Building on these conceptual connections, it would be interesting to identify appropriate lattices underlying decision spaces in problems such as ranking and causal discovery. Once the geometry of these lattices is well-understood, the existing techniques in the false discovery literature can be brought to fruition in these new domains.

Appendix A

PROOFS OF CHAPTER 3

Proof of Theorem 4

Proof. We first show that

$$\begin{aligned}\mathbb{E}\left[\text{trace}\left(\mathcal{P}_T\mathcal{P}_{T^{\star\perp}}\right)\right] &\leq \sum_{i=1}^{\dim(T^{\star\perp})} \mathbb{E}\left[\left\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}(M_i)\right\|_F\right]^2 + 4\sqrt{1-\alpha}\kappa_{\text{bag}} \\ &+ 2(1-\alpha)\mathbb{E}[\dim(T)].\end{aligned}$$

Let $\{M_i\}_{i=1}^{\dim(T^{\star\perp})}$ be a set of orthonormal basis elements for $T^{\star\perp}$. Then for any i and tangent space $\hat{T}(\mathcal{D}_j)$ estimated on a subsample, we have that

$$\begin{aligned}\text{trace}\left(\mathcal{P}_T\mathcal{P}_{\text{span}(M_i)}\right) = &\quad \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)}\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_j)}\mathcal{P}_{\text{span}(M_i)}\right) \\ &+ \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\mathcal{P}_{\text{span}(M_i)}\right) \\ &+ \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)}\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\mathcal{P}_{\text{span}(M_i)}\right) \\ &+ \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_j)}\mathcal{P}_{\text{span}(M_i)}\right)\end{aligned} \tag{A.1}$$

With some manipulations, we obtain

$$\begin{aligned}\text{trace}\left(\mathcal{P}_T\mathcal{P}_{\text{span}(M_i)}\right) \leq &\quad \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)}\mathcal{P}_{\text{span}(M_i)}\right) \\ &+ \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\mathcal{P}_{\text{span}(M_i)}\right) \\ &+ \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)}\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\mathcal{P}_{\text{span}(M_i)}\right) \\ &+ \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_j)}\mathcal{P}_{\text{span}(M_i)}\right),\end{aligned} \tag{A.2}$$

where the inequality is due the property that $\text{trace}(AB) \leq \text{trace}(A)\|B\|_2$ for $A \succeq 0$ and that the spectral norm of a projection operator is equal to one. As $\mathcal{D}_j$ was

arbitrary, we can minimize over the entire collection as follows:

$$\begin{aligned}
\text{trace}\left(\mathcal{P}_T\mathcal{P}_{T^{\star\perp}}\right) &\leq \sum_{i=1}^{\dim(T^{\star\perp})} \min_{j=1,2,\ldots,B/2} \min_{k=\{0,1\}} \Bigg\{ \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})}\mathcal{P}_{\text{span}(M_i)}\right) \\
&+ \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\mathcal{P}_{\text{span}(M_i)}\right) \\
&+ \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)}\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\mathcal{P}_{\text{span}(M_i)}\right) \\
&+ \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_j)}\mathcal{P}_{\text{span}(M_i)}\right) \Bigg\} \qquad \text{(A.3)} \\
&\leq \text{Term 1} + \text{Term 2} + \text{Term 3},
\end{aligned}$$

where

$$\begin{aligned}
\text{Term 1} &= \sum_{i=1}^{\dim(T^{\star\perp})} \frac{2}{B} \sum_{j=1}^{B/2} \min_{k=\{0,1\}} \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})}\mathcal{P}_{\text{span}(M_i)}\right) \\
\text{Term 2} &= \sum_{i=1}^{\dim(T^{\star\perp})} \frac{2}{B} \sum_{j=1}^{B/2} \sum_{k=\{0,1\}} \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})^\perp}\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})^\perp}\mathcal{P}_{\text{span}(M_i)}\right) \\
\text{Term 3} &= \sum_{i=1}^{\dim(T^{\star\perp})} \frac{2}{B} \sum_{j=1}^{B/2} \sum_{k=\{0,1\}} \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)}\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\mathcal{P}_{\text{span}(M_i)}\right) \\
&+ \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_j)}\mathcal{P}_{\text{span}(M_i)}\right)
\end{aligned}$$

The first inequality follows from (A.2) holding for every j and that $\mathcal{P}_{T^{\star\perp}} = \sum_{i=1}^{\dim(T^{\star\perp})} \mathcal{P}_{\text{span}(M_i)}$. The second inequality follows from the property that $\min\{a + b, c + d\} \leq \min\{a, c\} + b + d$ and that the minimum over a collection is bounded above by the average of the collection. We begin by bounding Term 1. Notice that

$$\begin{aligned}
\text{Term 1} &= \sum_{i=1}^{\dim(T^{\star\perp})} \frac{2}{B} \sum_{j=1}^{B/2} \min_{k=\{0,1\}} \left\| \mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})}\mathcal{P}_{\text{span}(M_i)} \right\|_F^2 \\
&= \sum_{i=1}^{\dim(T^{\star\perp})} \frac{2}{B} \sum_{j=1}^{B/2} \min_{k=\{0,1\}} \left\| \mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})}(M_i) \right\|_F^2 \\
&\leq \sum_{i=1}^{\dim(T^{\star\perp})} \frac{2}{B} \sum_{j=1}^{B/2} \prod_{k=\{0,1\}} \left\| \mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})}(M_i) \right\|_F
\end{aligned}$$

where the inequality follows from the fact that the minimum of two positive quantities is bounded above the product of their square roots. Bounding Term 2, we have:

$$\begin{aligned}
\text{Term 2} &= \sum_{i=1}^{\dim(T^{\star\perp})} \frac{2}{B} \sum_{j=1}^{B/2} \sum_{k=\{0,1\}} \operatorname{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})^{\perp}} \mathcal{P}_T \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})^{\perp}} \mathcal{P}_{\operatorname{span}(M_i)}\right) \\
&= \frac{2}{B} \sum_{j=1}^{B/2} \sum_{k=\{0,1\}} \operatorname{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})^{\perp}} \mathcal{P}_T \mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})^{\perp}} \mathcal{P}_{T^{\star\perp}}\right) \\
&\leq \frac{2}{B} \sum_{j=1}^{B} \operatorname{trace}\left(\mathcal{P}_T \mathcal{P}_{\hat{T}(\mathcal{D}_i)^{\perp}}\right) = 2 \operatorname{trace}\left(\mathcal{P}_T(\mathcal{I} - \mathcal{P}_{\mathsf{avg}})\mathcal{P}_T\right) \leq 2(1-\alpha)\dim(T).
\end{aligned}$$

The first inequality follows from $\operatorname{trace}(AB) \leq \operatorname{trace}(A)\|B\|_2$ for $A \succeq 0$ and that projection operators have spectral norm equal to one. The second inequality follows from the fact that $T \in \mathcal{T}_\alpha$. Finally, we consider Term 3:

$$\begin{aligned}
\text{Term 3} &\overset{(a)}{=} \sum_{i=1}^{\dim(T^{\star\perp})} \frac{2}{B} \sum_{j=1}^{B} \left[\operatorname{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)} \mathcal{P}_{\operatorname{span}(M_i)} \mathcal{P}_{\hat{T}(\mathcal{D}_j)^{\perp}} \mathcal{P}_T\right)\right. \\
&\quad + \left. \operatorname{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)^{\perp}} \mathcal{P}_{\operatorname{span}(M_i)} \mathcal{P}_{\hat{T}(\mathcal{D}_j)} \mathcal{P}_T\right)\right] \\
&\overset{(b)}{=} \frac{2}{B} \sum_{j=1}^{B} \left[\operatorname{trace}\left(\left[\mathcal{P}_{\hat{T}(\mathcal{D}_j)} \mathcal{P}_{T^{\star\perp}} - \mathcal{P}_{T^{\star\perp}} \mathcal{P}_{\hat{T}(\mathcal{D}_j)}\right] \mathcal{P}_{\hat{T}(\mathcal{D}_j)^{\perp}} \mathcal{P}_T\right)\right. \\
&\quad + \left. \operatorname{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)^{\perp}} \left[\mathcal{P}_{T^{\star\perp}} \mathcal{P}_{\hat{T}(\mathcal{D}_j)} - \mathcal{P}_{\hat{T}(\mathcal{D}_j)} \mathcal{P}_{T^{\star\perp}}\right] \mathcal{P}_T\right)\right] \\
&\overset{(c)}{\leq} \min\left\{\frac{4}{B} \sum_{j=1}^{B} \|\mathcal{P}_{\hat{T}(\mathcal{D}_j)^{\perp}} \mathcal{P}_T\|_F \|[\mathcal{P}_{\hat{T}(\mathcal{D}_j)}, \mathcal{P}_{T^{\star\perp}}]\|_F\right. \\
&\quad , \left. \frac{4}{B} \sum_{j=1}^{B} \|\mathcal{P}_{\hat{T}(\mathcal{D}_j)^{\perp}} \mathcal{P}_T\|_\star \|[\mathcal{P}_{\hat{T}(\mathcal{D}_j)}, \mathcal{P}_{T^{\star\perp}}]\|_2\right\} \\
&\overset{(d)}{\leq} 4 \min\left\{\sqrt{\frac{1}{B} \sum_{j=1}^{B} \|\mathcal{P}_{\hat{T}(\mathcal{D}_j)^{\perp}} \mathcal{P}_T\|_F^2} \sqrt{\frac{1}{B} \sum_{j=1}^{B} \|[\mathcal{P}_{\hat{T}(\mathcal{D}_j)}, \mathcal{P}_{T^{\star\perp}}]\|_F^2}\right. \\
&\quad , \left. \sqrt{\frac{1}{B} \sum_{j=1}^{B} \|\mathcal{P}_{\hat{T}(\mathcal{D}_j)^{\perp}} \mathcal{P}_T\|_\star^2} \sqrt{\frac{1}{B} \sum_{j=1}^{B} \|[\mathcal{P}_{\hat{T}(\mathcal{D}_j)}, \mathcal{P}_{T^{\star\perp}}]\|_2^2}\right\} \\
&\overset{(e)}{\leq} 4\sqrt{1-\alpha} \min\left\{\sqrt{\dim(T)} \sqrt{\frac{1}{B} \sum_{j=1}^{B} \|[\mathcal{P}_{\hat{T}(\mathcal{D}_j)}, \mathcal{P}_{T^{\star\perp}}]\|_F^2}\right. \\
&\quad , \left. \dim(T) \sqrt{\frac{1}{B} \sum_{j=1}^{B} \|[\mathcal{P}_{\hat{T}(\mathcal{D}_j)}, \mathcal{P}_{T^{\star\perp}}]\|_2^2}\right\}
\end{aligned}$$

Here $\overset{(a)}{=}$ follows from cyclicity of trace function; $\overset{(b)}{=}$ follows from the fact that $\sum_{i=1}^{\dim(T^{\star\perp})} \mathcal{P}_{\text{span}(M_i)} = \mathcal{P}_{T^{\star\perp}}$, $\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\mathcal{P}_{\hat{T}(\mathcal{D}_j)} = 0$ and the idempotence of projection operators; $\overset{(c)}{\leq}$ follows from the inequality $\text{trace}(AB) \leq \min\{\|A\|_F\|B\|_F, \|A\|_\star\|B\|_2\}$; $\overset{(d)}{\leq}$ follows from Cauchy-Shwarz inequality; and $\overset{(e)}{\leq}$ follows from $T \in \mathcal{T}_\alpha$ and that $\|A\|_\star \leq \|A\|_F\sqrt{\text{rank}(A)}$. Putting all the terms together and taking expectation yields:

$$\begin{aligned}
\mathbb{E}\left[\text{trace}\left(\mathcal{P}_T\mathcal{P}_{T^{\star\perp}}\right)\right] &\leq \mathbb{E}\left[\sum_{i=1}^{\dim(T^{\star\perp})} \frac{2}{B}\sum_{j=1}^{B/2}\prod_{k=\{0,1\}}\left\|\mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})}(M_i)\right\|_F\right] \\
&+ 4\sqrt{1-\alpha}\kappa_{\text{bag}} + 2(1-\alpha)\mathbb{E}[\dim(T)] \\
&= \sum_{i=1}^{\dim(T^{\star\perp})}\left[\mathbb{E}\left\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}(M_i)\right\|_F\right]^2 \\
&+ 4\sqrt{1-\alpha}\kappa_{\text{bag}} + 2(1-\alpha)\mathbb{E}[\dim(T)]
\end{aligned}$$

where the equality follows from observing that the quantities $\|\mathcal{P}_{\hat{T}(D_{2j-1})}(M_i)\|_F$ and $\|\mathcal{P}_{\hat{T}(D_{2j})}(M_i)\|_F$ are statistically independent due to complementary partitioning and noting that $\mathcal{P}_{\hat{T}(D_j)}(M_i)$ is identically distributed for all j.

We next show that

$$\begin{aligned}
\mathbb{E}\left[\text{trace}\left(\mathcal{P}_T\mathcal{P}_{T^{\star\perp}}\right)\right] &\leq \mathbb{E}\left[\text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\mathcal{P}_{T^{\star\perp}}\right)^{1/2}\right]^2 + 4\sqrt{1-\alpha}\kappa_{\text{bag}} \\
&+ 2(1-\alpha)\mathbb{E}[\dim(T)] + 2(1-\alpha)\mathbb{E}[\dim(T)].
\end{aligned}$$

To prove this relation, note that:

$$\begin{aligned}
\text{trace}\left(\mathcal{P}_T\mathcal{P}_{T^{\star\perp}}\right) &= \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)}\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_j)}\mathcal{P}_{T^{\star\perp}}\right) + \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\mathcal{P}_{T^{\star\perp}}\right) \\
&+ \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)}\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\mathcal{P}_{T^{\star\perp}}\right) + \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_j)}\mathcal{P}_{T^{\star\perp}}\right)
\end{aligned}$$

Using a similar logic as before, one can show

$$\begin{aligned}
\text{trace}\left(\mathcal{P}_T\mathcal{P}_{T^{\star\perp}}\right) &\leq \frac{2}{B}\sum_{j=1}^{B/2}\min_{k=\{0,1\}}\text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})}\mathcal{P}_{T^{\star\perp}}\right) + \text{Term 2} + \text{Term 3} \\
&\leq \frac{2}{B}\sum_{j=1}^{B/2}\prod_{k=\{0,1\}}\text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})}\mathcal{P}_{T^{\star\perp}}\right)^{1/2} + \text{Term 2} + \text{Term 3}
\end{aligned}$$

Taking expectations once again gives us the desired result. □

Proof of Variable Selection Bound

Proof. From decomposition (A.3), we have:

$$\begin{aligned}
\text{trace}\left(\mathcal{P}_T\mathcal{P}_{\text{span}(M_i)}\right) &\leq \min_{j=1,2,\ldots,B/2}\min_{k=\{0,1\}}\Bigg\{\text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})}\mathcal{P}_{\text{span}(M_i)}\right) \\
&+ \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\mathcal{P}_{\text{span}(M_i)}\right) \\
&+ \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)}\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\mathcal{P}_{\text{span}(M_i)}\right) \\
&+ \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_j)}\mathcal{P}_{\text{span}(M_i)}\right)\Bigg\} \\
&\leq \text{Term 1} + \text{Term 4}, \qquad \text{(A.4)}
\end{aligned}$$

where

$$\begin{aligned}
\text{Term 1} &= \min_{j=1,2,\ldots,B/2}\min_{k=\{0,1\}}\text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})}\mathcal{P}_{\text{span}(M_i)}\right) \\
&+ \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})^\perp}\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})^\perp}\mathcal{P}_{\text{span}(M_i)}\right) \\
\text{Term 4} &= \sum_{j=1}^{B}\text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)}\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\mathcal{P}_{\text{span}(M_i)}\right) \\
&+ \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_j)}\mathcal{P}_{\text{span}(M_i)}\right)
\end{aligned}$$

The second inequality follows from $\min\{a+b, c+d\} \leq \min\{a, c\} + b + d$. Since the projection operators $\mathcal{P}_{\hat{T}(\mathcal{D}_j)}, \mathcal{P}_{\text{span}(M_i)}, \mathcal{P}_T$ commute in variable selection, Term 4 = 0. Furthermore,

$$\begin{aligned}
&\text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})^\perp}\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})^\perp}\mathcal{P}_{\text{span}(M_i)}\right) \\
\overset{(a)}{=}\ &\text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})^\perp}\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})^\perp}\mathcal{P}_{\text{span}(M_i)}\mathcal{P}_{\text{span}(M_i)}\right) \\
\overset{(b)}{=}\ &\text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})^\perp}\mathcal{P}_T\mathcal{P}_{\text{span}(M_i)}\mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})^\perp}\mathcal{P}_{\text{span}(M_i)}\right) \\
=\ &\text{trace}\left(\mathcal{P}_{\text{span}(\mathcal{P}_T(M_i))}\mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})^\perp}\mathcal{P}_{\text{span}(\mathcal{P}_T(M_i))}\right)\|\mathcal{P}_T(M_j)\|_F^2 \\
\overset{(c)}{=}\ &\text{trace}\left(\mathcal{P}_{\text{span}(\mathcal{P}_T(M_i))}\mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})^\perp}\mathcal{P}_{\text{span}(\mathcal{P}_T(M_i))}\right)
\end{aligned}$$

Here $\overset{(a)}{=}$ follows idempotence of projection operators; $\overset{(b)}{=}$ follows from commutativity of the projection operators; and $\overset{(c)}{=}$ follows from the fact that $\|\mathcal{P}_T(M_j)\|_F = \{0, 1\}$.

