Covariance Estimation with Missing and Dependent Data

by

Roger Fan

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2020

Doctoral Committee:

Assistant Professor Yuekai Sun, Co-chair
Professor Shuheng Zhou, University of California, Riverside, Co-chair
Associate Professor Jian Kang
Professor Kerby Shedden

Roger Fan

rogerfan@umich.edu

ORCID id: 0000-0001-5744-4399

ACKNOWLEDGEMENTS

I would like to express my immense gratitude to my advisers Shuheng Zhou and Yuekai Sun for all of their guidance and mentorship during this process. The both provided an immense amount of knowledge, attention, and patience that were instrumental in my progress as a PhD student. They were both also exemplary role models as researchers and professionals, and I am deeply thankful and appreciative of all the support and encouragement I have received from them over the last several years. I especially want to thank Shuheng for continuing to provide such active help and for always being available for discussions and comments even after taking a position at the University of California, Riverside. I am very thankful for the additional effort and time committed to providind such supportive and detailed advising remotely. This thesis is joint work with both Yuekai Sun and Shuheng Zhou.

I would like to thank my committee member Professor Kerby Shedden for many insightul discussions and comments on this and other work, and especially for including and working with me on a previous project that lead to and inspired significant portions of this work. That work was lead by Michael Hornstein, who I would like to thank for including me on the initial project involving joint mean and covariance estimation.

I would like to thank Professor Jian Kang for agreeing to serve on my thesis committee and for this valuable time and discussions about this work.

I would like to thank my friend and colleague Byoungwook Jang for the innumerable hours of discussion and joint work that we have had together, as well for all the memories and experiences that his friendship has provided. Chapter 2, as well as portions of Chapters

3 and 4, are joint work with him. I would also like to thank Jonathan Goetz and Timothy Lycurgus for their unwavering support, encouragement, and friendship.

This thesis is based upon work supported by the National Science Foundation (NSF) under grants DMS-1316731 and DMS-1830247. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the NSF.

I would like to thank Stefania D'Amico for her mentorship and supervision at the Federal Reserve Bank of Chicago as I first learned to conduct academic research.

I would like to thank Megan Davis for her many years of friendship and support.

And, of course, I would like to thank my parents for their unconditional and unwavering love and support, as well as for inspiring me to pursue math and academia. I give all credit to them for building and constantly providing foundations that I have always been able to rely on.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

The estimation of conditional dependence graphs and precision matrices is one of the most relevant problems in modern statistics. There is a large body of work for estimation with fully observed and independent data. These are, however, often unrealistic assumptions for real-world data applications. So extensions are needed to accommodate data complications more suited for actual data analysis. In this thesis we address the methodology, theory, and applications of covariance estimation with these complications.

We focus on a data setting with both dependence and missingness. To model this, we use a matrix-variate model with a Kronecker product covariance structure and missing values. This model allows for correlations to exist both between the rows and between the columns, and is commonly used in fields as diverse as genetics, neuroimaging, psychology, and environmental science, where estimating and/or accounting for dependence is often a primary concern. We develop prototypical column- and row-wise precision matrix estimators for single data matrices with missing data. We show initial concentration of measure bounds on entry-wise consistency for data with mean structure and multiplicative errors, and develop corresponding rates of convergence for the joint mean and covariance estimation in high-dimensional settings.

To implement these estimators, we first solve a general implementation issue with graphical Lasso-type estimators designed for use with noisy and missing data. These estimators often result in non-positive semidefinite input matrices to the graphical Lasso, which can result in pathological optimization issues. We show how this problem can be fixed with modified objective functions and develop a feasible and efficient algorithm for solving the graphical Lasso with these modifications. This algorithm can be used not only for our

method, but also to implement a wide variety of graphical Lasso extensions that involve non-positive semidefinite inputs.

Finally, we use our methods to explore a dataset of voting records from the U.S. Senate, where we expect there to be connections both between similar or opposed senators as well as between bills that may share characteristics or topics. This dataset exhibits missing data and has mean structure due to the two-party system, and in particular we are interested in estimating relationships beyond just those dictated by this party structure.

CHAPTER I

Introduction

The estimation of conditional dependence graphs and precision matrices is one of the most relevant problems in modern statistics, with application domains spanning many fields ranging from genetics to neuroscience to economics and political science to environmental science. When assuming that data come from a multivariate Gaussian model, standard results show that estimating the conditional dependence graph can be done by estimating the structure of the associated inverse covariance, or precision, matrix. A wide body of work proposes methods to perform this estimation, most of which use a similar ℓ_1-penalized likelihood approach which we will refer to as the graphical LASSO (*Banerjee et al.*, 2008; *Friedman et al.*, 2008; *Ravikumar et al.*, 2011; *Rothman et al.*, 2008; *Zhou et al.*, 2010). These procedures generally follow the form of first constructing a positive semi-definite (PSD) estimate of the covariance or correlation matrix with favorable convergence properties, then using that as an input into a graphical LASSO or nodewise regression program that produces a sparse estimate of the precision matrix.

Classically, graphical Lasso-type estimators are M-estimators that take the form

$$\hat{\Theta} \in \underset{\Theta \succeq 0}{\arg\min} \left\{ \operatorname{tr}(\hat{\Gamma}_n \Theta) - \log\det(\Theta) + g_\lambda(\Theta) \right\},$$

given an input covariance estimate $\hat{\Gamma}_n$ and a penalty function $g_\lambda(\Theta)$, which is often the element-wise ℓ_1-norm.

Many of the results in this area focus on estimating the conditional dependencies between variables with fully observed independent and identically distributed (i.i.d.) observations with zero-mean. These are, however, clearly unrealistic assumptions for many real-world data applications. So modern methods often extend these methods to include data complications more suited for the complex data encountered in modern data analysis.