Plugging this finding into (A.4), we obtain:

$$\begin{aligned}\operatorname{trace}\left(\mathcal{P}_T\mathcal{P}_{\operatorname{span}(M_i)}\right) \quad &\leq \quad \min_{j=1,2,\ldots,B/2}\min_{k=\{0,1\}}\left\{\operatorname{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})}\mathcal{P}_{\operatorname{span}(M_i)}\right)\right. \\ &+ \quad \left.\operatorname{trace}\left(\mathcal{P}_{\operatorname{span}(\mathcal{P}_T(M_i))}\mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})^\perp}\mathcal{P}_{\operatorname{span}(\mathcal{P}_T(M_i))}\right)\right\} \quad \text{(A.5)}\end{aligned}$$

Since M_i is a standard basis element and $\mathcal{P}_T$ is diagonalized by standard basis elements, $\mathcal{P}_{\operatorname{span}(\mathcal{P}_T(M_i))}$ is again be diagonalized by standard basis elements and thus $\mathcal{P}_{\operatorname{span}(\mathcal{P}_T(M_i))}$ commutes with $\mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})^\perp}$. Hence, it follows immediately that $\operatorname{trace}\left(\mathcal{P}_{\operatorname{span}(\mathcal{P}_T(M_i))}\mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})^\perp}\mathcal{P}_{\operatorname{span}(\mathcal{P}_T(M_i))}\right) = \{0,1\}$. This leads to further bounding of false discovery:

$$\begin{aligned}\operatorname{trace}\left(\mathcal{P}_T\mathcal{P}_{\operatorname{span}(M_i)}\right) \quad &\leq \quad \min_{j=1,2,\ldots,B/2}\min_{k=\{0,1\}}\left\{\operatorname{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})}\mathcal{P}_{\operatorname{span}(M_i)}\right)\right. \\ &+ \quad \left.\operatorname{trace}\left(\mathcal{P}_{\operatorname{span}(\mathcal{P}_T(M_i))}\mathcal{P}_{\hat{T}(\mathcal{D}_{2i-k})^\perp}\mathcal{P}_{\operatorname{span}(\mathcal{P}_T(M_i))}\right)\right\} \\ &\overset{(a)}{\leq} \quad \min_{j=1,2,\ldots,B/2}\frac{\min_{k\in\{0,1\}}\left\|\mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})}\mathcal{P}_{\operatorname{span}(M_i)}\right\|_F^2}{\min_{k\in\{0,1\}}\operatorname{trace}\left(\mathcal{P}_{\operatorname{span}(\mathcal{P}_T(M_i))}\mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})}\mathcal{P}_{\operatorname{span}(\mathcal{P}_T(M_i))}\right)} \\ &\overset{(b)}{\leq} \quad \frac{\frac{2}{B}\sum_{j=1}^{B/2}\prod_{k\in\{0,1\}}\left\|\mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})}\mathcal{P}_{\operatorname{span}(M_i)}\right\|_F}{\frac{2}{B}\sum_{j=1}^{B}\prod_{k\in\{0,1\}}\operatorname{trace}\left(\mathcal{P}_{\operatorname{span}(\mathcal{P}_T(M_i))}\mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})}\mathcal{P}_{\operatorname{span}(\mathcal{P}_T(M_i))}\right)} \\ &\overset{(c)}{\leq} \quad \frac{\frac{2}{B}\sum_{j=1}^{B/2}\prod_{k\in\{0,1\}}\left\|\mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})}\mathcal{P}_{\operatorname{span}(M_i)}\right\|_F}{\frac{2}{B}\sum_{j=1}^{B}\sum_{k\in\{0,1\}}\operatorname{trace}\left(\mathcal{P}_{\operatorname{span}(\mathcal{P}_T(M_i))}\mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})}\mathcal{P}_{\operatorname{span}(\mathcal{P}_T(M_i))}\right)-1} \\ &\overset{(d)}{\leq} \quad \frac{\frac{2}{B}\sum_{j=1}^{B/2}\prod_{k\in\{0,1\}}\left\|\mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})}\mathcal{P}_{\operatorname{span}(M_i)}\right\|_F}{2\alpha-1}\end{aligned}$$

Here (a) follows from the property that $\min\{a+1-b, c+1-d\} \leq \frac{\min\{a,c\}}{\min\{b,d\}}$ for $a,b,c,d \in \{0,1\}$ with $\frac{0}{0}=1$, (b) uses the fact that $\min\{a,b\} = ab$ for $a,b \in \{0,1\}$ and the inequality $\min\{\frac{a}{b},\frac{c}{d}\} \leq \frac{a+c}{b+d}$ for $a,b,c,d \geq 0$, (c) is from $ab \geq a+b-1$ for $a,b \in (0,1)$, and (d) uses the fact that $\sigma_{\min}(\mathcal{P}_T\mathcal{P}_{\mathsf{avg}}\mathcal{P}_T) \geq \alpha$. Summing over all M_i, taking expectations, and using the fact that $\left\|\mathcal{P}_{\hat{T}(\mathcal{D}_{2j-1})}\mathcal{P}_{\operatorname{span}(M_i)}\right\|_F$ and $\left\|\mathcal{P}_{\hat{T}(\mathcal{D}_{2j})}\mathcal{P}_{\operatorname{span}(M_i)}\right\|_F$ are statistically independent and that $\mathcal{P}_{\hat{T}(\mathcal{D}_j)}$ are identically distributed for all j yields the desired result.

□

When are Assumptions 1 and 2 in (3.6) Satisfied?

Are there reasonable estimators and models in the low-rank setting that satisfy Assumptions 1 and 2 in (3.14) (main paper)? Assumption 1 is rather benign. Specifically, consider the following lemma whose proof we omit.

Lemma 1. *Let $k \leq \min\{p_1, p_2\}$ be fixed. Let $U \in \mathbb{R}^{p_1 \times k}$ and $V \in \mathbb{R}^{p_2 \times k}$ be drawn respectively from a Haar measure on the Steifel Manifold. Then the tangent space $\hat{T} = T(\text{span}(U), \text{span}(V))$ satisfies the following condition:*

$$\frac{\mathbb{E}\left[trace\left(\mathcal{P}_{T^{\star\perp}}\mathcal{P}_{\hat{T}}\right)\right]}{dim(T^{\star\perp})} = \frac{\mathbb{E}\left[trace\left(\mathcal{P}_{T^{\star}}\mathcal{P}_{\hat{T}}\right)\right]}{dim(T^{\star})}$$

In other words, as long as the low-rank estimator is better than the random selection procedure described in Lemma 1, Assumption 1 is satisfied. Assumption 2 on the other hand, is more stringent, although it is fulfilled in some natural classes of models/estimators. The lemmas below proves this result for the Gaussian linear measurement and matrix denoising settings.

Lemma 2. *Let $L^\star \in \mathbb{R}^{p_1 \times p_2}$ with rank$(L^\star) = k$. Further, let $L^\star$ have reduced-SVD $L^\star = U\Sigma V'$ for $\Sigma \in \mathbb{S}^k$ diagonal and $U \in \mathbb{R}^{p_1 \times k}$ and $V \in \mathbb{R}^{p_2 \times k}$ partial orthogonal matrices. Let $U_\perp \in \mathbb{R}^{p_1 \times p_1 - k}$ & $V_\perp \in \mathbb{R}^{p_2 \times p_2 - k}$ be partial orthogonal matrices that are orthogonal complements of U and V respectively. Consider the linear matrix regression setting $y_i = \langle \mathcal{A}_i, L^\star \rangle + \epsilon_i$. Here, $\mathcal{A}_i \in \mathbb{R}^{p_1 \times p_2}$ be iid Gaussian matrix & $\epsilon_i \in \mathbb{R}$ be chosen independently and identically distributed. Consider the following class of estimators:*

$$\hat{L} = \underset{L \in \mathbb{R}^{p_1 \times p_2}}{\text{argmin}} \quad \sum_{i \in S} (Y_i - \langle \mathcal{A}_i, L \rangle)^2 + \lambda\, \mathcal{R}(L), \tag{A.6}$$

which encompasses a convex estimator $\mathcal{R}(L) = \|L\|_\star$, as well as the alternating least squares estimator (4.2) (main paper) for $\mathcal{R}(L) = \|U\|_F^2 + \|V\|_F^2$ with $L = UV'$. In both of these settings, Assumption 2 in (3.14) *(main paper) is satisfied.*

Proof. We first examine the case when $\mathcal{R}(L) = \|L\|_\star$. Let $\hat{L}$ be the solution to (A.6). Then the objective function at $\hat{L}$, denoted by $f(\hat{L})$ takes on the value $f(\hat{L}) = \sum_{i \in S}(Y_i - \langle \mathcal{A}_i, \hat{L} \rangle)^2 + \lambda\, \|\hat{L}\|_\star$. We define the following linear operator and its adjoint for $Q_1 \in \mathbb{R}^{p_1 - k \times p_1 - k}$ and $Q_2 \in \mathbb{R}^{p_2 - k \times p_2 - k}$ orthogonal:

$$\begin{aligned}
\mathcal{L}(L; Q_1; Q_2) &= \left[\begin{pmatrix} U & U_\perp \end{pmatrix} \begin{pmatrix} \mathcal{I} & 0 \\ 0 & Q_1 \end{pmatrix}\right] \left[\begin{pmatrix} U & U_\perp \end{pmatrix}' L \begin{pmatrix} V & V_\perp \end{pmatrix}\right] \left[\begin{pmatrix} \mathcal{I} & 0 \\ 0 & Q_2 \end{pmatrix} \begin{pmatrix} V & V_\perp \end{pmatrix}\right]' \\
\mathcal{L}^\dagger(L; Q_1; Q_2) &= \left[\begin{pmatrix} U & U_\perp \end{pmatrix} \begin{pmatrix} \mathcal{I} & 0 \\ 0 & Q_1' \end{pmatrix}\right] \left[\begin{pmatrix} U & U_\perp \end{pmatrix}' L \begin{pmatrix} V & V_\perp \end{pmatrix}\right] \left[\begin{pmatrix} \mathcal{I} & 0 \\ 0 & Q_2' \end{pmatrix} \begin{pmatrix} V & V_\perp \end{pmatrix}\right]
\end{aligned} \tag{A.7}$$

We then evaluate the objective function at $\mathcal{L}(\hat{L}; Q_1; Q_2)$:

$$\begin{aligned} f(\mathcal{L}(\hat{L}; Q_1; Q_2)) &\overset{(a)}{=} \sum_{i \in S} (\langle \mathcal{A}_i, L^\star - \mathcal{L}(\hat{L}; Q_1; Q_2) \rangle + \epsilon_i)^2 + \lambda \, \|\mathcal{L}(\hat{L}; Q_1; Q_2)\|_\star \\ &\overset{(b)}{=} \sum_{i \in S} (\langle \mathcal{A}_i, \mathcal{L}(L^\star; Q_1; Q_2) \rangle - \mathcal{L}(\hat{L}; Q_1; Q_2) \rangle + \epsilon_i)^2 \\ &+ \lambda \, \|\mathcal{L}(\hat{L}; Q_1; Q_2)\|_\star \\ &\overset{(c)}{=} \sum_{i \in S} (\langle \mathcal{L}^\dagger(\mathcal{A}_i; Q_1; Q_2), L^\star - \hat{L} \rangle + \epsilon_i)^2 + \lambda \, \|\mathcal{L}(\hat{L}; Q_1; Q_2)\|_\star \\ &\overset{(d)}{=} \sum_{i \in S} (\langle \bar{\mathcal{A}}_i, L^\star - \hat{L} \rangle + \epsilon_i)^2 + \lambda \, \|\hat{L}\|_\star \end{aligned}$$

The equality $\overset{(a)}{=}$ follows from definition of the function f; $\overset{(b)}{=}$ follows from $\mathcal{L}(L^\star; Q_1; Q_2) = L^\star$; $\overset{(c)}{=}$ follows from the definition of the adjoint operator $\mathcal{L}^\dagger$; and $\overset{(d)}{=}$ follows from setting $\bar{\mathcal{A}}_i = \mathcal{L}^\dagger(\mathcal{A}_i; Q_1; Q_2)$ and that $\|\mathcal{L}(\hat{L}; Q_1; Q_2)\|_\star = \|\hat{L}\|_\star$ since the nuclear norm is invariant to multiplication by orthogonal matrices. Finally, it is straightforward to check that since i.i.d normal matrix is unitarily invariant, the distribution of $\bar{\mathcal{A}}_i$ will be the same as $\mathcal{A}_i$. Thus, $f(\mathcal{L}(\hat{L}; Q_1; Q_2))$ has the same distribution as $f(\hat{L})$, which implies that $\mathcal{L}(\hat{L}; Q_1; Q_2)$ will have the same distribution as $\hat{L}$. This invariance property has few immediate implications. Specifically, letting $Q_2 = \mathcal{I}$, it follows that $\mathcal{P}_{\hat{\mathcal{C}}(\mathcal{D}(n/2))}(u)$ has the same distribution for all $\|u\|_2 = 1$, $u \in C^{\star\perp}$. Similarly, letting $Q_1 = \mathcal{I}$, it follows that $\mathcal{P}_{\hat{\mathcal{R}}(\mathcal{D}(n/2))}(v)$ has the same distribution for all $\|v\|_2 = 1$, $v \in \mathcal{R}^{\star\perp}$. Putting these two facts together, and noting that any $M \in T^{\star\perp}$, $\text{rank}(M = 1)$, $\|M\|_F = 1$ has the form $M = uv'$, we have that $\mathcal{P}_{\hat{\mathcal{C}}(\mathcal{D}(n/2))} M \mathcal{P}_{\hat{\mathcal{R}}(\mathcal{D}(n/2))}$ is equally distributed for all $M \in T^{\star\perp}$, $\text{rank}(M) = 1$, and $\|M\|_F = 1$.

The proof of the setting $\mathcal{R}(L) = \|U\|_F^2 + \|V\|_F^2$ follows similarly and uses the property that the Frobenius norm is invariant to orthogonal multiplications. □

Lemma 3. *Let $L^\star \in \mathbb{R}^{p_1 \times p_2}$ be the underlying rank-k matrix with reduced-SVD $L^\star = U\Sigma V'$ where $U \in \mathbb{R}^{p_1 \times k}$ and $V \in \mathbb{R}^{p_2 \times k}$ are partial orthogonal and $\Sigma \in \mathbb{S}^k$ is diagonal. Furthermore, let $U_\perp \in \mathbb{R}^{p_1 \times p_1 - k}$ and $V_\perp \in \mathbb{R}^{p_2 \times p_2 - k}$ be the orthogonal complements of U and V respectively. Suppose we have n observations of $L^\star$ of the form $Y_i = L^\star + \epsilon_i$. Here $\epsilon_i \in \mathbb{R}^{p_1 \times p_2}$ is iid Gaussian matrix. Consider the hard thresholding estimator for $\hat{L}$ on the data matrix $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$. Then, Assumption 2 (main paper) is satisfied.*

Proof. The estimator can be stated as $\hat{L} = \underset{L \in \mathbb{R}^{p \times p}}{\arg\min} \sum_{i=1}^n \|Y_i - L\|_F^2 + \lambda \, \|L\|_\star$ for some choice of regularization parameter $\lambda > 0$. Let the objective function of the

above optimization be denoted by $f(\hat{L}) := \sum_{i=1}^{n} \|Y_i - L\|_F^2 + \lambda \|\hat{L}\|_\star$. Then for the linear operands (A.7), we have that the objective function $\mathcal{L}(\hat{L}; Q_1; Q_2)$ is given by:

$$\begin{aligned} f(\mathcal{L}(\hat{L}; Q_1; Q_2)) &\overset{(a)}{=} \sum_{i=1}^{n} \|\mathcal{L}(\hat{L}; Q_1; Q_2)\|_F^2 + \|L^\star\|_F^2 + \|\epsilon_i\|_F^2 - 2\langle \mathcal{L}(\hat{L}; Q_1; Q_2), L^\star \rangle \\ &- 2\langle \mathcal{L}(\hat{L}; Q_1; Q_2) - L^\star, \epsilon_i \rangle + \lambda \|\mathcal{L}(\hat{L}; Q_1; Q_2)\|_\star \\ &\overset{(b)}{=} \sum_{i=1}^{n} \|\hat{L}\|_F^2 + \|L^\star\|_F^2 + \|\epsilon_i\|_F^2 \\ &- 2\langle \hat{L}, \mathcal{L}^\dagger(L^\star; Q_1; Q_2) \rangle - 2\langle \hat{L} - L^\star, \mathcal{L}^\dagger(\epsilon_i; Q_1; Q_2) \rangle + \lambda \|\hat{L}\|_\star \\ &\overset{(c)}{=} \sum_{i=1}^{n} \|\hat{L}\|_F^2 + \|L^\star\|_F^2 + \|\epsilon_i\|_F^2 - 2\langle \hat{L}, L^\star \rangle - 2\langle \hat{L} - L^\star, \bar{\epsilon}_i] \rangle + \lambda \|\hat{L}\|_\star. \end{aligned}$$

Here $\overset{(a)}{=}$ follows from unwrapping the objective function; $\overset{(b)}{=}$ follows from the definition of an adjoint, the fact that $\mathcal{L}(L^\star; Q_1; Q_2) = L^\star$ and ; and $\overset{(c)}{=}$ follows from $\mathcal{L}^\dagger(L^\star; Q_1; Q_2) = L^\star$ and setting $\bar{\epsilon}_i = \mathcal{L}^\dagger(\epsilon_i; Q_1; Q_2)$. Since ϵ_i consists of i.i.d Gaussian entries, $\bar{\epsilon}_i$ will have the same distribution as ϵ_i. This observation implies that $f(\mathcal{L}(\hat{L}; Q_1; Q_2))$ has the same distribution as $f(\hat{L})$. Subsequently, the optimum $\hat{L}$ must have the property that $\hat{L}$ has the same distribution as $\mathcal{L}(\hat{L}; Q_1, Q_2)$. It then follows that the distributions of $\mathcal{P}_{\hat{C}}(u)$ is the same for all $u \in C^{\star\perp}$, $\|u\|_2 = 1$, and similarly the distributions of $\mathcal{P}_{\hat{\mathcal{R}}}(v)$ is the same for all $v \in \mathcal{R}^{\star\perp}$ with $\|v\|_2 = 1$. Putting these two facts together, and noting that any $M \in T^{\star\perp}$, $\text{rank}(M = 1)$, $\|M\|_F = 1$ has the form $M = uv'$, we have that $\mathcal{P}_{\hat{C}(\mathcal{D}(n/2))} M \mathcal{P}_{\hat{\mathcal{R}}(\mathcal{D}(n/2))}$ is equally distributed for all $M \in T^{\star\perp}$, $\text{rank}(M) = 1$, and $\|M\|_F = 1$. □

In Section A, we provide a PCA model and a corresponding estimator that would satisfy a version of Assumption 2 suitable for subspace estimation problems.

Proof of Proposition 1

Proof. The following lemma will be repeatedly employed.

Lemma 4. *Under assumption 2, the following hold:*

1. $\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}(M)\|_F$ *is equally distributed for all* $M \in T^{\star\perp}$, *rank*$(M) = 1$, *and* $\|M\|_F = 1$
2. $\|[\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}, \mathcal{P}_{\text{span}(M)}]\|_F$ *is equally distributed for all* $M \in T^{\star\perp}$, *rank*$(M) = 1$, *and* $\|M\|_F = 1$

Proof. Notice that for any $M \in T^{\star\perp}$ with $\text{rank}(M) = 1$ and $\|M\|_F = 1$, M can be decomposed as $M = uv'$ where $u \in C^{\star\perp}$, $v \in \mathcal{R}^{\star\perp}$, and $\|u\|_2 = \|v\|_2 = 1$. Hence the energy $\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}(M)\|_F^2$ can be reformulated as:

$$\begin{aligned}
\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}(M)\|_F^2 &= \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\mathcal{P}_{\text{span}(M)}\right) \\
&= \text{trace}\left(\mathcal{P}_{\hat{C}\mathcal{D}(n/2))}\mathcal{P}_{\text{col-space}(M)}\right) + \text{trace}\left(\mathcal{P}_{\hat{\mathcal{R}}(\mathcal{D}(n/2))}\mathcal{P}_{\text{row-space}(M)}\right) \\
&- \text{trace}\left(\mathcal{P}_{\hat{C}(\mathcal{D}(n/2))}\mathcal{P}_{\text{col-space}(M)}\right)\text{trace}\left(\mathcal{P}_{\hat{\mathcal{R}}(\mathcal{D}(n/2))}\mathcal{P}_{\text{row-space}(M)}\right) \\
&= \left\|\mathcal{P}_{\hat{C}(\mathcal{D}(n/2))}(u)\right\|_2^2 + \left\|\mathcal{P}_{\hat{\mathcal{R}}(\mathcal{D}(n/2))}(v)\right\|_2^2 \\
&- \left\|\mathcal{P}_{\hat{C}(\mathcal{D}(n/2))}(u)\right\|_2^2\left\|\mathcal{P}_{\hat{\mathcal{R}}(\mathcal{D}(n/2))}(v)\right\|_2^2
\end{aligned}$$

An immediate consequence of Assumption 2 is that $\|\mathcal{P}_{\hat{C}(\mathcal{D}(n/2))}(u)\|_2^2$ is equally distributed for all $u \in C^{\star\perp}$ with $\|u\|_2 = 1$, and $\|\mathcal{P}_{\hat{\mathcal{R}}(\mathcal{D}(n/2))}(v)\|_F^2$ is equally distributed for all $v \in \mathcal{R}^{\star\perp}$ with $\|v\|_2 = 1$. Hence, we conclude that $\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}(M)\|_F^2$ has the same distribution for all $M \in T^{\star\perp}$, $\text{rank}(M) = 1$, and $\|M\|_F = 1$. This proves item 1.

Next, recalling that $\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))} = \mathcal{P}_{\hat{C}(\mathcal{D}(n/2))} \otimes \mathcal{I} + \mathcal{I} \otimes \mathcal{P}_{\hat{\mathcal{R}}(\mathcal{D}(n/2))} - \mathcal{P}_{\hat{C}(\mathcal{D}(n/2))} \otimes \mathcal{P}_{\hat{\mathcal{R}}(\mathcal{D}(n/2))}$, and $\mathcal{P}_{\text{span}(M)} = \mathcal{P}_{\text{span}(u)} \otimes \mathcal{P}_{\text{span}(v)}$, we have:

$$\begin{aligned}
\|[\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}, \mathcal{P}_{\text{span}(M)}]\|_F^2 &= 2\,\text{trace}\left(\mathcal{P}_{\text{span}(M)}\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\right) \\
&- 2\,\text{trace}\left(\mathcal{P}_{\text{span}(M)}\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\mathcal{P}_{\text{span}(M)}\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\right) \\
&= 2\,\|\mathcal{P}_{\hat{C}(\mathcal{D}(n/2))}(u)\|_2^2 + 2\|\mathcal{P}_{\hat{\mathcal{R}}(\mathcal{D}(n/2))}(v)\|_2^2 \\
&- 2\,\|\mathcal{P}_{\hat{C}(\mathcal{D}(n/2))}(u)\|_2^2\|\mathcal{P}_{\hat{\mathcal{R}}(\mathcal{D}(n/2))}(v)\|_2^2 \\
&+ \|\mathcal{P}_{\hat{C}(\mathcal{D}(n/2))}(u)\|_2^4\left[1 - \|\mathcal{P}_{\hat{\mathcal{R}}(\mathcal{D}(n/2))}(v)\|_2^2\right]^2 \\
&+ \|\mathcal{P}_{\hat{\mathcal{R}}(\mathcal{D}(n/2))}(v)\|_2^4\left[1 - \|\mathcal{P}_{\hat{C}(\mathcal{D}(n/2))}(u)\|_2^2\right]^2
\end{aligned}$$

Since $\|\mathcal{P}_{\hat{C}(\mathcal{D}(n/2))}(u)\|_F^2$ is equally distributed for all $u \in C^{\star\perp}$ with $\|u\|_2 = 1$, and $\|\mathcal{P}_{\hat{\mathcal{R}}(\mathcal{D}(n/2))}(v)\|_F^2$ is equally distributed for all $v \in \mathcal{R}^{\star\perp}$ with $\|v\|_2 = 1$, we conclude that $\|[\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}, \mathcal{P}_{\text{span}(M)}]\|_F^2$ is equally distributed for all $M \in T^{\star\perp}$ with $\text{rank}(M) = 1$ and $\|M\|_F = 1$. □

We proceed with the proof of Proposition 1. Notice that

$$\mathbb{E}\left[\text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\mathcal{P}_{T^{\star\perp}}\right)\right] + \mathbb{E}\left[\text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\mathcal{P}_{T^{\star}}\right)\right] = \mathbb{E}[\dim(\hat{T}(\mathcal{D}(n/2)))]$$

Employing “better than random guessing" Assumption 1, we then find that:

$$\mathbb{E}\left[\text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\mathcal{P}_{T^{\star\perp}}\right)\right] \leq \frac{\mathbb{E}[\dim(\hat{T}(\mathcal{D}(n/2)))]}{p_1 p_2}\dim(T^{\star\perp}) \tag{A.8}$$

Since $\{M_i\}_{i=1}^{\dim(T^{\star\perp})}$ are orthonormal basis elements for $T^{\star\perp}$, we have that $\mathbb{E}\left[\text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\mathcal{P}_{T^{\star\perp}}\right)\right] = \sum_i \mathbb{E}\left[\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}(M_i)\|_F^2\right]$. Combining (A.8) with the first item of Lemma 4 yields that for any $M \in T^{\star\perp}$, $\text{rank}(M) = 1$, $\|M\|_F = 1$:

$$\mathbb{E}\left[\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}(M)\|_F^2\right] \leq \frac{\mathbb{E}[\dim(\hat{T}(\mathcal{D}(n/2)))]}{p_1 p_2} \tag{A.9}$$

Notice that:

$$\begin{aligned}
&\mathbb{E}\left[\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}(M)\|_F\right] - \mathbb{E}\left[\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}(M)\|_F^2\right] \\
\overset{(a)}{=}\ &\mathbb{E}\left[\left\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\mathcal{P}_{\text{span}(M)}\right\|_F\right] - \mathbb{E}\left[\left\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\mathcal{P}_{\text{span}(M)}\right\|_F^2\right] \\
\overset{(b)}{=}\ &\mathbb{E}\left[\left\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\mathcal{P}_{\text{span}(M)}\right\|_F\right] - \mathbb{E}\left[\left\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\mathcal{P}_{\text{span}(M)}\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\right\|_F\right] \\
\overset{(c)}{=}\ &\mathbb{E}\left[\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\mathcal{P}_{\text{span}(M)}]\|_F\right] \\
-\ &\mathbb{E}\left[\left\|\mathcal{P}_{\text{span}(M)}\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))} + \left[\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}, \mathcal{P}_{\text{span}(M)}\right]\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\right\|_F\right] \\
\overset{(d)}{\leq}\ &\mathbb{E}\left\|\left[\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}, \mathcal{P}_{\text{span}(M)}\right]\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\right\|_F \overset{(e)}{\leq} \mathbb{E}\left\|\left[\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}, \mathcal{P}_{\text{span}(M)}\right]\right\|_F
\end{aligned} \tag{A.10}$$

Here $\overset{(a)}{=}$ follows from the property that $\mathbb{E}\left[\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}(M)\|_F\right] = \mathbb{E}\left[\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\mathcal{P}_{\text{span}(M)}\|_F\right]$; $\overset{(b)}{=}$ follows from noting that $\mathcal{P}_{\text{span}(M)}$ has rank-1 by construction so that $\left\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\mathcal{P}_{\text{span}(M)}\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\right\|_F = \left\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}\mathcal{P}_{\text{span}(M)}\right\|_F^2$; $\overset{(c)}{=}$ follows from the definition of a commutator; $\overset{(d)}{\leq}$ follows from reverse triangle inequality; and $\overset{(e)}{\leq}$ follows idempotence of projection operators and $\text{trace}(AB) \leq \text{trace}(A)\|B\|_2$ for $A \geq 0$. One again applying item 1 of Lemma 4, we find that $\mathbb{E}\left[\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}(M)\|_F\right] - \mathbb{E}\left[\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}(M))\|_F^2\right]$ is the same for all $M \in T^{\star\perp}$, $\|M\|_F = 1$, $\text{rank}(M) = 1$. Combining this with the bound (A.10), we have that for all $M \in T^{\star\perp}$, $\|M\|_F = 1$, $\text{rank}(M) = 1$:

$$\begin{aligned}
&\mathbb{E}\left[\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}(M)\|_F - \|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}(M)\|_F^2\right] \\
\leq\ &\min_{\substack{M\in T^{\star\perp}, \|M\|_F=1 \\ \text{rank}(M)=1}} \mathbb{E}\left\|\left[\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}, \mathcal{P}_{\text{span}(M)}\right]\right\|_F.
\end{aligned} \tag{A.11}$$

Applying item 2 of Lemma 4, we know that the objective on the right-hand side of (A.11) is the same for all M, and thus the minimizer can be removed. Putting everything together, we have that for any $M \in T^{\star\perp}$, $\text{rank}(M) = 1$, and $\|M\|_F = 1$:

$$\begin{aligned}
&\sum_{i=1}^{\dim(T^{\star\perp})} \mathbb{E}\left[\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}(M_i)\|_F\right]^2 \\
= \;& \dim(T^{\star\perp})\,\mathbb{E}\left[\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}(M)\|_F\right]^2 \\
= \;& \dim(T^{\star\perp})\left\{\mathbb{E}\left[\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}(M)\|_F\right] - \mathbb{E}\left[\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}(M)\|_F\right]^2\right. \\
+ \;& \left.\mathbb{E}\left[\|\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}(M)\|_F\right]^2\right\}^2 \\
\leq \;& \left(\frac{\mathbb{E}[\dim(\hat{T}(\mathcal{D}(n/2)))]}{p_1 p_2} + \mathbb{E}\left\|\left[\mathcal{P}_{\hat{T}(\mathcal{D}(n/2))}, \mathcal{P}_{\text{span}(M)}\right]\right\|_F\right)^2 p_1 p_2 \qquad \text{(A.12)}
\end{aligned}$$

Hence, the term F in Theorem 4 can be bounded by the quantity in (A.12), giving the desired result. □

Proof of Proposition 2

There are two terms inside the Theorem 4 bound that are dependent on the number of bags: $\mathbb{E}[\dim(T)]$ and κ_{bag}. Recall that $\mathbb{E}[\dim(T)] \leq \frac{q}{\alpha}$. We next get a handle on the quantity

$\kappa_{\text{bag}} = \mathbb{E}[\min\{\sqrt{\dim(T)}\sqrt{\frac{1}{B}\sum_{j=1}^{B}\|[\mathcal{P}_{T^{\star\perp}}, \mathcal{P}_{\hat{T}(\mathcal{D}_j)}]\|_F^2}, \dim(T)\sqrt{\frac{1}{B}\sum_{j=1}^{B}\|[\mathcal{P}_{T^{\star\perp}}, \mathcal{P}_{\hat{T}(\mathcal{D}_j)}]\|_2^2}\}]$.
Since for any two subspaces T_1, T_2, $\|[\mathcal{P}_{T_1}, \mathcal{P}_{T_2}]\|_2 \leq \frac{1}{2}$, it immediately follows that $\kappa_{\text{bag}} \leq \frac{1}{2}\mathbb{E}[\dim(T)] \leq \frac{q}{2\alpha}$. This allows us to conclude the first results of Proposition 2.

Next, we prove the modified bound provided in the text following Proposition 2. This proof relies on the following lemma:

Lemma 5. *For any subset of indices $S \subset [1, B]$, with $|S| = B/2$: $\sigma_{min}\left(\mathcal{P}_T \mathcal{P}^S_{\mathsf{avg}} \mathcal{P}_T\right) \geq 2\alpha - 1$, where $\mathcal{P}^S_{\mathsf{avg}} = \frac{2}{B}\sum_{k\in S}\mathcal{P}_{\hat{T}(\mathcal{D}_k)}$.*

Proof. Notice that $\sigma_{\min}\left(\mathcal{P}_T\left[\frac{2}{B}\sum_{k\in S}\mathcal{P}_{\hat{T}(\mathcal{D}_k)} + \frac{2}{B}\sum_{k\in S^c}\mathcal{P}_{\hat{T}(\mathcal{D}_k)}\right]\mathcal{P}_T\right) \geq 2\alpha$. Since $\sigma_{\min}(A+B) \leq \sigma_{\min}(A) + \sigma_{\max}(B)$ and that $\sigma_{\max}\left(\mathcal{P}_T\left[\frac{2}{B}\sum_{k\in S^c}\mathcal{P}_{\hat{T}(\mathcal{D}_k)}\right]\mathcal{P}_T\right) \leq 1$, we conclude the desired result. □ □

Next, we prove a more refined bound. Consider the decomposition (A.1):

$$\begin{aligned}\text{trace}\left(\mathcal{P}_T\mathcal{P}_{\text{span}(M_i)}\right) &= \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)}\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_j)}\mathcal{P}_{\text{span}(M_i)}\right)\\ &+ \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\mathcal{P}_{\text{span}(M_i)}\right)\\ &+ \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)}\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\mathcal{P}_{\text{span}(M_i)}\right)\\ &+ \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_j)}\mathcal{P}_{\text{span}(M_i)}\right)\end{aligned}$$

Summing over all i and the idempotence of projection operators yields:

$$\begin{aligned}\text{trace}\left(\mathcal{P}_T\mathcal{P}_{T^{\star\perp}}\right) &= \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)}\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_j)}\mathcal{P}_{T^{\star\perp}}\right) + \text{trace}\left(\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\mathcal{P}_{T^{\star\perp}}\right)\\ &+ \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_j)}\mathcal{P}_{T^{\star\perp}}\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\right)\end{aligned}$$

Appealing to the inequalities $\text{trace}(AB) \le \|A\|_\star\|B\|_2$, $\text{trace}(AB) \le \text{trace}(A)\|B\|_2$ for $A \succeq 0$, projection operators have spectral norm equal to one yields the bound, and that for subspaces T_1 & T_2, $\|[\mathcal{P}_{T_1}, \mathcal{P}_{T_2}]\|_2 \le \frac{1}{2}$, we have:

$$\text{trace}\left(\mathcal{P}_T\mathcal{P}_{T^{\star\perp}}\right) \le \text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_j)}\mathcal{P}_{T^{\star\perp}}\right) + \frac{3}{2}\left\|\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_j)^\perp}\right\|_\star.$$

Since the choice of $\mathcal{D}_j$ was arbitrary, we minimizing over the entire collection $\mathcal{D}_j$ to find

$$\begin{aligned}\text{trace}\left(\mathcal{P}_T\mathcal{P}_{T^{\star\perp}}\right) &= \min_{j=1,2,\ldots,B/2}\min_{k=\{0,1\}}\left\{\text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})}\mathcal{P}_{T^{\star\perp}}\right) + \frac{3}{2}\left\|\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})^\perp}\right\|_\star\right\}\\ &\le \text{Term 1} + \text{Term 2}\end{aligned}$$

where

$$\begin{aligned}\text{Term 1} &= \left[\frac{2}{B}\sum_{j=1}^{B/2}\min_{k=\{0,1\}}\text{trace}\left(\mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})}\mathcal{P}_{T^{\star\perp}}\right)\right]\\ \text{Term 2} &= \frac{4}{B}\sum_{j=1}^{B/2}\max_{k=\{0,1\}}\left\|\mathcal{P}_T\mathcal{P}_{\hat{T}(\mathcal{D}_{2j-k})^\perp}\right\|_\star\end{aligned}$$

Here, we used the inequality $\min\{a+b, c+d\} \le \min\{a,c\} + \max\{b,d\}$. Term 1 is bounded in the proof of Theorem 4 in Section A. Examining Term 2, we define

$\hat{k}(j) = \arg\max_k \left\| \mathcal{P}_T \mathcal{P}_{\hat{T}(D_{2j-k})^\perp} \right\|_\star$. Then,

$$\begin{aligned}
\text{Term 2} &\overset{(a)}{\leq} \frac{3}{B}\sum_{j=1}^{B/2} \left\| \mathcal{P}_T \mathcal{P}_{\hat{T}\left(\mathcal{D}_{2j-\hat{k}(j)}\right)^\perp} \right\|_\star \overset{(b)}{\leq} \frac{3}{B}\sum_{j=1}^{B/2} \sqrt{\left\| \mathcal{P}_T \mathcal{P}_{\hat{T}\left(\mathcal{D}_{2j-\hat{k}(j)}\right)^\perp} \right\|_F^2}\sqrt{\dim(T)} \\
&\overset{(c)}{\leq} 3\sqrt{\operatorname{trace}\left(\mathcal{P}_T\left[\frac{2}{B}\sum_{j=1}^{B/2} \mathcal{P}_{\hat{T}\left(\mathcal{D}_{2j-\hat{k}(j)}\right)^\perp}\right]\mathcal{P}_T\right)}\sqrt{\dim(T)} \\
&\overset{(d)}{\leq} \frac{3}{2}\sqrt{\left\| \mathcal{P}_T\left[\frac{2}{B}\sum_{j=1}^{B/2} \mathcal{P}_{\hat{T}\left(\mathcal{D}_{2j-\hat{k}(j)}\right)^\perp}\right]\mathcal{P}_T \right\|_2 \dim(T)} \overset{(e)}{\leq} \frac{3}{2}\sqrt{2(1-\alpha)}\dim(T).
\end{aligned}$$

Here, $\overset{(a)}{\leq}$ follows from the definition of $\hat{k}(j)$; $\overset{(b)}{\leq}$ follows from $\|A\|_\star \leq \|A\|_F \operatorname{rank}(A)$; $\overset{(c)}{\leq}$ follows from concavity of the square root function; $\overset{(d)}{\leq}$ follows from idempotence of projection operators and that $\operatorname{trace}(AB) \leq \operatorname{trace}(A)\|B\|_2$ for $A \succeq 0$; and finally $\overset{(e)}{\leq}$ follows from Lemma 5. Taking expectations and employing the inequality $\mathbb{E}[\dim(T)] \leq \frac{q}{\alpha}$ gives the desired result.

Proof of Proposition 3

Let T be a tangent space produced by the modified algorithm with associated column and row spaces $(C, \mathcal{R})$. We proceed by obtaining an upper bound on $\|\mathcal{P}_T(\mathcal{I} - \mathcal{P}_{\mathsf{avg}})\mathcal{P}_T\|_2$, which gives a lower bound on $\sigma_{\min}(\mathcal{P}_T\mathcal{P}_{\mathsf{avg}}\mathcal{P}_T)$:

$$\begin{aligned}
\|\mathcal{P}_T(\mathcal{I} - \mathcal{P}_{\mathsf{avg}})\mathcal{P}_T\|_2 &= \max_{M\in T, \|M\|_F=1} \frac{1}{B}\operatorname{trace}\Big(\sum_{i=1}^{B} M'\mathcal{P}_{\hat{T}(\mathcal{D}_i)^\perp}(M)\Big) \\
&\overset{(a)}{=} \max_{M\in T, \|M\|_F=1} \frac{1}{B}\sum_{i=1}^{B} \|\mathcal{P}_{\hat{C}(\mathcal{D}_i)^\perp} M \mathcal{P}_{\hat{\mathcal{R}}(\mathcal{D}_i)^\perp}\|_F^2 \\
&\overset{(b)}{\leq} \max_{M\in T, \|M\|_F=1} \frac{2}{B}\sum_{i=1}^{B} \|\mathcal{P}_{\hat{C}(\mathcal{D}_i)^\perp}\mathcal{P}_C M \mathcal{P}_{\hat{\mathcal{R}}(\mathcal{D}_i)^\perp}\|_F^2 \\
&\quad + \frac{2}{B}\sum_{i=1}^{B} \|\mathcal{P}_{\hat{C}(\mathcal{D}_i)^\perp}\mathcal{P}_{C^\perp} M \mathcal{P}_{\mathcal{R}}\mathcal{P}_{\hat{\mathcal{R}}(\mathcal{D}_i)^\perp}\|_F^2 \\
&\overset{(c)}{\leq} \max_{M\in T, \|M\|_F=1} \frac{2}{B}\sum_{i=1}^{B} \|\mathcal{P}_{\hat{C}(\mathcal{D}_i)^\perp}\mathcal{P}_C M\|_F^2 + \frac{2}{B}\sum_{i=1}^{B} \|\mathcal{P}_{\mathcal{R}}\mathcal{P}_{\hat{\mathcal{R}}(\mathcal{D}_i)^\perp} M'\|_F^2 \\
&= \max_{M\in T, \|M\|_F=1} 2\operatorname{trace}(\mathcal{P}_C(\mathcal{I} - \mathcal{P}_{\mathsf{avg}})\mathcal{P}_C M M') \\
&\quad + 2\operatorname{trace}(\mathcal{P}_{\mathcal{R}}(\mathcal{I} - \mathcal{P}_{\mathsf{avg}})\mathcal{P}_{\mathcal{R}} M' M) \\
&\leq 2\|\mathcal{P}_C(\mathcal{I} - \mathcal{P}_{\mathsf{avg}})\mathcal{P}_C\|_2 + 2\|\mathcal{P}_{\mathcal{R}}(\mathcal{I} - \mathcal{P}_{\mathsf{avg}})\mathcal{P}_{\mathcal{R}}\|_2 \leq 4(1-\alpha).
\end{aligned}$$

Here (a) follows from the cyclicity of the trace functional and the idempotence of projection maps; (b) from the fact that $M \in T$ implies that $M = \mathcal{P}_C M + \mathcal{P}_{C^\perp} M \mathcal{P}_R$ and the elementary inequality $(a+b)^2 \leq 2a^2 + 2b^2$; and (c) from the property $\|A\mathcal{P}\|_F \leq \|A\|_F$ for any projection matrix $\mathcal{P}$.