Note that we focus on graphical Lasso estimators rather than the closely related neighborhood selection estimators, that generally utilize multiple node-wise regressions that are later combined with an AND or OR rule (*Meinshausen and Bühlmann*, 2006; *Yuan and Lin*, 2007; *Zhou et al.*, 2011). This is due to the masking strategies that we use throughout the work to handle missingness, that are in general more difficult to adapt to the nodewise estimators. Future work includes developing the methodology and theoretical results necessary to extend these methods to their corresponding node-wise estimators.

1.1 Precision Matrix Estimation with Noisy and Missing Data

One of the complications that arises with real data is missing, noisy, or incomplete data, and so it is natural to extend these methods to those settings. Many methods have been developed along these lines. *Lounici* (2014) and *Loh and Wainwright* (2015) perform graphical model estimation with missing or corrupted data using a modification to the covariance estimate first presented by *Hwang* (1986). There are many methods dealing with error-in-variables, including *Rudelson and Zhou* (2017), *Park et al.* (2017), *Belloni et al.* (2017) and *Greenewald et al.* (2017), which allow for various types of dependent noise.

A common issue, however, that arises in many of these models is that the input covariance estimate is no longer PSD. This makes it difficult to ensure that the optimization works as-desired, and until our work the effects of these non-PSD inputs were not well-understood. Note that these types of non-PSD inputs also arise in the closely related Gaussian copula models used for semiparametric graph estimation and estimation with ordinal or mixed data, which therefore face similar problems (*Liu et al.*, 2012; *Fan et al.*, 2017; *Feng and Ning*, 2019).

In Chapter II we explore this problem of performing precision matrix estimation with noisy and missing data, and in particular we develop optimization objectives and algorithms for use when the input matrices are not PSD. In order to ensure that these estimators are well-behaving with non-PSD input, we impose a side constraint of the form $\rho(\Theta) < R$, where ρ is a convex function, similar to the one suggested in *Loh and Wainwright* (2015). Here we focus on the estimator using the operator norm as a side constraint

$$\hat{\Theta} \in \underset{\Theta \succeq 0, \|\Theta\|_2 \leq R}{\arg\min} \left\{ \operatorname{tr}(\hat{\Gamma}_n \Theta) - \log\det(\Theta) + g_\lambda(\Theta) \right\}. \tag{1.1}$$

Unfortunately, this additional constraint precludes using existing methods to solve the penalized objective with non-PSD input. To close this gap, we develop an alternating direction method of multipliers (ADMM) algorithm to solve (1.1) efficiently. Although we focus on applications with missing and noisy data, this objective and algorithm can be applied to any graphical LASSO-type estimator with non-PSD input, and provides a practical method for solving these problems without relying on existing PSD-based solvers that can become degenerate in these scenarios.

We conduct empirical studies comparing this method to several other precision matrix estimators and show that it compares favorably. We also explore the use of non-convex regularizers such as SCAD and MCP, of which much recent work has focused on due to their favorable model selection properties (*Fan and Li*, 2001; *Zhang*, 2010; *Loh and Wainwright*, 2017) with fully observed data. Our simulation study reveals several trends in performance that are not present in the fully observed case, in particular showing that non-convex penalties tend to introduce undesireable instability and estimation error in the non-PSD setting.

1.2 Covariance Estimation for Matrix-Variate Data with Missing Values and Mean Structure

Another data complication of particular interest is adding dependence between the observations. The matrix-variate model, which allows for dependence along both axes of the data matrix, is an increasingly popular model for doing this. Its ability to model relationships between observations as well as covariates makes it useful for analyzing data with temporal, geographical, or other network relationships between them. Thus, applications for matrix-variate models often arise in biology, genetics, economics, climate science, and many other fields, where it is important to use methods that are at least robust to these types of dependencies. Specific applications include genomic data, where sample-side correlations can both be intentional due to the experimental design or unintentional as described by *Efron* (2009), spatial-temporal data such as brain imaging or environmental data (*Genton*, 2007; *Wang et al.*, 2016; *Qiu et al.*, 2016; *Shvartsman et al.*, 2018; *Glanz and Carvalho*, 2018), or panel data from surveys over time or financial panels (*Hatfield and Zaslavsky*, 2018; *Wang et al.*, 2019; *Chen et al.*, 2020). In the fully-observed setting there is a long line of work on estimating these models (*Dutilleul*, 1999; *Werner et al.*, 2008; *Yin and Li*, 2012; *Leng and Tang*, 2012; *Tsiligkaridis et al.*, 2013; *Zhou*, 2014; *Chen and Liu*, 2019). Note that this model also has a history in psychology and medicine, usually within the context of repeated measures studies (*Galecki*, 1994; *Naik and Rao*, 2001). We, however, will focus on applications with a single 2-dimensional data matrix.

Data in these settings, however, are also often collected with missing values (*Little and Rubin*, 2014). Factors as varied as nonresponse, equipment failure or limitations, measurement errors, human mistakes, or data corruption can all result in incomplete data matrices, which most methods are not prepared to handle. Since deleting incomplete cases is inefficient for even small missing rates, researchers generally impute the missing values. In a matrix-variate setting we cannot rely on independent observations, however, so existing methods for imputation are generally not appropriate. *Allen and Tibshirani* (2010) present a method for covariance estimation and imputation in this matrix variate setting. Their full EM-type algorithm is, however, computationally infeasible for even moderate datasets, and the focus of their work is therefore on approximate algorithms designed for imputation. We instead focus on parameter estimation in high-dimensional settings.