Sensitivity of Subspace Stability Selection to α

In this section, we explore the sensitivity of the subspace stability selection algorithm to the choice of α. Specifically, we consider the matrix completion setting in synthetic simulations of Section 4. We chose the rank of $L^\star \in \mathbb{R}^{100\times 100}$ in the set $\{1, 3, 5\}$ and the variance so that SNR is in the set $\{0.5, 0.8, 2\}$, for a total number of 9 problem instances. We fix $B = 100$ and vary $\alpha \in \alpha_{set}$ where $\alpha_{\text{set}} = \{0.6, 0.625, 0.65, 0.675, 0.7, 0.725, 0.75, 0.775, 0.8\}$. Figure A.1 demonstrates the variation in the normalized false discovery $\mathbb{E}\left[\text{trace}\left(\mathcal{P}_{T_{\text{S3}}(\alpha)}\mathcal{P}_{T^{\star\perp}}\right)\right]/\dim(T^{\star\perp})$ and normalized power $\mathbb{E}\left[\text{trace}\left(\mathcal{P}_{T_{\text{S3}}(\alpha)}\mathcal{P}_{T^\star}\right)\right]/\dim(T^\star)$ as a function of α. We notice that for moderate SNR regimes (e.g. SNR = 2), both the false discovery and power are very stable with respect to α for all ranks. Furthermore, for a fixed SNR (say SNR = 0.8), the stability to α is reduced for larger ranks. In summary, this experiment indicates that subspace stability selection algorithm tends to be robust to perturbations of α for moderate SNR regimes and small ranks.

Tangent Spaces for Subspace Estimation Problem

Consider a collection of n data points $\mathcal{D} = \{y_i\}_{i=1}^n \subset \mathbb{R}^p$ that lie on a low-dimensional subspace $C^\star$. Subspace estimation algorithms use these points as training data to obtain an estimated subspace $\hat{C}$. Let $\hat{L}$ be an $p \times n$ matrix formed by concatenating the vectors y_i and let r be the dimension of $\hat{C}$. By construction, $\hat{L}$ will have column space equal to $\hat{C}$ and lies inside the determinantal variety $\mathcal{V}(r)$. A first attempt at a tangent space formulation for $\hat{L}$ is simply the tangent space at L with respect to the determinantal variety. Specifically, parameterizing $\hat{L} = UV'$ with $U \in \mathbb{R}^{p\times r}$ and $V \in \mathbb{R}^{n\times r}$, the tangent space linearization is computed via perturbations $(U + \Delta_1)(V + \Delta_2)'$ around $\hat{L}$ that still lie in $\mathcal{V}(r)$ (since by construction $(U + \Delta_1)(V + \Delta_2)'$ has rank less than or equal to r). With this viewpoint of computing tangent spaces, the perturbations $(U + \Delta_1)(V + \Delta_2)'$ include matrices that have a component of their column-space inside $\hat{C}^\perp$, which is undesirable since our notion of discovery in the subspace estimation problem should not involve components orthogonal to the column space $\hat{C}$.

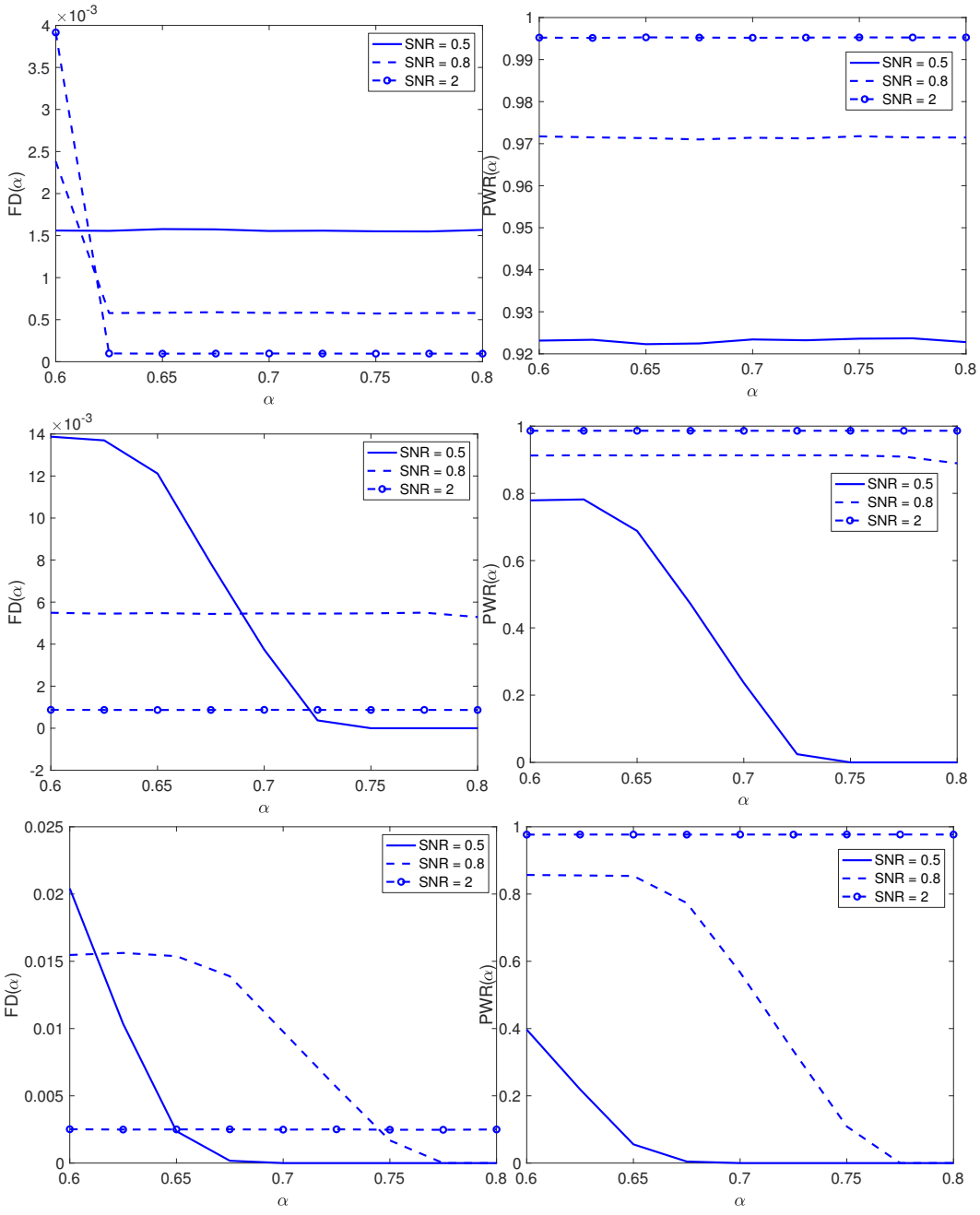


Figure A.1: Variation in false discovery $\mathbb{E}\left[\text{trace}\left(\mathcal{P}_{T_{\text{S3}}(\alpha)}\mathcal{P}_{T^{\star\perp}}\right)\right]$ and power $\mathbb{E}\left[\text{trace}\left(\mathcal{P}_{T_{\text{S3}}(\alpha)}\mathcal{P}_{T^{\star}}\right)\right]$ as a function of α for different SNR and rank regimes.

Based on this intuition, we quotient away unwanted components of the determinantal variety, and compute the tangent space with respect to this quotient manifold. In particular, we want to remove perturbations $\hat{L} + \mathcal{P}_{\hat{C}^\perp}\Delta$ that lie in the determinantal variety $\mathcal{V}(r)$. These perturbations are precisely of the form $\hat{L} + \mathcal{P}_{\hat{C}^\perp}\Delta\mathcal{P}_{\hat{\mathcal{R}}}$, where $\hat{\mathcal{R}}$ is the row-space of L. As such, we consider the following equivalence class for $\hat{L}$:

$$[\hat{L}] = \{\hat{L} + \mathcal{P}_{\hat{C}^\perp}\Delta\mathcal{P}_{\hat{\mathcal{R}}} \mid \Delta \in \mathbb{R}^{p\times n}\}, \tag{A.13}$$

which consist of all perturbation of $\hat{L}$ that contribute in the orthogonal subspace $\hat{C}^\perp$ and lie within the determinantal variety $\mathcal{V}(r)$. The quotient manifold, $\mathcal{V}(r)/[\hat{L}]$ collapses all the points within $[L]$, and our tangent space discovery is the tangent space with respect to this manifold.

The tangent spaces of $[\hat{L}]$ and $\mathcal{V}(r)/[\hat{L}]$ at $\hat{L}$ form complementary subspaces of the tangent space $T(\hat{C}, \hat{\mathcal{R}})$. Specifically, the tangent space with respect to an equivalence class is known as the *vertical space* and the tangent space with respect to the quotient manifold is known as the *horizontal space*. The vertical space (tangent space of (A.13)) equates to $T_{\text{vertical}}(\hat{C}, \hat{\mathcal{R}}) = \{\mathcal{P}_{\hat{C}^\perp}\Delta\mathcal{P}_{\hat{\mathcal{R}}} \mid \Delta \in \mathbb{R}^{p\times n}\}$ and the horizontal space (tangent space to quotient manifold $\mathcal{V}(r)/[\hat{L}]$) equates to $T_{\text{horizontal}}(\hat{C}) = \{\mathcal{P}_{\hat{C}}\Delta \mid \Delta \in \mathbb{R}^{p\times n}\}$ so that $T(\hat{C}, \hat{\mathcal{R}}) = T_{\text{vertical}}(\hat{C}, \hat{\mathcal{R}}) \oplus T_{\text{horizontal}}(\hat{C})$. Our tangent space discovery is thus the subspace $T_{\text{horizontal}}(\hat{C})$.

We are ready to define FD, Power and FDR. Noting that $\mathcal{P}_{T_{\text{horizontal}}(\hat{C})} = \mathcal{P}_{\hat{C}} \otimes \mathcal{I}$ and $\mathcal{P}_{T_{\text{horizontal}}(\hat{C})^\perp} = \mathcal{P}_{\hat{C}^\perp} \otimes \mathcal{I}$, the false discovery and true discovery metrics evaluate to:

$$\begin{aligned}
\text{FD} &= \mathbb{E}\left[\text{trace}\left(\mathcal{P}_{T_{\text{horizontal}}(\hat{C})}\mathcal{P}_{T_{\text{horizontal}}(C^\star)^\perp}\right)\right] = n\,\mathbb{E}\left[\text{trace}\left(\mathcal{P}_{\hat{C}}\mathcal{P}_{C^{\star\perp}}\right)\right] \\
\text{PWR} &= \mathbb{E}\left[\text{trace}\left(\mathcal{P}_{T_{\text{horizontal}}(\hat{C})}\mathcal{P}_{T_{\text{horizontal}}(C^\star)}\right)\right] = n\,\mathbb{E}\left[\text{trace}\left(\mathcal{P}_{\hat{C}}\mathcal{P}_{C^\star}\right)\right] \\
\text{FDR} &= \mathbb{E}\left[\frac{\text{trace}\left(\mathcal{P}_{T_{\text{horizontal}}(\hat{C})}\mathcal{P}_{T_{\text{horizontal}}(C^\star)^\perp}\right)}{\dim(T_{\text{horizontal}}(\hat{C}))}\right] = \mathbb{E}\left[\frac{\text{trace}\left(\mathcal{P}_{\hat{C}}\mathcal{P}_{C^{\star\perp}}\right)}{\dim(\hat{C})}\right].
\end{aligned} \tag{A.14}$$

The factor n in (A.14) is due to the fact that the row-space structure of L is a *free parameter* with respect to the tangent space discovery. Since this scaling is constant with respect to $\hat{C}$ & $C^\star$, we remove this factor in our characterization of false discovery and power.

Theoretical False Discovery Guarantees for Subspace Estimation

In this section, we provide false discovery control guarantees of subspace stability selection algorithm for subspace estimation problems. We suppose there exists a population subspace $C^\star \in \mathbb{R}^{p_1}$, and we are given i.i.d observations from a model parameterized by $C^\star$. Let $\hat{C}$ be a subspace estimator that operates on samples dram from the model parameterized by $C^\star$. Let $\mathcal{D}(n)$ denote a dataset consisting of n i.i.d observations from these models; we assume n is even and that we are given B subsamples $\{\mathcal{D}_i\}_{i=1}^B$ via complementary partitions of $\mathcal{D}(n)$.

We omit the proof of each of these statements as their proof is similar in spirit to those from the main paper.

Theorem 6 (False Discovery Control of Subspace Stability Selection)**.** *Consider the setup described above. Let $\hat{C}(\mathcal{D}_j)$ denote the subspace estimates obtained from each of the subsamples, and let $\mathcal{P}^C_{\mathsf{avg}}$ denote the associated average projection operator computed via (3.2) (main paper). Fix any $\alpha \in (0,1)$ and let C denote any selection of an element of the associated set $\mathcal{T}_\alpha$ of stable tangent spaces. Then for any collection of orthonormal basis elements $\{M_i\}_{i=1}^{\dim(C^{\star\perp})}$ of $C^{\star\perp}$*

$$\mathbb{E}\left[\text{trace}\left(\mathcal{P}_C \mathcal{P}_{C^{\star\perp}}\right)\right] \leq F + 4\sqrt{1-\alpha}\kappa_{avg} + 2(1-\alpha)\mathbb{E}[\dim(C)], \tag{A.15}$$

where,

$$\begin{aligned} F &= \min\left\{ \sum_{i=1}^{\dim(C^{\star\perp})} \mathbb{E}\left[\left\|\mathcal{P}_{\hat{C}(\mathcal{D}(n/2))}(M_i)\right\|_F\right]^2, \mathbb{E}\left[\text{trace}\left(\mathcal{P}_{\hat{C}(\mathcal{D}(n/2))}\mathcal{P}_{C^{\star\perp}}\right)^{1/2}\right]^2 \right\} \\ \kappa_{bag} &= \mathbb{E}\Bigg[\min\Bigg\{ \sqrt{\dim(C)}\sqrt{\frac{1}{B}\sum_{j=1}^B \|[\mathcal{P}_{C^{\star\perp}}, \mathcal{P}_{\hat{C}(\mathcal{D}_j)}]\|_F^2} \\ &\quad , \ \dim(C)\sqrt{\frac{1}{B}\sum_{j=1}^B \|[\mathcal{P}_{C^{\star\perp}}, \mathcal{P}_{\hat{C}(\mathcal{D}_j)}]\|_2^2} \Bigg\}\Bigg] \end{aligned}$$

Here the expectation is with respect to randomness in the observations. The set $\mathcal{D}(n/2)$ denotes a collection of $n/2$ i.i.d. observations drawn from the model parametrized by $C^\star$.

The next proposition provides a refined bound under "better than random guess-

ing" and exchangeability assumptions:

$$\text{Assumption 3: } \frac{\mathbb{E}\left[\text{trace}\left(\mathcal{P}_{C^{\star\perp}}\mathcal{P}_{\hat{C}(\mathcal{D}(n/2))}\right)\right]}{\dim(C^{\star\perp})} \leq \frac{\mathbb{E}\left[\text{trace}\left(\mathcal{P}_{C^{\star}}\mathcal{P}_{\hat{C}(\mathcal{D}(n/2))}\right)\right]}{\dim(C^{\star})}$$
$$\text{Assumption 4: The distribution of } \mathcal{P}_{\hat{C}(\mathcal{D}(n/2))}(M) \text{ is the same for all} \quad (A.16)$$
$$M \in C^{\star\perp}, \|M\|_F = 1.$$

The idea behind these two assumptions are similar to (3.14) in Section A. In particular, a similar argument as the one in Lemma 1 demonstrate that Assumption 3 is very benign. The following Lemma examines a PCA model that satisfies Assumption 4.

Lemma 6. *Consider the PCA model $y = \mathcal{B}^\star z + \epsilon$ for $\mathcal{B}^\star \in \mathbb{R}^{p_1 \times k}$ and $\epsilon \in \mathbb{R}^{p_1}$ having i.i.d Gaussian entries. Consider the PCA-estimator that finds top components of the empirical covariance of y from observations. Then the estimator satisfies Assumption 4 in* (A.16).

Proof. Letting $Y \in \mathbb{R}^{p \times n}$ be the concatenation of the data points, a way to interpret the PCA estimator of the column space is via the optimization problem: $\hat{L} = \underset{L \in \mathbb{R}^{p_1 \times n}}{\text{argmin}} \ \|Y - L\|_F^2 + \lambda \|L\|_\star$. Here the PCA components are captured by the column-space of $\hat{L}$ and the number of components is tuned via the parameter λ. Then the objective function at $\hat{L}$, denoted by $f(\hat{L})$ takes on the value $f(\hat{L}) = \|Y - L\|_F^2 + \lambda \|\hat{L}\|_\star$. Let $U \in \mathbb{R}^{p_1 \times k}$ be an orthonormal basis set for $C^\star$ and $U_\perp \in \mathbb{R}^{p_1 \times p_1 - k}$ be an orthonormal basis set for $C^{\star\perp}$. We define the following linear operator and its adjoint for $Q_1 \in \mathbb{R}^{p_1 - k \times p_1 - k}$ orthogonal:

$$\begin{aligned}
\mathcal{L}(L; Q_1) &= \left[\begin{pmatrix} U & U_\perp \end{pmatrix} \begin{pmatrix} \mathcal{I} & 0 \\ 0 & Q_1 \end{pmatrix}\right] \begin{pmatrix} U & U_\perp \end{pmatrix}' L \\
\mathcal{L}^\dagger(L; Q_1) &= \left[\begin{pmatrix} U & U_\perp \end{pmatrix} \begin{pmatrix} \mathcal{I} & 0 \\ 0 & Q_1' \end{pmatrix}\right] \begin{pmatrix} U & U_\perp \end{pmatrix}' L
\end{aligned}$$

Let $Z \in \mathbb{R}^{k \times n}$ be the collection of latent observations and $E \in \mathbb{R}^{p \times n}$ be the concatenation of the noise variables ϵ across the samples. We then evaluate the objective function at $\mathcal{L}(\hat{L}; Q_1)$:

$$\begin{aligned}
f(\mathcal{L}(\hat{L}; Q_1)) &\overset{(a)}{=} \|\mathcal{L}(\hat{L}; Q_1)\|_F^2 + \|\mathcal{B}^\star Z\|_F^2 + \|E\|_F^2 - 2\langle \mathcal{L}(\hat{L}; Q_1), \mathcal{B}^\star Z \rangle \\
&- 2\langle \mathcal{L}(\hat{L}; Q_1) - \mathcal{B}^\star Z, E \rangle + \lambda \|\mathcal{L}(\hat{L}; Q_1)\|_\star \\
&\overset{(b)}{=} \|\hat{L}\|_F^2 + \|\mathcal{B}^\star Z\|_F^2 + \|E\|_F^2 - 2\langle \hat{L}, \mathcal{L}^\dagger(\mathcal{B}^\star Z; Q_1) \rangle \\
&- 2\langle \hat{L} - \mathcal{B}^\star Z, \mathcal{L}^\dagger(E; Q_1) \rangle + \lambda \|\hat{L}\|_\star \\
&\overset{(c)}{=} \|\hat{L}\|_F^2 + \|\mathcal{B}^\star Z\|_F^2 + \|E\|_F^2 - 2\langle \hat{L}, \mathcal{B}^\star Z \rangle - 2\langle \hat{L} - \mathcal{B}^\star Z, \tilde{E} \rangle + \lambda \|\hat{L}\|_\star.
\end{aligned}$$

Here $\overset{(a)}{=}$ follows from unwrapping the objective function; $\overset{(b)}{=}$ follows from the definition of an adjoint, the fact that $\mathcal{L}(\mathcal{B}^\star Z; Q_1) = \mathcal{B}^\star Z$ and ; and $\overset{(c)}{=}$ follows from $\mathcal{L}^\dagger(\mathcal{B}^\star Z; Q_1) = \mathcal{B}^\star Z$ and setting $\tilde{E} = \mathcal{L}^\dagger(E; Q_1)$. Since E consists of i.i.d Gaussian entries, $\tilde{E}$ will have the same distribution as E. This observation implies that $f(\mathcal{L}(\hat{L}; Q_1))$ has the same distribution as $f(\hat{L})$. Subsequently, the optimum $\hat{L}$ must have the property that $\hat{L}$ has the same distribution as $\mathcal{L}(\hat{L}; Q_1)$. It then follows that the distribution of $\mathcal{P}_{\hat{C}}(u)$ is the same for any $u \in C^{\star\perp}$ with $\|u\|_2 = 1$. □

Proposition 9 (Refind False Discovery Bound)**.** *Suppose that Assumptions 3 and 4 in* (A.16) *are satisfied. Let the average number of discoveries from $n/2$ observations be denoted by $q := \mathbb{E}[\dim(\hat{C}(\mathcal{D}(n/2)))]$. Then, for any $M \in C^{\star\perp}$ with $\|M\|_2 = 1$, the false discovery of a stable tangent space is bounded by:*

$$\mathbb{E}\left[\text{trace}\left(\mathcal{P}_C \mathcal{P}_{C^{\star\perp}}\right)\right] \leq \frac{q^2}{p_1} + f\left(\kappa_{indiv}\right) + 4\sqrt{1-\alpha}\kappa_{bag} + 2(1-\alpha)\mathbb{E}[\dim(C)], \tag{A.17}$$

where $\kappa_{indiv} := \mathbb{E}\left[\|[\mathcal{P}_{\text{span}(M)}, \mathcal{P}_{C^{\star\perp}}]\|_F\right]$ and $f(\kappa_{indiv}) = p_1 \kappa_{indiv}^2 + 2q\kappa_{indiv}$.