In Chapter III we propose methods for estimating both the row- and column-wise precision and covariance matrices in matrix-variate data settings with missing data. In particular, we incorporate missing values with varying sample rates by column. These are based on the graphical Lasso estimator and assume sparsity in the inverse covariance (or precision) matrix and therefore also in the undirected graphs that they encode (in the Gaussian case). We establish the conditions required for consistency and present the convergence rates of our estimators, which attain the same rates as in the fully observed setting when the missing rates are fixed. Proofs for the results in this chapter are deferred to Chapter V.

In *Hornstein et al.* (2019), we developed an extension of the matrix-variate estimators in *Zhou* (2014) to a setting with two-group means in the fully observed case. Motivated by an persistent problem in genomics research, where test statistics for mean differences are often observed to be miscalibrated, likely due to row-wise dependence (*Efron*, 2009), we developed methods for joint mean and covariance estimation in this setting.

We therefore also extend our matrix-variate estimators with missing data to similar settings with two-group mean structure. We show that we can still prove consistency and convergence rates when jointly estimating both the mean and precision matrices, despite significant complications that arise around the interaction of missing values and the joint mean and covariance estimation. We present methodology for extending these estimators to setting with unknown groups labels or more flexible mean structures.

In Chapter IV we apply these methods to a dataset of voting records from the U.S. Senate, where we expect there to be connections both between similar or opposed senators as well as between bills that may share characteristics or topics. Here, by removing the party means for each bill, we focus on connections beyond those of encoded in the mean structure. Instead, we isolate relationships in the covariance structure of the errors, and we use these to conduct an exploratory analysis, finding both expected and interesting novel relationships and patterns.

1.3 Notation

For a matrix $A = (a_{ij}) \in \mathbb{R}^{n \times m}$, we denote the spectral or operator norm as $\|A\|_2$, the Frobenius norm as $\|A\|_F$, the entry-wise max norm as $\|A\|_\infty$, the entry-wise ℓ_1 norm as $|A|_1$, the entry-wise ℓ_0 norm as $|A|_0$, and the matrix one-norm as $\|A\|_1 = \max_j \sum_{i=1}^n |a_{ij}|$. Let $|A|_{1,\text{off}}$ and $|A|_{0,\text{off}}$ denote these norms applied to the nondiagonal entries. Let $|A|$ denote the determinant and $\text{tr}(A)$ denote the trace. Let $\phi_i(A)$ denote the eigenvalues of A, with $\phi_{\max}(A)$ and $\phi_{\min}(A)$ being the largest and smallest eigenvalues, and $\kappa(A)$ being the condition number. Let $\text{diag}(A)$ be the diagonal matrix with the same diagonal as A, and let I be the identity matrix. For $a, b \in \mathbb{R}$ we denote $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. For matrices $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{p \times q}$ we denote the Kronecker product as

$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1m}B \\ \vdots & \ddots & \vdots \\ a_{n1}B & \cdots & a_{nm}B \end{pmatrix} \in R^{np \times mq}$$

For $A, B \in R^{n \times m}$ we denote element-wise (Hadamard) multiplication and division as $A \circ B$ and $A \oslash B$, respectively.

CHAPTER II

Precision Matrix Estimation with Noisy and Missing Data

Undirected graphs are often used to describe high-dimensional distributions. Under sparsity conditions, these graphs can be estimated using penalized methods such as

$$\hat{\Theta} \in \arg\min_{\Theta \succeq 0} \left\{ \operatorname{tr}(\hat{\Gamma}_n \Theta) - \log\det(\Theta) + g_\lambda(\Theta) \right\}, \tag{2.1}$$

where $\hat{\Gamma}_n$ is the sample covariance or correlation matrix and g_λ is a separable (entry-wise) sparsity-inducing penalty function. Although this approach has proven successful in a variety of application areas such as neuroscience and genomics, its soundness hinges on the positive semidefiniteness (PSD) of $\hat{\Gamma}_n$. If $\hat{\Gamma}_n$ is indefinite, the objective may be unbounded from below.

In order to ensure this penalized M-estimator is well-behaving, *Loh and Wainwright* (2015) impose a side constraint of the form $\rho(\Theta) < R$, where ρ is a convex function. Here we focus on the estimator using the operator norm as a side constraint

$$\hat{\Theta} \in \arg\min_{\Theta \succeq 0, \|\Theta\|_2 \leq R} \left\{ \operatorname{tr}(\hat{\Gamma}_n \Theta) - \log\det(\Theta) + g_\lambda(\Theta) \right\}. \tag{2.2}$$

Loh and Wainwright (2017) adopt this method and show in theory the superior statistical properties of this constrained estimator. Their results suggest that the addition of a side constraint is not only sufficient but also almost necessary to effectively untangle the aforementioned complications.

Unfortunately, this additional constraint precludes using existing methods to solve the penalized objective with non-PSD input. To close this gap, we develop an alternating direction method of multipliers (ADMM) algorithm to implement (2.2) efficiently.

The remainder of this chapter is organized as follows. In Section 2.1, we provide an overview of existing related work and describe in detail the optimization issues that arise from indefinite inputs and nonconvex penalties. In Section 2.2, we present the proposed ADMM algorithm and present some convergence results. Section 2.3 provides numerical examples and comparisons. Finally, we summarize the empirical results and their practical implications regarding choice of method in Section 2.4.

2.1 Problem formulation and existing work

There is a wide body of work proposing methods to perform precision matrix estimation in the fully observed case, including *Meinshausen and Bühlmann* (2006), *Yuan and Lin* (2007), *Rothman et al.* (2008), *Friedman et al.* (2008), *Banerjee et al.* (2008), and *Zhou et al.* (2010), most of which are essentially a ℓ_1-penalized likelihood approach (2.1) which we will refer to as the graphical Lasso.