Finally, we have the following bag-independent result.

Proposition 10 (Bag Independent Result)**.** *False discovery of the stable tangent space for any $B \geq 2$ is bounded by*

$$\mathbb{E}\left[\text{trace}\left(\mathcal{P}_C \mathcal{P}_{C^{\star\perp}}\right)\right] \leq F + \frac{2q}{\alpha}[1 - \alpha + \sqrt{1-\alpha}] \tag{A.18}$$

and additionally, $\mathbb{E}\left[\text{trace}\left(\mathcal{P}_C \mathcal{P}_{C^{\star\perp}}\right)\right] \leq \mathbb{E}\left[\text{trace}\left(\mathcal{P}_{\hat{C}(\mathcal{D}(n/2))} \mathcal{P}_{C^{\star\perp}}\right)^{1/2}\right]^2 + \frac{3q}{\sqrt{2}\alpha}\sqrt{1-\alpha}$.

Appendix B

PROOFS OF CHAPTER 5

B.1 Proof of Proposition 4

Let $x_1, x_2, \ldots, x_n$ be the scalar random variables from the 1-D markov chain. Then, the naive estimator for θ is given by:

$$\hat{\theta}_n = \frac{1}{n}\text{tr}\left(\begin{pmatrix} x_1 \\ x_2 \\ \ldots \\ x_n \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ \ldots \\ x_n \end{pmatrix}'\right) = \frac{1}{n}(x_1^2 + x_2^2 + \ldots x_n^2)$$

Note that:

$$\lim_{n\to\infty} \mathbb{E}[\hat{\theta}_n] = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[x_i^2] = \lim_{n\to\infty} \frac{1}{n}\text{tr}\left(K_{(n)}^{-1}\right)$$

Thus, we must show that for $b \neq 0$, $\lim_{n\to\infty} \frac{1}{n}\text{tr}\left(K_{(n)}^{-1}\right) > \theta$. By Theorem 1 of [BG96], we have that for any positive definite matrix $\mathcal{A}$ with eigenvalues in the range $[\alpha, \beta]$

$$\text{tr}(\mathcal{A}^{-1}) \geq \begin{pmatrix} \text{tr}(A) & n \end{pmatrix}\begin{pmatrix} \|A\|_F^2 & \text{tr}(A) \\ \beta^2 & \beta \end{pmatrix}^{-1}\begin{pmatrix} n \\ 1 \end{pmatrix}$$

Since $K_{(n)}$ is a Toeplitz tridiagonal matrix, it's largest eigenvalue is precisely characterized as:

$$\|K_{(n)}\|_2 = a - 2b\cos\left(\frac{\pi n}{n+1}\right)$$

We set $\beta_n = \|K_{(n)}\|_2$ and apply this theorem to our setting to obtain

$$\begin{aligned} \frac{1}{n}\text{tr}\left(K_{(n)}^{-1}\right) &\geq \frac{1}{n}\left(\frac{n^2(\beta_n a - \beta^2) + 2n^2b^2 - 2nb^2}{n\beta_n a^2 + 2n\beta_n b\frac{n-1}{n} - n\beta_n^2 a}\right) \\ &= \frac{\beta_n a - \beta^2 + 2b^2 - \frac{2b^2}{n}}{\beta_n a^2 + 2\beta_n b\frac{n-1}{n} - \beta_n^2 a} \end{aligned}$$

Noting that $\lim_{n\to\infty} \beta_n = a + 2b$, it then follows that:

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n}\mathrm{tr}\left(K_{(n)}^{-1}\right) &\geq \lim_{n\to\infty} \frac{\beta_n a - \beta_n^2 + 2b^2 - \frac{2b^2}{n}}{\beta_n a^2 + 2\beta_n b^2 \frac{n-1}{n} - \beta_n^2 a} \\
&= \frac{a(a+2b) - (a+2b)^2 + 2b^2}{(a+2b)a^2 + 2(a+2b)b^2 - (a+2b)^2 a} \\
&= \frac{a+b}{a^2 + ba - 2b^2} > \frac{1}{a}
\end{aligned}$$

Appendix C

PROOF OF CHAPTER 6

C.1 A Numerical Approach for Verfying Assumptions 1, 2, and 3

We begin by considering Assumption 1 in (6.15). Let $f_1 \triangleq 2\max\{\sqrt{p}, \sqrt{2k_u}, \sqrt{2k_x}\gamma, \sqrt{q}\}$, $f_2 \triangleq \max\{\sqrt{2k_u}, \sqrt{2k_x}\gamma\}$ and $\omega \triangleq \max\{\omega_y, \omega_{yx}\}$. Let $Z = (Z_1, Z_2, Z_3, Z_4) \in \mathbb{H}'$ with $\Phi_\gamma(Z_1, Z_2, Z_3, Z_4) = 1$. It is straightforward to check that:

$$\begin{aligned}\Phi_\gamma[\mathcal{P}_{\mathbb{H}'}\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}\mathcal{P}_{\mathbb{H}'}(Z_1, Z_2, Z_3, Z_4)] \;&\geq\; f_1^{-1}\sigma_{\min}(\mathcal{P}_{\mathbb{H}^\star}\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}\mathcal{P}_{\mathbb{H}^\star}) \\ &-\; \max\left\{1, \frac{1}{\gamma}\right\}(\sqrt{3}\omega + \omega + \sqrt{3}\omega^2)f_2\psi^2 \triangleq T_1\end{aligned}$$

Notice that the quantity $\sigma_{\min}(\mathcal{P}_{\mathbb{H}^\star}\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}\mathcal{P}_{\mathbb{H}^\star})$ (and henceforth the quantity T_1) is computable given the population model. Thus a trivial lower bound for α is given by:

$$\inf_{\mathbb{H}'\in U(\omega_y,\omega_{yx})} \chi(\mathbb{H}', \Phi_\gamma) \geq \alpha \geq T_1$$

We now consider Assumption 2 in (6.16). Let $Z = (Z_1, Z_2) \in \mathbb{H}[2,3]'$ with $\Gamma_\gamma(Z_1, Z_2) = 1$. Using triangle inequality, it is straightforward to check the following bound:

$$\begin{aligned}\Gamma_\gamma[\mathcal{P}_{\mathbb{H}[2,3]'}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{G}\mathcal{P}_{\mathbb{H}[2,3]'}(Z_1, Z_2)] \;&\geq\; \min\left\{1, \frac{1}{\gamma}\right\}(\sqrt{3}f_2)^{-1} \\ &\quad\; \sigma_{\min}(\mathcal{P}_{\mathbb{H}[2,3]^\star}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{G}\mathcal{P}_{\mathbb{H}[2,3]^\star}) \\ &-\; \max\left\{1, \frac{1}{\gamma}\right\}(\sqrt{3}\omega + \omega + \sqrt{3}\omega^2)f_2\psi^2 \triangleq T2\end{aligned}$$

Notice that the quantity T_2 is computable giving the population model. Then,

$$\inf_{\mathbb{H}'\in U(\omega_y,\omega_{yx})} \Xi(\mathbb{H}', \Gamma_\gamma) \geq T2$$

Now we consider Assumption 3 in (6.17). Using triangle inequality, it is straightforward to check that:

$$\begin{aligned}\Gamma_\gamma[\mathcal{P}_{\mathbb{H}[2,3]'^\perp}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{G}\mathcal{P}_{\mathbb{H}[2,3]'}(Z_1, Z_2)] \;&\leq\; \sqrt{3}f_2\max\left\{1, \frac{1}{\gamma}\right\} \\ &\quad\; \sigma_{\max}(\mathcal{P}_{\mathbb{H}[2,3]^{\star\perp}}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{G}\mathcal{P}_{\mathbb{H}[2,3]^\star}) \\ &+\; \max\left\{1, \frac{1}{\gamma}\right\}(\sqrt{3}\omega + \omega + \sqrt{3}\omega^2)f_2\psi^2 \triangleq T3\end{aligned}$$

Similarly, the quantity T_3 can be computed given the population model. Then, an upper bound for $\varphi(\mathbb{H}', \Gamma_\gamma)$ is given by:

$$\sup_{\mathbb{H}' \in U(\omega_{yx}, \omega_{yx})} \varphi(\mathbb{H}', \Gamma_\gamma) \leq 1 - \frac{2}{1+\beta} \leq \frac{T3}{T2} \Longrightarrow \beta \leq \frac{2}{1 - \frac{T3}{T2}} - 1$$

C.2 *Proof of Proposition 5*

Proof. We note that:

$$\|\Delta\|_2 \quad \leq \quad \|\Delta D_y\|_2 + \|\Delta L_y\|_2 + \|\Delta \Theta_{yx}\|_2 + \|\Delta \Theta_x\|_2 \leq (3+\gamma)\Phi_\gamma(\Delta)$$

Furthermore, recall that

$$R_{\Sigma^\star}(\mathcal{F}(\Delta)) = \Sigma^{\star -1}\Big[\sum_{k=2}^{\infty}(-\mathcal{F}(\Delta)\Sigma^{\star -1})^k\Big].$$

Using this observation and some algebra, we have that:

$$\begin{aligned}\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\mathcal{F}(\Delta))] \leq m\psi\Big[\sum_{k=2}^{\infty}(\psi\|\Delta\|_2)^k\Big] \quad &\leq \quad m\psi^3\frac{(3+\gamma)^2\Phi_\gamma[\Delta]^2}{1-(3+\gamma)\Phi_\gamma[\Delta]\psi} \\ &\leq \quad 2m\psi C'^2\Phi_\gamma[\Delta]^2\end{aligned}$$

□

C.3 *Proof of Proposition 6*

Proof. The proof of this result uses Brouwer's fixed-point theorem, and is inspired by the proof of a similar result in [Rav+11a; CPW12]. The optimality conditions of (20) (main paper) suggest that there exist Lagrange multipliers $Q_{D_y} \in \mathcal{W}$, $Q_{T_y} \in T_y'^{\perp}$, and $Q_{T_{yx}} \in T_{yx}'^{\perp}$ such that

$$\begin{aligned}[\Sigma_n - \tilde{\Theta}^{-1}]_y + Q_{D_y} = 0; \quad [\Sigma_n - \tilde{\Theta}^{-1}]_y + Q_{T_y} \quad &\in \quad \lambda_n \partial\|\tilde{L}_y\|_\star \\ [\Sigma_n - \tilde{\Theta}^{-1}]_{yx} + Q_{T_{yx}} \in -\lambda_n\gamma\partial\|\tilde{\Theta}_{yx}\|_\star; \quad [\Sigma_n - \tilde{\Theta}^{-1}]_x \quad &= \quad 0\end{aligned}$$

Letting the SVD of $\tilde{L}$ and $\tilde{\Theta}_{yx}$ be given by $\tilde{L}_y = \bar{U}\bar{D}\bar{V}'$ and $\tilde{\Theta}_{yx} = \breve{U}\breve{D}\breve{V}'$ respectively, and $Z \triangleq (0, \ \lambda_n\bar{U}\bar{V}', \ -\lambda_n\gamma\breve{U}\breve{V}', \ 0)$, we can restrict the optimality conditions of (15) (main paper) to the space $\mathbb{H}'$ to obtain, $\mathcal{P}_{\mathbb{H}'}\mathcal{F}^\dagger(\Sigma_n - \tilde{\Theta}^{-1}) = Z$. Further, by appealing to the matrix inversion lemma, this condition can be restated as $\mathcal{P}_{\mathbb{H}_\mathcal{M}}\mathcal{F}^\dagger(E_n - R_{\Sigma^\star}(\mathcal{F}\Delta) + \mathbb{I}^\star\mathcal{F}(\Delta)) = Z$. Based on the Fisher information Assumption 1 in (6.15) (main paper), the optimum of (20) (main paper) is unique

(this is because the Hessian of the negative log-likelihood term is positive definite restricted to the tangent space constraints). Moreover, using standard Lagrangian duality, one can show that the set of variables $(\tilde{\Theta}, \tilde{D}_y, \tilde{L}_y)$ that satisfy the restricted optimality conditions are unique. Consider the following function $S(\underline{\delta})$ restricted to $\underline{\delta} \in \mathcal{W} \times T'_y \times T'_{yx} \times \mathbb{S}^q$ with $\rho(T(L^\star_y), T'_y) \leq \omega_y$ and $\rho(T(\Theta^\star_{yx}), T'_{yx}) \leq \omega_{yx}$:

$$S(\underline{\delta}) = \underline{\delta} - (\mathcal{P}_{\mathbb{H}'}\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}\mathcal{P}_{\mathbb{H}'})^{-1}\Big(\mathcal{P}_{\mathbb{H}'}\mathcal{F}^\dagger[E_n \quad - \quad R_{\Sigma^\star}\mathcal{F}(\underline{\delta} + C_T) \\ +\mathbb{I}^\star\mathcal{F}(\underline{\delta} + C_T)] - Z\Big)$$

The function $S(\underline{\delta})$ is well-defined since the operator $\mathcal{P}_{\mathbb{H}'}\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}\mathcal{P}_{\mathbb{H}'}$ is bijective due to Fisher information Assumption 1 in (6.15) (main paper). As a result, $\underline{\delta}$ is a fixed point of $S(\underline{\delta})$ if and only if $\mathcal{P}_{\mathbb{H}'}\mathcal{F}^\dagger[E_n - R_{\Sigma^\star}(\mathcal{F}(\underline{\delta} + C_T)) + \mathbb{I}^\star\mathcal{F}(\underline{\delta} + C_T)] = Z$. Since the pair $(\tilde{\Theta}, \tilde{D}_y, \tilde{L}_y)$ are the unique solution to (20) (main paper), the only fixed point of S is $\mathcal{P}_{\mathbb{H}'}[\Delta]$. Next we show that this unique optimum lives inside the ball $\mathbb{B}_{r_1^u, r_2^u} = \{\underline{\delta} \mid \max\{\|\delta_2\|_2, \frac{1}{\gamma}\|\delta_3\|_2\} \leq r_1^u, \max\{\|\delta_1\|_2, \|\delta_4\|_2\} \leq r_2^u \ \underline{\delta} \in \mathbb{H}'\}$. In particular, we show that under the map S, the image of $\mathbb{B}_{r_1^u, r_2^u}$ lies in $\mathbb{B}_{r_1^u, r_2^u}$ and appeal to Brouwer's fixed point theorem to conclude that $\mathcal{P}_{\mathbb{H}'}[\Delta] \in \mathbb{B}_{r_1^u, r_2^u}$. For $\underline{\delta} \in \mathbb{B}_{r_1^u, r_2^u}$, the first component of $S(\underline{\delta})$, denoted by $S(\underline{\delta})_1$, can be bounded as follows:

$$\begin{aligned} \|S(\underline{\delta})_1\|_2 &= \Big\|\Big[(\mathcal{P}_{\mathbb{H}'}\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}\mathcal{P}_{\mathbb{H}'})^{-1}\Big(\mathcal{P}_{\mathbb{H}'}\mathcal{F}^\dagger[E_n - R_{\Sigma^\star}(\mathcal{F}(\underline{\delta} + C_T)) \\ &+ \mathbb{I}^\star\mathcal{F}C_T] + Z\Big)\Big]_1\Big\|_2 \leq \frac{2}{\alpha}\Big[\Phi_\gamma[\mathcal{F}^\dagger(E_n + \mathbb{I}^\star\mathcal{F}(C_T))]\Big] \\ &+ \frac{2}{\alpha}\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\underline{\delta} + C_T)] \leq \frac{r_2^u}{2} + \frac{2}{\alpha}\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\underline{\delta} + C_T)] \end{aligned}$$

The first inequality holds because of Fisher Information Assumption 1 in (6.15) (main paper), and the properties that $\Phi_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}}(.)] \leq 2\Phi_\gamma(.)$ (since projecting into the tangent space of a low-rank matrix variety increases the spectral norm by a factor of at most two) and $\Phi_\gamma(Z) = \lambda_n$. Moreover, since $r_1^u \leq \frac{1}{4C'}$, we have $\Phi_\gamma(\underline{\delta} + C_T) \leq \Phi_\gamma(\underline{\delta}) + \Phi_\gamma(C_T) \leq 2r_1^u \leq \frac{1}{2C'}$. Moreover, $r_1^u \leq r_2^u \max\{1 + \frac{\kappa}{2}, \frac{\alpha}{8}\}$. We can now appeal to Proposition 5 to obtain:

$$\begin{aligned} \frac{2}{\alpha}\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\underline{\delta} + C_T)] &\leq \frac{4}{\alpha}m\psi C'^2[\Phi_\gamma(\underline{\delta} + C_T)]^2 \\ &\leq \frac{16}{\alpha}m\psi C'^2(r_2^u)^2\max\{1 + \frac{\kappa}{2}, \frac{\alpha}{8}\}^2 \\ &\leq \frac{r_2^u}{2} \end{aligned}$$

Thus, we conclude that $\|S(\delta)_1\|_2 \leq r_2^u$. Similarly, we check that:

$$\begin{aligned}\|[S(\underline{\delta})_2]\|_2 &= \Big\|\Big[(\mathcal{P}_{\mathbb{H}'}\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}\mathcal{P}_{\mathbb{H}'})^{-1}\Big(\mathcal{P}_{\mathbb{H}'}\mathcal{F}^\dagger[E_n - R_{\Sigma^\star}(\mathcal{F}(\underline{\delta}+C_T)) \\ &+ \mathbb{I}^\star\mathcal{F}C_T] + Z\Big)\Big]_2\Big\|_2 \leq \frac{2}{\alpha}\Big[\Phi_\gamma[\mathcal{F}^\dagger(E_n + \mathbb{I}^\star\mathcal{F}(C_T)] + \lambda_n\Big] \\ &+ \frac{2}{\alpha}\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\underline{\delta}+C_T)] \leq \frac{r_1^u}{2} + \frac{2}{\alpha}\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\underline{\delta}+C_T)] \leq r_1^u\end{aligned}$$

Using a similar approach, we can conclude that $\frac{1}{\gamma}\|S(\delta)_3\|_2 \leq r_1^u$ and $\|S(\delta)_3\|_2 \leq r_2^u$. Therefore, Brouwer's fixed point theorem suggests that $\mathcal{P}_{\mathbb{H}'}(\Delta) \in \mathcal{B}_{r_1^u, r_2^u}$. Hence,
$\|\Delta_1\|_2 \leq r_2^u$, $\|\Delta_4\|_2 \leq r_2^u$,
$\|\Delta_2\|_2 \leq \|\mathcal{P}_{\mathbb{H}'[2]}(\Delta_2)\|_2 + \|\mathcal{P}_{\mathbb{H}'[2]^\perp}(\Delta_2)\|_2 \leq 2r_1^u$, and
$\frac{1}{\gamma}\|\Delta_3\|_2 \leq \frac{1}{\gamma}\|\mathcal{P}_{\mathbb{H}'[3]}(\Delta_3)\|_2 + \frac{1}{\gamma}\|\mathcal{P}_{\mathbb{H}'[3]^\perp}(\Delta_2)\|_2 \leq 2r_1^u$. □

C.4 Proof of Proposition 7

Below, we outline our proof strategy:

1. We proceed by analyzing (19) (main paper) with additional constraints that the variables L_y, and Θ_{yx} belong to the algebraic varieties low-rank matrices (specified by rank of $L_y^\star$, and $\Theta_{yx}^\star$) , and that the tangent spaces $T(L_y)$, $T(\Theta_{yx})$ are close to the nominal tangent spaces $T(L_y^\star)$, and $T(\Theta_{yx}^\star)$ respectively. We prove that under suitable conditions on the minimum nonzero singular value of $L_y^\star$, and minimum nonzero singular value of $\Theta_{yx}^\star$, any optimum pair of variables (Θ, D_y, L_y) of this non-convex program are smooth points of the underlying varieties; that is $\text{rank}(L_y) = \text{rank}(L_y^\star)$ and $\text{rank}(\Theta_{yx}) = \text{rank}(\Theta_{yx}^\star)$. Further, we show that L_y has the same inertia as $L_y^\star$ so that $L_y \succeq 0$.