Recent work has focused on using nonconvex regularizers such as SCAD and MCP for model selection in the regression setting (*Fan and Li*, 2001; *Zhang*, 2010; *Breheny and Huang*, 2011; *Zhang and Zhang*, 2012). *Loh and Wainwright* (2015, 2017) extend this analysis to general M-estimators, including variants of the graphical Lasso objective, and show their statistical convergence and support recovery properties. Estimators with these penalties have been shown to attain model selection under weaker theoretical conditions, but require more sophisticated optimization algorithms to solve, such as the local linear approximation (LLA) method of *Fan et al.* (2014).

In a fully observed and noiseless setting, $\hat{\Gamma}_n$ is the sample covariance and guaranteed to be at least positive semidefinite. Then, if g_λ is the ℓ_1-penalty, the objective of (2.1) is convex and bounded from below. In this setting, one can show that for $\lambda > 0$ a unique optimum $\hat{\Theta}$ exists with bounded eigenvalues and that the iterates for any descent algorithm will also have bounded eigenvalues (for example, see Lemma 2 in *Hsieh et al.*, 2014).

When working with missing, corrupted, and dependent data, the likelihood is nonconvex, and the expectation-maximization (EM) algorithm has traditionally been used to perform statistical inference. However, in these noisy settings, the convergence of the EM algorithm is difficult to guarantee and is often slow in practice. For instance, *Städler and Bühlmann* (2012) implement a likelihood-based method for inverse covariance estimation with missing values, but their EM algorithm requires solving a full graphical Lasso optimization problem in each M-step.

An alternative approach is to develop M-estimators that account for missing and corrupted data. For graphical models, *Loh and Wainwright* (2015) establish that the graphical Lasso, including when using nonconvex penalties, can be modified to accommodate noisy or missing data by adjusting the sample covariance estimate.

These modified estimators depend on the observation that statistical theory for the graphical Lasso generally requires that $\|\hat{\Gamma}_n - \Sigma\|_\infty$ converges to zero at a sufficiently fast rate (e.g. *Rothman et al.*, 2008; *Zhou et al.*, 2010; *Loh and Wainwright*, 2017). When considering missing or corrupted data, it is often possible to construct covariance estimates $\hat{\Gamma}_n$ that satisfy this convergence criteria but are not necessarily positive semidefinite. In fact, in high-dimensional settings $\hat{\Gamma}_n$ may even be guaranteed to be indefinite. Attempting to input these indefinite covariance estimates into the graphical Lasso, however, presents novel optimization issues.

Unbounded objective. When attempting to move beyond the ℓ_1 penalized case with positive semidefinite input, the problem in (2.1) becomes unbounded from below, so an optimum may not necessarily exist. This issue comes from two potential sources: 1) negative eigenvalues in $\hat{\Gamma}_n$, or 2) zero eigenvalues combined with the boundedness of the nonconvex penalty g_λ. For example, consider the restriction of the objective in (2.1) to a ray defined by an eigenvalue-vector pair σ_1, v_1 of $\hat{\Gamma}_n$:

$$\begin{aligned} f(I + tv_1v_1^T) &= \operatorname{tr}(\hat{\Gamma}_n) + t\operatorname{tr}(\hat{\Gamma}_n v_1 v_1^T) - \log(1+t) + g_\lambda(tv_1v_1^T) \\ &= \operatorname{tr}(\hat{\Gamma}_n) + t\sigma_1 - \log(1+t) + g_\lambda(tv_1v_1^T). \end{aligned} \tag{2.3}$$

If $\sigma_1 < 0$, we see that f is unbounded from below due to the $t\sigma_1$ and $-\log(1+t)$ terms. In fact, if $\sigma_1 = 0$ and g_λ is bounded from above, as is the case when using standard nonconvex penalties, the objective is also unbounded from below.

So unboundedness can occur anytime there is a negative eigenvalue in the input matrix, or whenever there are zero eigenvalues combined with a nonconvex penalty function g_λ. Unboundedness creates optimization issues, as an optimum no longer necessarily exists.

Handling unboundedness. In order to guarantee that an optimum exists for (2.1), an additional constraint of the form $\rho(\Theta) \leq R$ can be imposed, where ρ is some convex function. In this paper, we consider the estimator (2.2), which uses a side constraint of the form $\|\Theta\|_2 \leq R$. *Loh and Wainwright* (2017) show the rates of convergence of this estimator (2.2) and show that it can attain model selection consistency and spectral norm convergence without the incoherence assumption when used with a nonconvex penalty (see Appendix E therein), but do not discuss implementation or optimization aspects of the problem.

To our knowledge, there is currently no feasible optimization algorithm for the estimator defined in (2.2), particularly when the input is indefinite. *Loh and Wainwright* (2015) present a composite gradient descent method for optimizing a subset of side-constrained versions of (2.1). However, their algorithm requires a side constraint of the specific form $\rho(\Theta) = \frac{1}{\lambda}(g_\lambda(\Theta) + \frac{\mu}{2}\|\Theta\|_F^2)$, which does not include the spectral norm constraint and therefore cannot attain the better theoretical results it achieves. It may be possible to develop heuristic algorithms that alternate performing a proximal gradient update ignoring the side constraint and projecting to the constraint set, but as far as we know there has not been any analysis of algorithms of this type (we discuss this in more detail in Section 2.3.3).

An alternative approach to solving this unbounded issue is to project the input matrix $\hat{\Gamma}_n$ to the positive semidefinite cone before inputting into (2.1). We discuss this further in Section 2.3.1, but this only solves the unbounded issue when using the ℓ_1 penalty; nonconvex penalties still require a side constraint to have a bounded objective and therefore our algorithm is still useful even for the projected methods.