2. Conclusions of the previous step imply the the variety constraints can be "linearized" at the optimum of the non-convex program to obtain tangent-space constraints. Under the specified conditions on the regularization parameter λ_n, we prove that with high probability, the unique optimum of this "linearized" program coincides with the global optimum of the non-convex program.

3. Finally, we show that the tangent-space constraints of the linearized program are inactive at the optimum. Therefore the optimal solution of (19) (main paper) has the property that with high probability: $\text{rank}(\bar{L}_y) = \text{rank}(L_y^\star)$ and $\text{rank}(\bar{\Theta}_{yx}) = \text{rank}(\Theta_{yx}^\star)$. Since $\bar{L}_y \succeq 0$, we conclude that the variables $(\bar{\Theta}, \bar{D}_y, \bar{L}_y)$ are the unique optimum of (6.4).

Variety Constrained Program

We begin by considering a variety-constrained optimization program. Letting $(M, N, P, Q) \subset \mathbb{S}^p \times \mathbb{S}^p \times \mathbb{R}^{p\times q} \times \mathbb{S}^q$, we denote $\mathcal{P}_{[2,3]}(M, N, P, Q) = (N, P) \subset \mathbb{S}^p \times \mathbb{R}^{p\times q}$. The variety-constrained optimization program is given by:

$$(\Theta^{\mathcal{M}}, D_y^{\mathcal{M}}, L_y^{\mathcal{M}}) = \underset{\substack{\Theta\in\mathbb{S}^{q+p},\ \Theta\succ 0\\ D_y, L_y\in\mathbb{S}^p}}{\operatorname{argmin}} \quad -\ell(\Theta; \mathcal{D}_n^+) + \lambda_n[\|L_y\|_\star + \gamma\|\Theta_{yx}\|_\star]$$
$$\text{s.t.} \quad \Theta_y = D_y - L_y, (\Theta, D_y, L_y) \in \mathcal{M}. \tag{C.1}$$

Here, the set $\mathcal{M} = \mathcal{M}_1 \cap \mathcal{M}_2$, where the sets $\mathcal{M}_1$ and $\mathcal{M}_2$ are given by:

$$\begin{aligned}
\mathcal{M}_1 \quad &\triangleq \quad \Big\{(\Theta, D_y, L_y) \in \mathbb{S}^{(p+q)} \times \mathbb{S}^p \times \mathbb{S}^p \Big| D_y \text{ is diagonal}, \ \ \operatorname{rank}(L_y) \le \operatorname{rank}(L_y^\star) \\
&\operatorname{rank} \ (\Theta_{yx}) \le \operatorname{rank}(\Theta_{yx}^\star); \|\mathcal{P}_{T(L_y^\star)^\perp}(L_y - L_y^\star)\|_2 \le \frac{\lambda_n}{2\psi^2} \\
&\|\mathcal{P}_{T(\Theta_{yx}^\star)^\perp}(\Theta_{yx} - \Theta_{yx}^\star)\|_2 \le \frac{\lambda_n}{2\psi^2}\Big\} \\
\mathcal{M}_2 \quad &\triangleq \quad \Big\{(\Theta, D_y, L_y) \in \mathbb{S}^{(p+q)} \times \mathbb{S}^p \times \mathbb{S}^p \Big| \\
&\|\mathbb{I}^\star\mathcal{F}(\Delta)\|_2 \le 6\bar{m}\psi^2\lambda_n\Big(\frac{8}{\alpha\kappa} + \frac{4}{\alpha} + \frac{1}{\kappa}\Big)\Big\},
\end{aligned}$$

The optimization program (C.1) is non-convex due to the rank constraints $\operatorname{rank}(L_y) \le \operatorname{rank}(L_y^\star)$ and $\operatorname{rank}(\Theta_{yx}) \le \operatorname{rank}(\Theta_{yx}^\star)$ in the set $\mathcal{M}$. These constraints ensure that the matrices L_y, and Θ_{yx} belong to appropriate varieties. The constraints in $\mathcal{M}$ along $T(L_y^\star)^\perp$ and $T(\Theta_{yx}^\star)^\perp$ ensure that the tangent spaces $T(L_y)$ and $T(\Theta_{yx})$ are "close" to $T(L_y^\star)$ and $T(\Theta_{yx}^\star)$ respectively. Finally, the last conditions roughly controls the error. We begin by proving the following useful proposition:

Proposition 11. *Let* (Θ, D_y, L_y) *be a set of feasible variables of* (C.1). *Let* $\Delta = (D_y - D_y^\star, L_y - L_y^\star, \Theta_{yx} - \Theta_{yx}^\star, \Theta_x - \Theta_x^\star)$ *and recall that* $C_1' = \frac{2\bar{m}m}{\kappa\alpha}\Big(6\psi^2 + \frac{5}{\alpha}\psi^2 + \frac{46\psi^2\kappa}{\alpha} + \kappa\Big) + \frac{1}{\psi^2}$. *Then,* $\Phi_\gamma[\Delta] \le C_1'\lambda_n$

Proof. Let $\mathbb{H}^\star = \mathcal{W} \times T(L_y^\star) \times T(\Theta_{yx}^\star) \times \mathbb{S}^q$. Then,

$$\begin{aligned}
\Phi_\gamma[\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}\mathcal{P}_{\mathbb{H}^\star}(\Delta)] \ &\le \ \Phi_\gamma[\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}(\Delta)] + \Phi_\gamma[\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}\mathcal{P}_{\mathbb{H}^{\star\perp}}(\Delta)] \\
&\le \ 6\bar{m}m\psi^2\lambda_n\Big(\frac{8}{\alpha\kappa} + \frac{4}{\alpha} + \frac{1}{\kappa}\Big) \\
&+ \ m\psi^2\Big(\frac{\omega_y\lambda_n}{2\psi^2} + \frac{\omega_{yx}\lambda_n}{2\psi^2}\Big) \\
&\le \ \frac{\bar{m}m\lambda_n}{\kappa}\Big(6\psi^2 + \frac{24}{\alpha}\psi^2 + \frac{48\psi^2\kappa}{\alpha} + \kappa\Big)
\end{aligned}$$

Since $\Phi_\gamma[\mathcal{P}_{\mathbb{H}^\star}(\cdot)] \leq 2\Phi_\gamma(\cdot)$, we have that $\Phi_\gamma[\mathcal{P}_{\mathbb{H}^\star}\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}\mathcal{P}_{\mathbb{H}^\star}(\Delta)] \leq \frac{2\bar{m}m\lambda_n}{\kappa\alpha}\Big(6\psi^2 + \frac{24}{\alpha}\psi^2 + \frac{48\psi^2\kappa}{\alpha} + \kappa\Big)$. Consequently, we apply Fisher Information Assumption 1 in (6.15) (main paper) to conclude that $\Phi_\gamma[\mathcal{P}_{\mathbb{H}^\star}(\Delta)] \leq \frac{2\bar{m}m\lambda_n}{\kappa\alpha}\Big(6\psi^2 + \frac{24}{\alpha}\psi^2 + \frac{48\psi^2\kappa}{\alpha} + \kappa\Big)$. Moreover:

$$\begin{aligned}\Phi_\gamma[\Delta] \leq \Phi_\gamma[\mathcal{P}_{\mathbb{H}^\star}(\Delta)] + \Phi_\gamma[\mathcal{P}_{\mathbb{H}^{\star\perp}}(\Delta)] \quad &\leq \quad \frac{2\bar{m}m\lambda_n}{\kappa\alpha}\Big(6\psi^2 + \frac{24}{\alpha}\psi^2 + \frac{48\psi^2\kappa}{\alpha} + \kappa\Big) \\ &+ \quad \frac{\lambda_n}{\psi^2} = C_1'\lambda_n\end{aligned}$$

□

Proposition 11 leads to powerful implications. In particular, under additional conditions on the minimum nonzero singular values of $L_y^\star$ and $\Theta_{yx}^\star$, any feasible set of variables (Θ, D_y, L_y) of (C.1) has two key properties: (a) The variables (Θ_{yx}, L_y) are smooth points of the underlying varieties, (b) The constraints in $\mathcal{M}$ along $T(L_y^\star)^\perp$ and $T(\Theta_{yx}^\star)^\perp$ are locally inactive at Θ_{yx} and L_y. These properties, among others, are proved in the following corollary.

Corollary 1. *Consider any feasible variables* (Θ, D_y, L_y) *of* (C.1). *Let* σ_y *be the smallest nonzero singular value of* $L_y^\star$ *and* σ_{yx} *be the smallest nonzero singular value of* $\Theta_{yx}^\star$. *Let* $\mathbb{H}' = \mathcal{W} \times T(L_y) \times T(\Theta_{yx}) \times \mathbb{S}^q$ *and* $C_{T'} = \mathcal{P}_{\mathbb{H}'^\perp}(0, L_y^\star, \Theta_{yx}^\star, 0)$. *Furthermore, recall that* $C_1' = \frac{2\bar{m}m}{\kappa\alpha}\Big(6\psi^2 + \frac{24}{\alpha}\psi^2 + \frac{48\psi^2\kappa}{\alpha} + \kappa\Big) + \frac{1}{\psi^2}$, $C_2' = \frac{4}{\alpha}(1 + \frac{2}{\kappa})$, $C_{\sigma_y}' = C_1'^2\psi^2\max\{2\kappa + 1, \frac{2}{C_2'\psi^2} + 1\}$ *and* $C_{\sigma_{yx}}' = C_1'^2\psi^2\max\{2\kappa + \frac{\kappa}{\gamma}, \frac{2}{C_2'\psi^2} + \frac{\kappa}{\gamma}\}$. *Suppose that the following inequalities are met:* $\sigma_y \geq \frac{m}{\omega_y}C_{\sigma_y}\lambda_n$, $\sigma_{yx} \geq \frac{m\gamma^2}{\omega_{yx}}C_{\sigma_{yx}}'\lambda_n$. *Then,*

1. L_y *and* Θ_{yx} *are smooth points of their underlying varieties, i.e.* $rank(L_y) = rank(L_y^\star)$, $rank(\Theta_{yx}) = rank(\Theta_{yx}^\star)$*; Moreover* L_y *has the same inertia as* $L_y^\star$.
2. $\|\mathcal{P}_{T(L_y^\star)^\perp}(L_y - L_y^\star)\|_2 \leq \frac{\lambda_n}{48m\psi^2}$ *and* $\|\mathcal{P}_{T(\Theta_{yx}^\star)^\perp}(\Theta_{yx} - \Theta_{yx}^\star)\|_2 \leq \frac{\lambda_n}{48m\psi^2}$
3. $\rho(T(L_y), T(L_y^\star)) \leq \omega_y$; $\rho(T(\Theta_{yx}), T(\Theta_{yx}^\star)) \leq \omega_{yx}$*; that is, the tangent spaces at* L_y *and* Θ_{yx} *is "close" to the tangent space* $L_y^\star$ *and* $\Theta_{yx}^\star$.
4. $\Phi_\gamma[C_{T'}] \leq \min\{\frac{\lambda_n}{\kappa\psi^2}, C_2'\lambda_n\}$

Proof. We note the following relations before proving each step: $C_1' \geq \frac{1}{\psi^2} \geq \frac{1}{m\psi^2}$, $\omega_y, \omega_{yx} \in (0, 1)$, and $\kappa \geq 6$. We also appeal to the results of regarding perturbation

analysis of the low-rank matrix variety [Bac08].

1. Based on the assumptions regarding the minimum nonzero singular values of $L_y^\star$ and $\Theta_{yx}^\star$, one can check that:

$$\begin{aligned}\sigma_y &\geq \frac{C_1'^2\lambda_n}{\omega_y} m\psi^2(\kappa+1) \geq \frac{C_1'\lambda_n}{\omega_y}(2\kappa+1) \geq 8\|L - L_y^\star\|_2 \\ \sigma_{yx} &\geq \frac{C_1'^2\lambda_n}{\omega_{yx}}\gamma^2 m\psi^2\left(\frac{6\beta}{\gamma}+2\kappa\right) \geq 8\|\Theta_{yx} - \Theta_{yx}^\star\|_2\end{aligned}$$

Combining these results and Proposition 11, we conclude that L_y and Θ_{yx} are smooth points of their respective varieties, i.e. $\operatorname{rank}(L_y) = \operatorname{rank}(L_y^\star)$, and $\operatorname{rank}(\Theta_{yx}) = \operatorname{rank}(\Theta_{yx}^\star)$. Furthermore, L_y has the same inertia as $L_y^\star$.

2. Since $\sigma_y \geq 8\|L_y - L_y^\star\|_2$, and $\sigma_{yx} \geq 8\|\Theta_{yx} - \Theta_{yx}^\star\|_2$, we can appeal to Proposition 2.2 of [CPW12] to conclude that the constraints in $\mathcal{M}$ along $\mathcal{P}_{T(L_y^\star)^\perp}$ and $\mathcal{P}_{T(\Theta_{yx}^\star)^\perp}$ are strictly feasible:

$$\begin{aligned}\|\mathcal{P}_{T(L_y^\star)^\perp}(L_y - L_y^\star)\|_2 &\leq \frac{\|L_y - L_y^\star\|_2^2}{\sigma_y} \leq \frac{\lambda_n}{48m\psi^2} \\ \|\mathcal{P}_{T(\Theta_{yx}^\star)^\perp}(\Theta_{yx} - \Theta_{yx}^\star)\|_2 &\leq \frac{\|\Theta_{yx} - \Theta_{yx}^\star\|_2^2}{\sigma_{yx}} \leq \frac{\lambda_n}{48m\psi^2}\end{aligned}$$

3. Appealing to Proposition 2.1 of [CPW12], we prove that the tangent spaces $T(L_y)$ and $T(\Theta_{yx})$ are close to $T(L_y^\star)$ and $T(\Theta_{yx}^\star)$ respectively:

$$\begin{aligned}\rho(T(L_y), T(L_y^\star)) &\leq \frac{2\|L_y - L_y^\star\|_2}{\sigma_y} \leq \frac{2C_1'\lambda_n\omega_y}{C_1'^2\lambda_n m\psi^2(2\kappa+1)} \leq \omega_y \\ \rho(T(\Theta_{yx}), T(\Theta_{yx}^\star)) &\leq \frac{2\|\Theta_{yx} - \Theta_{yx}^\star\|_2}{\sigma_{yx}} \leq \frac{2C_1'\lambda_n\gamma\omega_{yx}}{\frac{C_1'^2\lambda_n}{\omega_{yx}}\gamma^2 m\psi^2\left(\frac{\kappa}{\gamma}+2\kappa\right)} \leq \omega_{yx}\end{aligned}$$

4. Letting σ_y' and σ_{yx}' be the minimum nonzero singular value of L_y and Θ_{yx} respectively, one can check that:

$$\sigma_y' \geq \sigma_y - \|L_y - L_y^\star\|_2 \geq 8C_1'\lambda_n \geq 8\|L_y - L_y^\star\|_2$$

$$\sigma_{yx}' \geq \sigma_{yx} - \|\Theta_{yx} - \Theta_{yx}^\star\|_2 \geq 8C_1'\lambda_n\gamma \geq 8\|\Theta_{yx} - \Theta_{yx}^\star\|_2$$

Once again appealing to Proposition 2.2 of [CPW12] and simple algebra, we have:

$$\begin{aligned}\Phi_\gamma(C_{T'}) &\leq m\|\mathcal{P}_{T(L_y)^\perp}(L_y - L_y^\star)\|_2 + m\|\mathcal{P}_{T(\Theta_{yx})^\perp}(\Theta_{yx} - \Theta_{yx}^\star)\|_2 \\ &\leq m\frac{\|L_y - L_y^\star\|_2^2}{\sigma'_y} + m\frac{\|\Theta_{yx} - \Theta_{yx}^\star\|_2^2}{\sigma'_{yx}} \leq \min\left\{\frac{\lambda_n}{\kappa\psi^2}, C'_2\lambda_n\right\}\end{aligned}$$

□

Variety Constrained Program to Tangent Space Constrained Program

Consider any optimal solution $(\Theta^{\mathcal{M}}, D_y^{\mathcal{M}}, L_y^{\mathcal{M}})$ of (C.1). In Corollary 1, we concluded that the variables $(\Theta_{yx}^{\mathcal{M}}, L_y^{\mathcal{M}})$ are smooth points of their respective varieties. As a result, the rank constraints $\text{rank}(L_y) \leq \text{rank}(L_y^\star)$ and $\text{rank}(\Theta_{yx}) \leq \text{rank}(\Theta_{yx}^\star)$ can be "linearized" to $L_y \in T(L_y^{\mathcal{M}})$ and $\Theta_{yx} \in T(\Theta_{yx}^{\mathcal{M}})$ respectively. Since all the remaining constraints are convex, the optimum of this linearized program is also the optimum of (C.1). Moreover, we once more appeal to Corollary 1 to conclude that the constraints in $\mathcal{M}$ along $\mathcal{P}_{T(L_y^\star)^\perp}$ and $\mathcal{P}_{T(\Theta_{yx}^\star)^\perp}$ are strictly feasible at $(\Theta^{\mathcal{M}}, D_y^{\mathcal{M}}, L_y^{\mathcal{M}})$. As a result, these constraints are locally inactive and can be removed without changing the optimum. Therefore the constraint $(\Theta^{\mathcal{M}}, D_y^{\mathcal{M}}, L_y^{\mathcal{M}}) \in \mathcal{M}_1$ is inactive and can be removed. We now argue that the constraint $(\Theta^{\mathcal{M}}, D_y^{\mathcal{M}}, L_y^{\mathcal{M}}) \in \mathcal{M}_2$ in (C.1) can also removed in this "linearized" convex program. In particular, letting $\mathbb{H}_{\mathcal{M}} \triangleq \mathcal{W} \times T(L_y^{\mathcal{M}}) \times T(\Theta_{yx}^{\mathcal{M}}) \times \mathbb{S}^q$, consider the following convex optimization program:

$$\begin{aligned}(\tilde{\Theta}, \tilde{D}_y, \tilde{L}_y) = \underset{\substack{\Theta\in\mathbb{S}^{q+p},\ \Theta\succ 0 \\ D_y, L_y\in\mathbb{S}^p}}{\text{argmin}} \quad & -\ell(\Theta; \mathcal{D}_n^+) + \lambda_n[\|L_y\|_\star + \gamma\|\Theta_{yx}\|_\star] \\ \text{s.t.} \quad & \Theta_y = D_y - L_y,\ (D_y, L_y, \Theta_{yx}, \Theta_x) \in \mathbb{H}_{\mathcal{M}} \qquad \text{(C.2)}\end{aligned}$$

We prove that under conditions imposed on the regularization parameter λ_n, the pair of variables $(\Theta^{\mathcal{M}}, D_y^{\mathcal{M}}, L_y^{\mathcal{M}})$ is the unique optimum of (C.2). That is, we show that

1. $\|\mathbb{I}^\star\mathcal{F}(\Delta)\|_2 < 6\bar{m}\psi^2\lambda_n\left(\frac{8}{\alpha\kappa} + \frac{4}{\alpha} + \frac{1}{\kappa}\right)$

Appealing to Corollary 1 and Proposition 8, we have that $\Phi_\gamma[\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}C_{T_{\mathcal{M}}}] \leq \frac{\lambda_n}{\kappa}$, $\Phi_\gamma[C_{T_{\mathcal{M}}}] \leq C'_2\lambda_n$ and (with high probability) $\Phi_\gamma[\mathcal{F}^\dagger E_n] \leq \frac{\lambda_n}{\kappa}$. Consequently, based on the bound on λ_n in assumption of Theorem 5, it is straightforward to show that $r_1^u \leq \min\left\{\frac{1}{4C'}, \frac{\alpha}{32\max\{1+\frac{\kappa}{2}, \frac{\alpha}{8}\}^2 m\psi C'^2}\right\}$ so that $\Phi_\gamma[\Delta] \leq \frac{1}{2C'}$. Hence by Proposition 6, we have that $\|\Delta_1\|_2, \|\Delta_4\|_2 \leq r_2^u < r_1^u$, $\|\Delta_2\|_2 \leq 2r_1^u$ and $\|\Delta\|_3 \leq 2\gamma r_1^u$. Therefore:

$$\begin{aligned}\|\mathbb{I}^\star\mathcal{F}(\Delta)\|_2 &\leq \psi^2(\|\Delta_1\|_2 + \|\Delta_2\|_2 + \|\Delta_3\|_2 + \|\Delta_4\|_2) \\ &< 6\bar{m}\psi^2 r_1^u \leq 6\bar{m}\psi^2\lambda_n\left(\frac{8}{\alpha\kappa} + \frac{4}{\alpha} + \frac{1}{\kappa}\right)\end{aligned}$$

From Tangent Space Constraints to the Original Problem

The optimality conditions of (C.2) suggest that there exist Lagrange multipliers $Q_{D_y} \in \mathcal{W}$, $Q_{T_y} \in T(L_y^{\mathcal{M}})^{\perp}$, and $Q_{T_{yx}} \in T(\Theta_{yx}^{\mathcal{M}})^{\perp}$ such that

$$\begin{aligned} [\Sigma_n - \tilde{\Theta}^{-1}]_y + Q_{D_y} = 0; \;\; [\Sigma_n - \tilde{\Theta}^{-1}]_y + Q_{T_y} &\in \lambda_n \partial \|\tilde{L}_y\|_\star \\ [\Sigma_n - \tilde{\Theta}^{-1}]_{yx} + Q_{T_{yx}} \in -\lambda_n \gamma \partial \|\tilde{\Theta}_{yx}\|_\star; \;\; [\Sigma_n - \tilde{\Theta}^{-1}]_x &= 0 \end{aligned}$$

Letting the SVD of $\tilde{L}_y$ and $\tilde{\Theta}_{yx}$ be given by $\tilde{L}_y = \bar{U}\bar{O}\bar{V}'$ and $\tilde{\Theta}_{yx} = \breve{U}\breve{O}\breve{V}'$ respectively, and $Z \triangleq (0, \;\; \lambda_n \bar{U}\bar{V}', \;\; -\lambda_n \gamma \breve{U}\breve{V}', \;\; 0)$, we can restrict the optimality conditions to the space $\mathbb{H}_{\mathcal{M}}$ to obtain, $\mathcal{P}_{\mathbb{H}_{\mathcal{M}}}\mathcal{F}^{\dagger}(\Sigma_n - \tilde{\Theta}^{-1}) = Z$. We proceed by proving that the variables $(\tilde{\Theta}, \tilde{D}_y, \tilde{L}_y)$ satisfy the optimality conditions of the original convex program (6.4). That is:

1. $\mathcal{P}_{\mathbb{H}_{\mathcal{M}}}\mathcal{F}^{\dagger}(\Sigma_n - \tilde{\Theta}^{-1}) = Z$
2. $\max\left\{\|\mathcal{P}_{T_y'^{\perp}}(\Sigma_n - \tilde{\Theta}^{-1})_y\|_2, \frac{1}{\gamma}\|\mathcal{P}_{T_{yx}'^{\perp}}(\Sigma_n - \tilde{\Theta}^{-1})_{yx}\|_2\right\} < \lambda_n$

It is clear that the first condition is satisfied since the pair $(\tilde{\Theta}, \tilde{S}_y, \tilde{L}_y)$ is optimum for (C.2). To prove that the second condition, we must prove that $\Gamma_\gamma[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}^{\perp}[2,3]}\mathcal{G}^{\dagger}(\Sigma_n - \tilde{\Theta}^{-1})] < \lambda_n$. In particular, denoting $\Delta = (\tilde{D}_y - D_y^{\star}, \tilde{L}_y - L_y^{\star}, \tilde{\Theta}_{yx} - \Theta_{yx}^{\star}, \tilde{\Theta}_x - \Theta_x^{\star})$ we show that:

$$\begin{aligned} \Gamma_\gamma[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}^{\perp}[2,3]}\mathcal{G}^{\dagger}\mathbb{I}^{\star}\mathcal{G}\mathcal{P}_{\mathbb{H}_{\mathcal{M}}[2,3]}(\Delta)] &< \lambda_n - \Phi_\gamma[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}^{\perp}}\mathcal{F}^{\dagger}E_n] \\ &- \Phi_\gamma[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}^{\perp}}\mathcal{F}^{\dagger}R_{\Sigma^{\star}}(\mathcal{F}(\Delta))] \\ &- \Phi_\gamma[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}^{\perp}}\mathcal{F}^{\dagger}\mathbb{I}^{\star}\mathcal{F}C_{T_{\mathcal{M}}}] \\ &- \Gamma_\gamma[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}[2,3]^{\perp}}\mathcal{G}^{\dagger}\mathbb{I}^{\star}\mathcal{F}(\Delta_1, 0, 0, \Delta_4)] \end{aligned} \tag{C.3}$$

Using the fact that $\Gamma_\gamma[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}[2,3]^{\perp}}\mathcal{G}^{\dagger}(N)] \leq \Phi_\gamma[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}^{\perp}}\mathcal{F}^{\dagger}(N)]$ for any matrix $N \in \mathbb{S}^{p+q}$, this would in turn imply that:

$$\begin{aligned} \Gamma_\gamma[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}[2,3]^{\perp}}\mathcal{G}^{\dagger}\mathbb{I}^{\star}\mathcal{G}\mathcal{P}_{\mathbb{H}_{\mathcal{M}}[2,3]}(\Delta)] &< \lambda_n - \Gamma_\gamma[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}[2,3]^{\perp}}\mathcal{G}^{\dagger}E_n] \\ &- \Gamma_\gamma[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}[2,3]^{\perp}}\mathcal{G}^{\dagger}R_{\Sigma^{\star}}(\mathcal{F}(\Delta))] \\ &- \Gamma_\gamma[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}[2,3]^{\perp}}\mathcal{G}^{\dagger}\mathbb{I}^{\star}\mathcal{F}C_{T_{\mathcal{M}}}] \\ &- \Gamma_\gamma[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}[2,3]^{\perp}}\mathcal{G}^{\dagger}\mathbb{I}^{\star}\mathcal{F}(\Delta_1, 0, 0, \Delta_4)] \end{aligned} \tag{C.4}$$

Indeed (4) implies that the second optimality condition is satisfied. So we focus on showing that (4) is satisfied. Since $\Phi_\gamma[\Delta] \leq \frac{1}{2C'}$, we can appeal to Proposition 5 and the bound on λ_n to conclude $\Phi_\gamma[\mathcal{F}^{\dagger}R_{\Sigma^{\star}}(\mathcal{F}(\Delta))] \leq 2m\psi C'^2\Phi_\gamma[\Delta]^2 \leq$

$2m\psi C'^2 C_1'^2 \lambda_n^2 \leq \frac{\lambda_n}{\kappa}$. Using the first optimality condition, the fact that projecting into tangent spaces with respect to rank variety increase the spectral norm by at most a factor of two (i.e. $\Phi_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}}(\cdot)] \leq 2\Phi_\gamma(\cdot)$),the fact that $\Gamma_\gamma[\mathcal{G}^\dagger(\cdot)] \leq \Phi_\gamma[\mathcal{F}^\dagger(\cdot)]$, and that $\kappa = \beta(6 + \frac{16\psi^2 m}{\alpha})$, we have that:

$$\begin{aligned}
\Gamma_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}[2,3]}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{G}\mathcal{P}_{\mathbb{H}_\mathcal{M}[2,3]}(\Delta)] &\leq \lambda_n + 2\Gamma_\gamma[\mathcal{G}^\dagger R_{\Sigma^\star}(\Delta)] + 2\Gamma_\gamma[\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{F}C_{T_M}] \\
&+ 2\Gamma_\gamma[\mathcal{G}^\dagger E_n] + \Gamma_\gamma[\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{F}(\Delta_1, 0, 0, \Delta_4)] \\
&\leq \lambda_n + 2\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\Delta)] + 2\Phi_\gamma[\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}C_{T_M}] \\
&+ 2\Phi_\gamma[\mathcal{F}^\dagger E_n] + \Phi_\gamma[\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}(\Delta_1, 0, 0, \Delta_4)] \\
&\leq \lambda_n + \frac{\lambda_n}{\beta}
\end{aligned}$$

Applying Fisher Information Assumption 2 in (6.16), we obtain:

$$\begin{aligned}
\Gamma_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}[2,3]^\perp}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{G}\mathcal{P}_{\mathbb{H}_\mathcal{M}[2,3]}(\Delta)] &\leq \frac{(\beta+1)\lambda_n}{\beta}\Big(1 - \frac{2}{\beta+1}\Big) = \lambda_n - \frac{\lambda_n}{\beta} \\
&< \lambda_n - \frac{\lambda_n}{2\beta} \\
&\leq \lambda_n - \Phi_\gamma[\mathcal{F}^\dagger R_\Sigma(\mathcal{F}(\Delta))] - \Phi_\gamma[\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}C_{T_\mathcal{M}}] \\
&- \Phi_\gamma[\mathcal{F}^\dagger E_n] - \Gamma_\gamma[\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{F}(\Delta_1, 0, 0, \Delta_4)] \\
&\leq \lambda_n - \Phi_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}^\perp}\mathcal{F}^\dagger R_{\Sigma^\star}(\mathcal{F}(\Delta))] \\
&- \Phi_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}^\perp}\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}C_{T_\mathcal{M}}] \\
&- \Phi_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}^\perp}\mathcal{F}^\dagger E_n] \\
&- \Gamma_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}[2,3]^\perp}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{F}(\Delta_1, 0, 0, \Delta_4)]
\end{aligned}$$

Here, we used the fact that $\|\mathcal{P}_{T^\perp}(.)\|_2 \leq \|.\|_2$ for a tangent space T of the low-rank matrix variety.

C.5 *Proof of Proposition 8*

We must study the rate of convergence of the sample covariance matrix to the population covariance matrix. The following result from [DS01] plays a key role in obtaining this result.

Proposition 12. *Given natural numbers n, p with $p \leq n$, Let Γ be a $p \times n$ matrix with i.i.d Gaussian entries that have zero-mean and variance $\frac{1}{n}$. Then the largest and smallest singular values $\sigma_1(\Gamma)$ and $\sigma_p(\Gamma)$ of Γ are such that:*

$$\max\left\{Prob[\sigma_1(\Gamma) \leq 1 + \sqrt{\frac{p}{n}} + t], Prob[\sigma_p(\Gamma) \leq 1 - \sqrt{\frac{p}{n}} - t]\right\}$$

We now proceed with proving Proposition 8. First, note that $\Phi_\gamma[\mathcal{F}^\dagger E_n] \leq m\|\Sigma_n - \Sigma^\star\|_2$. Using Proposition 12 and the fact that $\frac{\lambda_n}{m\kappa} \leq 8\psi$ and $n \geq \frac{64\kappa^2(p+q)m^2\psi^2}{\lambda_n^2}$, the following bound holds: $\Pr[m\|\Sigma_n - \Sigma^\star\|_2 \geq \frac{\lambda_n}{\kappa}] \leq 2\exp\Big\{-\frac{n\lambda_n^2}{128\kappa^2m^2\psi^2}\Big\}$. Thus, $\Phi_\gamma[\mathcal{F}^\dagger E_n] \leq \frac{\lambda_n}{\kappa}$ with probability greater than $1 - 2\exp\Big\{-\frac{n\lambda_n^2}{128\kappa^2m^2\psi^2}\Big\}$

C.6 Consistency of the Convex Program (6.18)

In this section, we prove the consistency of convex program (6.18) for estimating a factor model. We first introduce some notation. We define the linear operator: $\tilde{\mathcal{F}} : \mathbb{S}^p \times \mathbb{S}^p \to \mathbb{S}^p$ and its adjoint $\tilde{\mathcal{F}}^\dagger : \mathbb{S}^p \to \mathbb{S}^p \times \mathbb{S}^p$ as follows:

$$\tilde{\mathcal{F}}(M, K) \triangleq M - K, \qquad \tilde{\mathcal{F}}^\dagger(Q) \triangleq (Q, Q) \tag{C.5}$$

We consider a population composite factor model (6.4) $y = \mathcal{A}^\star x + \mathcal{B}_u^\star \zeta_u + \epsilon$ underlying a pair of random vectors $(y, x) \in \mathbb{R}^p \times \mathbb{R}^q$, with $\text{rank}(\mathcal{A}^\star) = k_x$, $\mathcal{B}_u^\star \in \mathbb{R}^{p\times k_u}$, and column-space$(\mathcal{A}^\star) \cap$ column-space$(\mathcal{B}_u^\star) = \{0\}$. As the convex relaxation (6.18) is solved in the precision matrix parametrization, the conditions for our theorems are more naturally stated in terms of the joint precision matrix $\Theta^\star \in \mathbb{S}^{p+q}$, $\Theta^\star \succ 0$ of (y, x). The algebraic aspects of the parameters underlying the factor model translate to algebraic properties of submatrices of $\Theta^\star$. In particular, the submatrix $\Theta_{yx}^\star$ has rank equal to k_x, and the submatrix $\Theta_y^\star$ is decomposable as $D_y^\star - L_y^\star$ with $D_y^\star$ being diagonal and $L_y^\star \succeq 0$ having rank equal to k_u. Finally, the transversality of column-space$(\mathcal{A}^\star)$ and column-space$(\mathcal{B}_u^\star)$ translates to the fact that column-space$(\Theta_{yx}^\star) \cap$ column-space$(L_y^\star) = \{0\}$ have a transverse intersection. We consider the factor model underlying the random vector $y \in \mathbb{R}^p$ that is induced upon marginalization of x. In particular, the precision matrix of y is given by $\tilde{\Theta}_y^\star = D_y^\star - [L_y^\star + \Theta_{yx}^\star(\Theta_x^\star)^{-1}\Theta_{xy}^\star]$. To learn an accurate factor model, we seek an estimate $(\hat{\tilde{D}}_y, \hat{\tilde{L}}_y)$ from the convex program (6.18) such that $\text{rank}(\hat{\tilde{L}}_y = \text{rank}(L_y^\star + \Theta_{yx}^\star\Theta_x^{\star -1}\Theta_{xy}^\star)$, and the errors $\|\hat{\tilde{D}}_y - D_y^\star\|_2$, $\|\hat{\tilde{L}}_y - [L_y^\star + \Theta_{yx}^\star(\Theta_x^\star)^{-1}\Theta_{xy}^\star]\|_2$ are small.

Following the same reasoning as the Fisher information conditions for consistency of the convex program (6.4), A natural set of conditions on the population Fisher

information at $\tilde{\Theta}_y^\star$ defined as $\mathbb{I}_y^\star = (\tilde{\Theta}_y^\star)^{-1} \otimes (\tilde{\Theta}_y^\star)^{-1}$ are given by:

$$\text{Assumption 4} \quad : \quad \inf_{\mathbb{H}' \in \tilde{U}(\tilde{\omega}_y)} \tilde{\chi}(\mathbb{H}', \tilde{\Phi}) \geq \tilde{\alpha}, \quad \text{for some } \tilde{\alpha} > 0 \tag{C.6}$$

$$\text{Assumption 5} \quad : \quad \inf_{\mathbb{H}' \in \tilde{U}(\tilde{\omega}_y)} \tilde{\Xi}(\mathbb{H}') > 0 \tag{C.7}$$

$$\text{Assumption 6} \quad : \quad \sup_{\mathbb{H}' \in \tilde{U}(\tilde{\omega}_y)} \tilde{\varphi}(\mathbb{H}') \leq 1 - \frac{2}{\tilde{\beta}+1} \quad \text{for some } \tilde{\beta} \geq 2, \tag{C.8}$$

where,

$$\begin{aligned}
\tilde{\chi}(\mathbb{H}, \|.\|_\Upsilon) &\triangleq \min_{\substack{Z \in \mathbb{H} \\ \|Z\|_\Upsilon = 1}} \|\mathcal{P}_{\mathbb{H}} \tilde{\mathcal{F}}^\dagger \mathbb{I}_y^\star \tilde{\mathcal{F}} \mathcal{P}_{\mathbb{H}}(Z)\|_\Upsilon \\
\tilde{\Xi}(\mathbb{H}) &\triangleq \min_{\substack{Z \in \mathbb{H}[2] \\ \|Z\|_2 = 1}} \|\mathcal{P}_{\mathbb{H}[2]} \mathbb{I}_y^\star \mathcal{P}_{\mathbb{H}[2]}(Z)\|_2 \\
\tilde{\varphi}(\mathbb{H}) &\triangleq \max_{\substack{Z \in \mathbb{H}[2] \\ \|Z\|_2 = 1}} \|\mathcal{P}_{\mathbb{H}^\perp[2]} \mathbb{I}_y^\star \mathcal{P}_{\mathbb{H}[2]} (\mathcal{P}_{\mathbb{H}[2]} \mathbb{I}_y^\star \mathcal{P}_{\mathbb{H}[2]})^{-1}(Z)\|_2 \\
\tilde{U}(\tilde{\omega}_y) &\triangleq \left\{ \mathcal{W} \times T' \mid \rho(T', T(L_y^\star + \Theta_{yx}^\star (\Theta_x^\star)^{-1} \Theta_{xy}^\star)) \leq \tilde{\omega}_y \right\} \\
\tilde{\Phi}(D, L) &\triangleq \max \{ \|D\|_2, \|L\|_2 \} .
\end{aligned}$$

Assumption 4 controls the gain of the Fisher information $\mathbb{I}_y^\star$ restricted to appropriate subspaces and Assumption 5 and 6 are in the spirit of irrepresentability conditions. As the variety of low-rank matrices is locally curved around $T(L_y^\star + \Theta_{yx}^\star (\Theta_x^\star)^{-1} \Theta_{xy}^\star)$, we control the Fisher information $\mathbb{I}_y^\star$ at nearby tangent spaces T' where $\rho(T', T(L_y^\star + \Theta_{yx}^\star (\Theta_x^\star)^{-1} \Theta_{xy}^\star) \leq \tilde{\omega}_y$. We also note that measuring the gains of Fisher information $\mathbb{I}_y^\star$ with the norm $\tilde{\Phi}$ and $\|\cdot\|_2$ is natural as these are closely tied with dual norm of the regularizer trace$(\tilde{L}_y)$ in (6.18).