2.1.1 Nonconvex penalties

The nonconvex penalties we will focus on are the SCAD and MCP functions, introduced in *Fan and Li* (2001) and *Zhang* (2010), respectively. Following *Loh and Wainwright* (2015), we make the following assumptions regarding the (univariate) penalty function $g_\lambda \colon \mathbb{R} \to \mathbb{R}$.

(i) $g_\lambda(0) = 0$ and $g_\lambda(t) = g_\lambda(-t)$.

(ii) $g_\lambda(w)$ is nondecreasing for $w >= 0$.

(iii) $g_\lambda(w)/w$ is nonincreasing for $w > 0$.

(iv) $g'_\lambda(w)$ exists for all $w \neq 0$ and $\lim_{w \to 0^+} g'_\lambda(w) = \lambda$.

(v) g_λ is weakly convex, i.e. there exists $\mu > 0$ such that $g_\lambda(w) + (\mu/2)w^2$ is convex.

Note that *Loh and Wainwright* (2017) show stronger model selection results under the following additional assumption.

(vi) There exists a constant $\gamma < \infty$ such that $g'_\lambda(w) = 0$ for all $w > \gamma\lambda$.

This excludes the ℓ_1 penalty, but is satisfied by the nonconvex penalties we consider.

The SCAD penalty takes the form

$$g_\lambda(w) = \begin{cases} \lambda|w| & \text{if } |w| \leq \lambda \\ -\frac{w^2 - 2a\lambda|w| + \lambda^2}{2(a-1)} & \text{if } \lambda < |w| \leq a\lambda \\ \frac{(a+1)\lambda^2}{2} & \text{if } a\lambda < |w| \end{cases} \tag{2.4}$$

for some parameter $a > 2$. Note that this penalty is weakly convex with constant $\mu = 1/(a-1)$.

The MCP penalty has the form

$$g_\lambda(w) = \text{sign}(w)\lambda \int_0^{|w|} \left(1 - \frac{z}{\lambda a}\right)_+ dz \tag{2.5}$$

for some parameter $a > 0$. This penalty is weakly convex with $\mu = 1/a$.

2.2 ADMM Algorithm

Our algorithm is similar to the algorithm in *Guo and Zhang* (2017), which applies ADMM to the closely related problem of condition number-constrained sparse precision matrix estimation using the same splitting scheme as below. We discuss their method in more detail in Section A.1. The following algorithm is specialized to the case where the spectral norm is used as the side constraint. In Section A.2 we derive a similar ADMM algorithm that can be used for any side constraint with a computable projection operator.

Rewrite the objective from (2.2) as

$$f(\Theta) = \text{tr}(\hat{\Gamma}_n\Theta) - \log\det(\Theta) + g_\lambda(\Theta) + \mathbb{1}_{\mathcal{X}_R}(\Theta) \tag{2.6}$$

where $\mathcal{X}_R = \{\Theta : \Theta \succeq 0, \|\Theta\|_2 \leq R\}$ and $\mathbb{1}_{\mathcal{X}}(\Theta) = 0$ if $\Theta \in \mathcal{X}$ and ∞ otherwise.

Let $\rho > 0$ be a penalty parameter and let $\text{Prox}_{g_\lambda/\rho}$ be the prox operator of g_λ/ρ. We derive these updates for SCAD and MCP in Section 2.2.1. Let $T_\rho(A)$ be the following prox operator for $-\log\det\Theta + \mathbb{1}_{\mathcal{X}_R}(\Theta)$, which we derive in Section 2.2.1,

$$\begin{aligned} T_\rho(A) &= T_\rho(UMU^T) = U\tilde{D}U^T \\ &\quad \text{where } \tilde{D}_{ii} = \min\left\{\frac{M_{ii} + (M_{ii}^2 + \frac{4}{\rho})^{1/2}}{2}, R\right\}, \end{aligned}$$

Algorithm 1: ADMM for graphical Lasso with a side constraint

```
**Input:** $\hat{\Gamma}_n$, $\rho$, $g_\lambda$, $R$
**Output:** $\hat{\Theta}$
Initialize $V^0 = \Theta^0 \succ 0$, $\Lambda^0 = \mathbf{0}$ ;
**while** *not converged* **do**
    $V^{k+1} = \text{Prox}_{g_\lambda/\rho}\left(\frac{\rho\Theta^k + \Lambda^k}{\rho}\right)$
    $\Theta^{k+1} = T_\rho\left(\frac{\rho V^{k+1} - \hat{\Gamma}_n - \Lambda^k}{\rho}\right)$
    $\Lambda^{k+1} = \Lambda^k + \rho(\Theta^{k+1} - V^{k+1})$
**end**
```

where UMU^T is the eigendecomposition of A. Then the ADMM algorithm for solving (2.6), which we derive in Section 2.2.1, is described in Algorithm 1. Computationally this algorithm is dominated by the eigendecomposition used to evaluate T_ρ, and therefore has a complexity of $O(m^3)$, which matches the scaling of other graphical Lasso solvers (e.g. *Meinshausen and Bühlmann*, 2006; *Friedman et al.*, 2008; *Hsieh et al.*, 2014).

2.2.1 Derivation of Algorithm 1

ADMM algorithm. Recall that we can rewrite the objective as

$$f(\Theta) = \text{tr}(\hat{\Gamma}_n\Theta) - \log\det(\Theta) + g_\lambda(\Theta) + \mathbb{1}_{\mathcal{X}_R}(\Theta)$$

where $\mathcal{X}_R = \{\Theta : \Theta \succeq 0, \|\Theta\|_2 \leq R\}$ and $\mathbb{1}_{\mathcal{X}}(\Theta) = 0$ if $\Theta \in \mathcal{X}$ and ∞ otherwise.