We present a theorem of consistency of the convex relaxation (6.18) under Assumptions 4, 5 and 6. We let σ denote the minimum nonzero singular value of $L_y^\star + \Theta_{yx}^\star (\Theta_x^\star)^{-1} \Theta_{xy}^\star$. The proof strategy is similar in spirit to the strategy for proving the consistency of the convex relaxation (6.4).

Theorem 7. *Suppose that there exists $\tilde{\alpha} > 0$, $\tilde{\beta} \geq 2$, $\tilde{\omega}_y \in (0, 1)$ so that the population Fisher information $\mathbb{I}_y^\star$ satisfies Assumptions 4, 5 and 6 in* (C.6),(C.7) *and* (C.8). *Suppose that the following conditions hold:*

1. $n \gtrsim \left\lceil \frac{\tilde{\beta}^2}{\tilde{\alpha}^2} \right\rceil (p)$
2. $\tilde{\lambda}_n \sim \frac{\tilde{\beta}}{\tilde{\alpha}} \sqrt{\frac{p}{n}}$

3. $\sigma \gtrsim \frac{\tilde{\beta}}{\tilde{\alpha}^5 \tilde{\omega}_y} \tilde{\lambda}_n$

Then with probability greater than $1 - 2\exp\left\{ - C\frac{\tilde{\alpha}}{\tilde{\beta}} n \tilde{\lambda}_n^2 \right\}$, *the optimal solution* $(\hat{\Theta}, \hat{\hat{D}}_y, \hat{\hat{L}}_y)$ *of* (6.18) *with i.i.d. observations* $\mathcal{D}_n = \{y^{(i)}\}_{i=1}^n$ *satisfies the following properties:*

1. $rank(\hat{\hat{L}}_y) = rank(L_y^\star + \Theta_{yx}^\star (\Theta_x^\star)^{-1} \Theta_{xy}^\star)$

2. $\|\hat{\hat{D}}_y - D_y^\star\|_2 \lesssim \frac{\tilde{\lambda}_n}{\tilde{\alpha}^2}$, $\|\hat{\hat{L}}_y - L_y^\star - \Theta_{yx}^\star (\Theta_x^\star)^{-1} \Theta_{xy}^\star\|_2 \lesssim \frac{\tilde{\lambda}_n}{\tilde{\alpha}^2}$

BIBLIOGRAPHY

[Agh+14] A. AghaKouchak et al. “Global warming and changes in risk and concurrent climate extremes: Insights from the 2014 California drought”. In: *Geophysical Research Letters* 41 (2014), pp. 8847–8852.

[AL13] G. Allen and Z. Liu. “A local poisson graphical model for inferring networks from sequencing data”. In: *IEEE Transactions on NanoBioscience* 12 (2013), pp. 1–10.

[Bac08] F. Bach. “Consistency of trace norm minimization”. In: *Journal of Machine Learning Research* 9 (2008), pp. 1019–1048.

[BD07] P. Bickel and K. Doksum. *Mathematical statistics, basic ideas and selected topics*. Prentice-Hall, 2007.

[Bes75] J. Besag. “Statistical Analysis of Non-Lattice Data”. In: *The Statistician* 24 (1975), pp. 179–195.

[BG11] P. Bühlmann and S. van de Geer. *Statistics for high dimensional data*. Springer Series in Statistics, 2011.

[BG73] A. Björck and G. Golub. “Numerical Methods for Computing Angles Between Linear Subspaces”. In: *Mathematics of Computations* 27 (1973).

[BG96] Z Bai and G. Golub. “Bounds for the trace of the inverse and the determinant of symmetric positive definite matrices”. In: *Annals of Numerical Mathematics* 4 (1996), pp. 29–38.

[BHH03] B. Bazartseren, G. Hildebrandt, and K. Holz. “Short-term water level prediction using neural networks and neuro-fuzzy approach”. In: *Neurocomputing* 55 (2003), pp. 439–450.

[BP08] T. Barnett and D. Pierce. “When will Lake Mead go dry?” In: *Water Resources Research* 44 (2008).

[Cav+08] C. Cavalho et al. “High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics”. In: *Journal of the American Statistical Association* 103 (2008), pp. 1438–1456.

[Cha+12] V. Chandrasekaran et al. “Convex geometry of linear inverse problems”. In: *Foundations of Computational Mathematics* 12 (2012), pp. 805–849.

[Che+15] C. Cheng et al. “Heuristic methods for reservoir monthly inflow forecasting: A case study of Xinfengjiang reservoir in Pearl river, China”. In: *Water* 7 (2015), pp. 4477–4495.

[Chr+06] N. Christensen et al. “Effects of climate change on the hydrology and water resources of the Colorado basin”. In: *Climate Change* 62 (2006), pp. 337–363.

[CL04] N. Christensen and D. Lettenmaier. “A multimodel ensemble approach to climate change impacts on the hydrology and water resources of the Colorado River Basin”. In: *Hydrology of Earth System Sciences* 3 (2004), pp. 1–44.

[CPW12] V. Chandrasekaran, P. A. Parillo, and A. S. Willsky. “Latent variable graphical model selection via convex optimization”. In: *Annals of Statistics* 40 (2012), pp. 1935–1967.

[CR09] E. Candés and B. Recht. “Exact Matrix Completion via Convex Optimization”. In: *Foundations of Computational Mathematics* 27 (2009).

[CTT17] Y. Choi, J. Taylor, and R. Tibshirani. “Selecting the number of principal components: estimation of the true rank of a noisy matrix”. In: *Annals of Statistics* 45 (2017), pp. 2590–2617.

[CWS15] S. Chen, D. Witten, and A. Shojaie. “Selection and Estimation for Mixed Graphical Models”. In: *Biometrica* 102 (2015), pp. 47–64.

[DIS17] Matthew M Dunlop, Marco A Iglesias, and Andrew M Stuart. “Hierarchical Bayesian level set inversion”. In: *Statistics and Computing* 27.6 (2017), pp. 1555–1584.

[DS01] K. Davidson and S. Szarek. “Local operator theory, random matrices and Banach space”. In: *Handbook of the Geometry of Banach Spaces* 1 (2001), pp. 317–366.

[Dua+08] M. Duarte et al. “Single-pixel imaging via compressive sampling”. In: *IEEE Signal Processing Magazine* 25 (2008), pp. 83–91.

[Fam14] J. Famiglietti. “The global groundwater crisis”. In: *Nature Climate Change* 4 (2014), pp. 945–948.

[Faz02] M. Fazel. “Matrix rank minimization with applications”. PhD thesis. Stanford, 2002.

[FHT08] J. Friedman, T. Hastie, and R. Tibshirani. “Sparse inverse covariance estimation with the graphical lasso”. In: *Biostatistics* 9 (2008), pp. 432–441.

[FJM17] B. Frot, L. Jostins, and G. McVean. “Latent variable model selection for Gaussian conditional random fields”. In: *arXiv:1512.06412* (2017).

[FL11] R. Fa and R. Lamare. “Reduced-Rank STAP Algorithms using Joint Iterative Optimization of Filters”. In: *IEEE Transactions on Aerospace and Electronic Systems* 47 (2011), pp. 1668–1684.

[GA14] D. Griffin and K. Anchukaitis. “How unusual is the 2012–2014 California drought?” In: *Geophysical Research Letters* 41 (2014), pp. 9017–9023.

[Gol+92] D. Goldberg et al. “Reduced-Rank STAP Algorithms using Joint Iterative Optimization of Filters”. In: *Using collaborative filtering to weave an information tapestry* 35 (1992), pp. 61–70.

[Gra99] W. Graf. “Dam nation: A geographic census of American dams and their large-scale hydrologic impacts”. In: *Water Resources Research* 35 (1999), pp. 1305–1311.

[Har95] J. Harris. *Algebraic Geometry: A First Course*. Berlin: Springer-Verlog, 1995.

[HE07] M. Hoerling and J. Eischeid. “Past peak water in the West”. In: *Southwest Hydrology* 6 (2007), pp. 18–19.

[Hot02] H. Hotelling. “Relations between two sets of variants”. In: *Biometrika* 28 (2002), pp. 321–377.

[How+14] R. Howitt et al. “Economic analysis of the 2014 drought for California agriculture”. In: *Center for Watershed Sciences, University of California, Davis* (2014).

[HS09] M. A. Herman and T. Strohmer. “High-resolution radar via compressed sensing”. In: *IEEE Transactions on Signal Processing* 57 (2009), pp. 2275–2284.

[HT09] H. Hoffling and R. Tibshirani. “Estimation of Sparse Binary Pairwise Markov Networks using Pseudo-likelihoods”. In: *Journal of Machine Learning Research* 10 (2009), pp. 883–906.

[HT14] H. Hoffling and R. Tibshirani. “The Optimal Hard Threshold for Singular Values is $4/sqrt3$”. In: *IEEE Transcations on Information Theory* 60 (2014), pp. 5040–5053.

[HTF09] T. Hastie, T. Tibshirani, and R. Friedman. *The elements of statistical learning*. Springer, 2009.

[Jor04] Jordan. “Graphical Models”. In: *Statistical Science* 42 (2004), pp. 140–155.

[Kay98] S. Kay. *Modern spectral estimation, theory and application*. Prentice-Hall, 1998.

[Led40] W. Ledermann. “On a problem concerning matrices with variable diagonal elements”. In: *Proceeding of Royal Society Edinburgh* 60 (1940), pp. 1–17.

[LHR14] S. Lasanen, J. Huttunen, and L. Roininen. “Whittle-Matérn priors for Bayesian statistical inversion with applications in electrical impedance tomography”. In: *Inverse Problems Imaging* 8 (2014), pp. 561–586.

[LJ08] P. Liang and M. Jordan. “An Asymptotic Analysis of Generative, Discriminative, and Pseudolikelihood Estimators”. In: *International Conference in Machine Learning* (2008).

[LL18] Z. Liu and X. Lin. “A Geometric Perspective on the Power of Principal Component Association Tests in Multiple Phenotype Studies”. In: *Journal of the American Statistical Association* (2018), pp. 1–36.

[LRL11] Finn Lindgren, Håvard Rue, and Johan Lindström. “An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.4 (2011), pp. 423–498.

[LRW10] H. Liu, K. Roeder, and L. Wasserman. “Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models”. In: *International Conference on Neural Information Processing Systems* 2 (2010), pp. 1432–1440.

[Lus+08] M. Lustig et al. “Compressed sensing MRI”. In: *IEEE Signal Processing Magazine* 25 (2008), pp. 72–82.

[LV09] Z. Liu and L. Vandenberghe. “Interior-point method for nuclear norm approximation with application to system identification”. In: *SIAM Journal on Matrix Analysis and Applications* 31 (2009), pp. 1235–1256.

[Man03] D. Manolakis. “Detection algorithms for hyperspectral imaging applications: a signal processing perspective”. In: *IEEE Workshop on Advances in Techniques for Analysis of Remotely Sensed Data* (2003), pp. 378–384.

[MB06] N. Meinshausen and P. Bühlmann. “High dimensional graphs and variable selection with the Lasso”. In: *Annals of Statistics* 34 (2006), pp. 1436–1462.

[MB10] N. Meinshausen and P. Bühlmann. “Stability Selection”. In: *Journal of Royal Statistical Methodology (Series B)* 72 (2010), pp. 417–473.

[MN83] P. McCullagh and J. Nedler. *Generalized Linear Models*. Chapman-Hall, 1983.

[MP97] M. Mesbahi and G. Papavassilopoulos. “On the rank minimization problem over a positive semidefinite linear matrix inequality”. In: *IEEE Transactions on Automatic Control* 42 (1997), pp. 239–243.

[Nat95] B. Nataranjan. “Sparse approximate solutions to linear systems”. In: *SIAM Journal of Computing* 24 (1995), pp. 227–234.

[NG91] L. Nash and P. Gleick. “The sensitivity of stream flow in the Colorado Basin to climatic changes”. In: *Journal of Hydrology* 124 (1991), pp. 221–241.

[NG93] L. Nash and P. Gleick. “The Colorado Basin and climate change”. In: *Rep. EPA 230-R-93-009, United States Environmental Protection Agency, Washington, DC* (1993).

[NW15] A. Nazemi and H. S. Wheater. “On inclusion of water resource management in earth system models – Part 1: Problem definition and representation of water demand”. In: *Hydrology and Earth System Science* 19 (2015), pp. 33–61.

[Pha89] R. Phatafod. “Riverflow and reservoir storage models”. In: *Mathematical Computational Modeling* 12 (1989), pp. 1057–1077.

[PK94] Y. Pati and T. Kailath. “Phase-shifting masks for microlithography: Automated design and mask requirements”. In: *Journal of the Optical Society of America A* 11 (1994).

[Rav+11a] P. Ravikumar et al. “High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence”. In: *Electronic Journal of Statistics* 4 (2011), pp. 935–980.

[Rav+11b] P. Ravikumar et al. “High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence”. In: *Electronic Journal of Statistics* 5 (2011), pp. 935–980.

[RFP10] B. Recht, M. Fazel, and P. Parrilo. “Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization”. In: *SIAM Review* 52 (2010), pp. 471–501.

[RW83] P. Revelle and P. Waggoner. “Effects of carbon dioxide-induced climatic change on water supplies in the western United States”. In: *Changing Climate, Carbon Dioxide Assessment Communication* (1983).

[Sau+12] J. Saunderson et al. “Diagonal and low-rank matrix decompositions, correlation matrices, and ellipsoid fitting”. In: *SIAM Journal on Matrix Analysis* 33 (2012), pp. 1395–1416.

[Sha04] A. Shapiro. “Identifiability of factor analysis: Some results and open problems”. In: *Linear Algebra Applications* 15 (1904), pp. 201–292.

[Sha82a] A. Shapiro. “Rank-reducibility of a symmetric matrix and sampling theory of minimum trace factor analysis”. In: *Psychometrika* 47 (1982), pp. 187–199.

[Sha82b] A. Shapiro. “Weighted minimum trace factor analysis”. In: *Psychometrika* 47 (1982), pp. 243–264.

[Sha82c] A. Shapiro. “Weighted minimum trace factor analysis”. In: *Psychometrika* 77 (1982), pp. 243–264.

[SJ05] J. Rennie Srebro N. and T. Jaakkola. “Maximum Margin Matrix Factorizations”. In: *Advances in Neural Information Processing Systems* 17 (2005).

[Sol+16] K. Solander et al. “Simulating human water regulation: The development of an optimal complexity, climate-adaptive reservoir management model for an LSM”. In: *Journal of Hydrometeorology* 17 (2016), pp. 725–744.

[Spe04] C. Spearman. “‘General intelligence’, objectively determined and measured”. In: *American Journal of Psychology* 15 (1904), pp. 201–292.

[SS05] N. Srebro and A. Shraibman. “Rank, trace-norm and max-norm”. In: *Proceedings of the 18th Annual Conference on Learning Theory* (2005), pp. 545–560.

[SS13] R. Shah and J. Samworth. “Variable selection and error control: another look at stability selection”. In: *Journal of Royal Statistical Methodology (Series B)* 75 (2013), pp. 55–80.

[SS18] J. Song and S. Shin. “Stability approach to selecting the number of principal components”. In: *Computational Statistics* 33 (2018), pp. 1923–1938.

[Tae+17] A. Taeb, J.T. Reager, M. Turmon, and V. Chandrasekaran. “A Satistical Model of the California Reservoir System”. In: *Water Resources Research* 53.11 (2017), pp. 9721–9739. DOI: 10.1002/2017WR020412.

[TC15] A. Taeb and V. Chandrasekaran. “Sufficient dimension reduction and modeling responses conditioned on covariates: an integrated approach via convex optimization”. In: *arXiv:1508.03852* (2015).

[TC18] A. Taeb and V. Chandrasekaran. “Interpreting Latent Variables in Factor Analysis via Convex Optimization”. In: *Mathematical Programming* 167.1 (2018), pp. 129–154. DOI: 10.1007/s10107-017-1187-7.

[Tho+15] D. Thompson et al. “Real-time remote detection and measurement for airborne imaging spectroscopy: a case study with methane”. In: *Atmospheric Measurement Techniques* 8 (2015), pp. 4383–4397.

[TSC19] A. Taeb, P. Shah, and V. Chandrasekaran. “False Discovery and Its Control in Low-Rank Estimation”. In: *arXiv* (2019). DOI: arXiv:1810.08595v1.

[TTT16] K. C Toh, M. J. Todd, and R. H. Tutuncu. “SDPT3 - A MATLAB software package for semidefinite-quadratic-linear programming”. 2016.

[Tur60] G. Turin. “An introduction to matched filters”. In: *IRE Transactions on Information Theory* 6 (1960), pp. 311–329.

[Wac03] H. Wackernagel. *TMultivariate geostatistics: an introduction with applications*. Springer, 2003.

[Wai09] M. J. Wainwright. "Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (Lasso)". In: *IEEE Transactions on Information Theory* 55 (2009), pp. 2183–2202.

[Wai14] M. Wainwright. "Structured regularizers for high-dimensional problems: Statistical and computational issues". In: *Annual Review of Statistics and its Applications* 1 (2014), pp. 233–253.

[Wis+10] D. Wisser et al. "Reconstructing 20th century global hydrography: A contribution to the Global Terrestrial Network". In: *Hydrology (GTN-H)* (2010).

[Yan+15] E. Yang et al. "On Graphical Models via Univariate Exponential Family Distributions". In: *Journal of Machine Learning Research* 16 (2015), pp. 3813–3847.

[YH95] Benjamini Y. and Y. Hockberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of Royal Statistical Society Series B* 57 (1995), pp. 289–300.

[YL07] M. Yuan and Y. Lin. "Model selection and estimation in the Gaussian graphical model". In: *Biometrika* 94 (2007), pp. 19–35.

[ZK14] L Zhang and S. Kim. "Learning Gene Networks under SNP Perturbations Using eQTL Datasets". In: *PLOS Computational Biology* 10.1 (2014).

[ZY06] P. Zhao and B. Yu. "On model selection consistency of the lasso". In: *Journal of Machine Learning Research* 7 (2006), pp. 2541–2567.