We then introduce an auxiliary optimization variable $V \in \mathbb{R}^{m\times m}$ and reformulate the problem as

$$\hat{\Theta} = \underset{\Theta, V \in \mathbb{R}^{m\times m}}{\arg\max} \left\{ \text{tr}(\hat{\Gamma}_n\Theta) - \log\det(\Theta) + \mathbb{1}_{\mathcal{X}_\mathcal{R}}(\Theta) + g_\lambda(V) \right\} \text{ s.t. } \Theta = V$$

For a penalty parameter $\rho > 0$ and Lagrange multiplier $\Lambda \in \mathbb{R}^{m\times m}$, we consider the augmented Lagrangian

$$\mathcal{L}_\rho(\Theta, V, \Lambda) = \operatorname{tr}(\hat{\Gamma}_n\Theta) - \log\det(\Theta) + \mathbb{1}_{\mathcal{X}_R}(\Theta) + g_\lambda(V) + \frac{\rho}{2}\|\Theta - V\|_F^2 + \langle \Lambda, \Theta - V\rangle \tag{2.7}$$

The ADMM algorithm is then, given current iterates Θ^k, V^k, and Λ^k,

$$V^{k+1} = \operatorname*{arg\,min}_{V\in\mathbb{R}^{m\times m}} \left\{ g_\lambda(V) + \frac{\rho}{2}\|\Theta^k - V\|_F^2 + \langle \Lambda^k, \Theta^k - V\rangle \right\} \tag{2.8}$$

$$\begin{aligned}\Theta^{k+1} = \operatorname*{arg\,min}_{\Theta\in\mathbb{R}^{m\times m}} \Big\{ & -\log\det\Theta + \operatorname{tr}(\hat{\Gamma}_n\Theta) + \mathbb{1}_{\mathcal{X}_R}(\Theta) \\ & + \frac{\rho}{2}\|\Theta - V^{k+1}\|_F^2 + \langle \Lambda^k, \Theta - V^{k+1}\rangle \Big\}\end{aligned} \tag{2.9}$$

$$\Lambda^{k+1} = \Lambda^k + \rho(\Theta^{k+1} - V^{k+1}) \tag{2.10}$$

Considering the V-subproblem, we can show that the minimization problem in (2.8) is equivalent to

$$V^{k+1} = \operatorname*{arg\,min}_{V\in\mathbb{R}^{m\times m}} \left\{ \frac{1}{\rho} g_\lambda(V) + \frac{1}{2}\left\| V - \frac{\rho\Theta^k + \Lambda^k}{\rho} \right\|_F^2 \right\}.$$

Which is a prox operator of g_λ/ρ. Let $W = \frac{\rho\Theta^k + \Lambda^k}{\rho}$ and $\nu = 1/\rho$. If g_λ is the ℓ_1 penalty then these updates simply soft-threshold the elements of W at level λ/ρ. For SCAD, these updates have the element-wise form

$$\operatorname{Prox}_{g_\lambda/\rho}(w) = \begin{cases} 0 & \text{if } |w| \leq \nu\lambda \\ w - \operatorname{sign}(w)\nu\lambda & \text{if } \nu\lambda \leq |w| \leq (\nu+1)\lambda \\ \frac{w - \operatorname{sign}(w)\frac{a\nu\lambda}{a-1}}{1 - \frac{\nu}{a-1}} & \text{if } (\nu+1)\lambda \leq |w| \leq a\lambda \\ w & \text{if } a\lambda \leq |w| \end{cases} \tag{2.11}$$

While for MCP the updates are

$$\mathrm{Prox}_{g_\lambda/\rho}(w) = \begin{cases} 0 & \text{if } |w| \le \nu\lambda \\ \frac{w-\mathrm{sign}(w)\nu\lambda}{1-\nu/a} & \text{if } \nu\lambda \le |w| \le a\lambda \\ w & \text{if } a\lambda \le |w| \end{cases} \tag{2.12}$$

See *Loh and Wainwright* (2015) for the derivations of these updates.

For the Θ-subproblem, we can similarly show that (2.9) is equivalent to

$$\Theta^{k+1} = \underset{\Theta \in \mathbb{R}^{m\times m}}{\arg\min} \left\{ -\log\det\Theta + \mathbb{1}_{\mathcal{X}_R}(\Theta) + \frac{\rho}{2}\left\| \Theta - \frac{\rho V^{k+1} - \hat{\Gamma}_n - \Lambda^k}{\rho} \right\|_F^2 \right\} \tag{2.13}$$

For any matrix A with corresponding eigendecomposition $A = RMR^T$ let us define the operator

$$\begin{aligned} T_\rho(A) = T_\rho(UMU^T) &= \arg\min_\Theta \left\{ -\log\det\Theta + \mathbb{1}_{\mathcal{X}_R}(\Theta) + \frac{\rho}{2}\|\Theta - A\|_F^2 \right\} \\ &= U\tilde{D}U^T \text{ where } \tilde{D}_{ii} = \min\left\{ \frac{M_{ii} + (M_{ii}^2 + \frac{4}{\rho})^{1/2}}{2}, R \right\} \end{aligned} \tag{2.14}$$

whose solution is derived below. Then the solution to (2.9) is $T_\rho((\rho V^{k+1} - \hat{\Gamma}_n - \Lambda^k)/\rho)$.

Using these results, the algorithm in (2.8)-(2.10) becomes

$$\begin{aligned} V^{k+1} &= \mathrm{Prox}_{g_\lambda/\rho}\left(\frac{\rho\Theta^k + \Lambda^k}{\rho} \right) \\ \Theta^{k+1} &= T_\rho\left(\frac{\rho V^{k+1} - \hat{\Gamma}_n - \Lambda^k}{\rho} \right) \\ \Lambda^{k+1} &= \Lambda^k + \rho(\Theta^{k+1} - V^{k+1}) \end{aligned} \tag{2.15}$$

Solution of T_ρ Recall that in (2.14) we define

$$T_\rho(A) = \arg\min_\Theta \left\{ -\log\det\Theta + \mathbb{1}_{\mathcal{X}_R}(\Theta) + \frac{\rho}{2}\|\Theta - A\|_F^2 \right\}$$

Let $\Theta = WDW^T$ and $A = UMU^T$ be the eigen-decompositions of the optimization variable and A. Then, similar to the derivation in *Guo and Zhang* (2017), we can rewrite this problem as

$$\begin{aligned} T_\rho(A) &= \underset{\Theta \in \mathbb{R}^{m \times m}}{\arg\min} - \log\det\Theta + \frac{\rho}{2}\operatorname{tr}(\Theta\Theta) - \rho\operatorname{tr}(\Theta A) + \mathbb{1}_{\mathcal{X}_R}(\Theta) \\ &= \underset{\Theta = WDW^T}{\arg\min} \; - \log\det D + \frac{\rho}{2}\operatorname{tr}(DD) - \rho\operatorname{tr}(WDW^TUMU^T) + \mathbb{1}_{\mathcal{X}_R}(D) \\ &= \underset{\Theta = WDW^T, W=U}{\arg\min} \; - \log\det D + \frac{\rho}{2}\operatorname{tr}(DD) - \rho\operatorname{tr}(DM) + \mathbb{1}_{\mathcal{X}_R}(D) \end{aligned}$$

The final line is since, if we denote $O(m)$ to be the set of $m \times m$ orthonormal matrices,

$$\operatorname{tr}(WDW^TUMU^T) = \operatorname{tr}((U^TW)D(U^TW)^TM) \le \sup_{Q \in O(m)} \operatorname{tr}(QDQ^TM) = \operatorname{tr}(DM)$$

Which holds with equality when $W = U$. Note that the last equality here is from Theorem 14.3.2 of *Farrell* (1985).

We therefore get that $T_\rho(A) = U\tilde{D}U^T$ where

$$\begin{aligned} \tilde{D} &= \underset{D \text{ diagonal}}{\arg\min} \; - \log\det D + \frac{\rho}{2}\operatorname{tr}(D^2) - \rho\operatorname{tr}(DM) + \mathbb{1}_{\mathcal{X}_R}(D) \\ &= \underset{D \text{ diagonal}}{\arg\min} \sum_{i=1}^{m} \Big(- \log D_{ii} + \frac{\rho}{2}D_{ii}^2 - \rho D_{ii}M_{ii} + \mathbb{1}(0 \le D_{ii} \le R) \Big) \end{aligned}$$

We can see that this is separable by element. Let

$$q(d; M_{ii}) = -\log d + \frac{\rho}{2}d^2 - \rho d M_{ii}$$

So $\tilde{D}_{ii} = \arg\min_d q(d; M_{ii}) + \mathbb{1}(0 \le d \le R)$. Ignoring the constraints in the indicator function for now, we can set the derivative of q equal to zero to get that

$$0 = -\frac{1}{d} + \rho d - \rho M_{ii} \implies 0 = d^2 - M_{ii}d - \frac{1}{\rho}$$

Which we can solve with the quadratic formula to show that $q(d; M_{ii})$ has a unique minimizer over $d > 0$ at

$$\arg\min_{d} q(d; M_{ii}) = \frac{M_{ii} + (M_{ii}^2 + \frac{4}{\rho})^{1/2}}{2}$$

Adding $\mathbb{1}(0 \leq d \leq R)$ back and noting that $q(d; M_{ii})$ is strictly convex over $d > 0$, we get that we simply need to truncate this value at R. Therefore we get that

$$T_\rho(UMU^T) = U\tilde{D}U^T \text{ where } \tilde{D}_{ii} = \min\left\{\frac{M_{ii} + (M_{ii}^2 + \frac{4}{\rho})^{1/2}}{2}, R\right\}$$

2.2.2 Convergence

The following proposition applies standard results on the convergence of ADMM for convex problems to show convergence when the ℓ_1 penalty is used. Details are in Section A.1.

Proposition 1. *If the penalty is convex and satisfies the conditions in Section 2.1.1, Algorithm 1 converges to a global minimum of* (2.6).

Remark. Regarding the nonconvex penalty, recent work has established ADMM convergence results in some nonconvex settings (see *Hong et al.*, 2016; *Wang et al.*, 2015), but to our knowledge there is no convergence result that encompasses this nonsmooth and nonconvex application. We can show convergence if a fairly strong assumption is made on the iterates, but we are currently working on extending existing results to this case.

Proposition 2 shows that any limiting point of Algorithm 1 is a stationary point of the original objective (2.6). This is proved in Section A.1. When using the ℓ_1 penalty or a nonconvex penalty with $R \leq \sqrt{2/\mu}$, where μ is the weak convexity constant of g_λ, the objective f is convex and therefore any stationary point is unique and also the global optimum. See Section A.3 for a more detailed discussion.

Proposition 2. *Assume that the penalty g_λ satisfies the conditions in Section 2.1.1. Then for any limit point $(\Theta^*, V^*, \Lambda^*)$ of the ADMM algorithm defined in Algorithm 1, Θ^* is also a stationary point of the objective f as defined in* (2.6).

The assumptions on g_λ in Section 2.1.1 are the same as those assumed in *Loh and Wainwright* (2015, 2017), and are satisfied by the Lasso, SCAD, and MCP functions.

Note that if a limiting point is found to exist when using a nonconvex penalty the result in Proposition 2 will still hold. Empirically we find that the algorithm performs well and converges consistently when used with nonconvex penalties, but there is no existing theoretical guarantee that a limiting point of ADMM will exist in that setting.

2.3 Simulations

We evaluate the proposed estimators using the relative Frobenius norm and the sum of the false positive rate and false negative rate (FPR + FNR). We present results over a range of λ values, noting that all the compared methods would use similar techniques to perform model tuning. We also present an example of how to use BIC or cross-validation to tune these methods. We present results using covariance matrices from auto-regressive and Erdős-Rényi random graph models. See Section A.3 for descriptions of these models as well as additional simulation results.

2.3.1 Alternative methods

When faced with indefinite input, there are two alternative graphical Lasso-style estimators that can be used besides (2.2), which involve either ℓ_∞ projection to the positive semidefinite cone or nodewise regression in the style of *Meinshausen and Bühlmann* (2006).

Projection. Given an indefinite input matrix $\hat{\Gamma}_n$, *Park* (2016) and *Greenewald et al.* (2017) propose performing the projection $\hat{\Gamma}_n^+ = \arg\min_{\Gamma \succeq 0} \|\Gamma - \hat{\Gamma}_n\|_\infty$. They then input $\hat{\Gamma}_n^+$ into the optimization problem (2.1). This is similar to the projection done in *Datta and Zou* (2017). In terms of the upper bound on statistical convergence rates, this method pays a constant factor cost, though in practice projection may result in a loss of information and therefore a decrease in efficiency.

After projecting the input, existing algorithms can be used to optimize (2.1) with the ℓ_1

penalty. However, as mentioned in Section 2.1, using a nonconvex penalty still leads to an unbounded objective and therefore still requires using our ADMM algorithm to solve (2.2).

Nodewise regression. *Loh and Wainwright* (2012) and *Rudelson and Zhou* (2017) both study the statistical and computational convergence properties of using errors-in-variables regression to handle indefinite input matrices in high-dimensional settings. Following the nodewise regression ideas of *Meinshausen and Bühlmann* (2006) and *Yuan* (2010), we can perform m Lasso-type regressions to obtain estimates $\hat{\beta}_j$ and form estimates $\hat{a}_j$, where

$$\begin{aligned} \hat{\beta}_j &\in \operatorname*{arg\,min}_{\|\beta\|_1 \le R} \left\{ \frac{1}{2}\beta^T \hat{\Gamma}_{n,-j,-j}\beta - \langle \hat{\Gamma}_{n,-j,j}, \beta \rangle + \lambda \|\beta\|_1 \right\} \\ \hat{a}_j &= -(\hat{\Gamma}_{n,j,j} - \langle \hat{\Gamma}_{n,-j,j}, \hat{\beta}_j \rangle)^{-1} \end{aligned} \tag{2.16}$$

and combine to get $\tilde{\Theta}$ with $\tilde{\Theta}_{-j,j} = \hat{a}_j \hat{\beta}_j$ and $\tilde{\Theta}_{j,j} = -\hat{a}_j$. Finally, we symmetrize the result to obtain $\hat{\Theta} = \arg\min_{\Theta \in S^m} \|\Theta - \tilde{\Theta}\|_1$, where S^m is the set of symmetric matrices.

These types of nodewise estimators have gained popularity as they require less restrictive incoherence conditions to attain model selection consistency and often perform better in practice in the fully observed case. They have not, however, been as well studied when used with indefinite input.

2.3.2 Data models

We test these methods on two models that result in indefinite covariance estimators, the non-separable Kronecker sum model from *Rudelson and Zhou* (2017) and the missing data graphical model described in *Loh and Wainwright* (2015).

Missing data (MD). As discussed above, *Loh and Wainwright* (2013, 2015) propose an estimator for a graphical model with missing-completely-at-random observations.

Let $W \in \mathbb{R}^{n \times m}$ be a mean-zero subgaussian random matrix. Let $U \in \{0,1\}^{n \times m}$ where $U_{ij} \sim \text{Bernoulli}(\zeta_j)$ are independent of W. This corresponds to entries of the jth column of the data matrix being observed with probability ζ_j. Then we have an unobserved matrix

Z and observed matrix X generated by $Z = WA^{1/2}$ and $X = U \circ X$, where $\circ$ denotes the Hadamard, or element-wise, product. Here the covariance estimate for A is

$$\hat{\Gamma}_n = \frac{1}{n} X^T X \oslash M \text{ where } M_{k\ell} = \begin{cases} \zeta_k & \text{if } k = \ell \\ \zeta_k \zeta_\ell & \text{if } k \neq \ell \end{cases} \tag{2.17}$$

where $\oslash$ denotes element-wise division. As we divide off-diagonal entries by smaller values, $\hat{\Gamma}_n$ will not necessarily be positive semidefinite.

Kronecker Sum (KS). *Park et al.* (2017) present a graphical model with additive noise that is dependent across observations. This noise structure was first studied in the regression setting in *Rudelson and Zhou* (2017) with a Kronecker sum covariance.

$X \sim \mathcal{M}_{n,m}(0, A \oplus B)$, where $\mathcal{M}_{n,m}$ is the matrix variate normal distribution and for covariance matrices $A \in \mathbb{R}^{m \times m}$ and $B \in \mathbb{R}^{n \times n}$. Note that $A \oplus B = A \otimes I_n + I_m \otimes B$, where $\otimes$ denotes the Kronecker product. We are interested in estimating the signal precision matrix $\Theta = A^{-1}$, which has sparse off-diagonal entries. For our simulations, we normalize the noise covariance B so that $\mathrm{tr}(B) = n\tau_B$, where τ_B is a measure of the noise level. Then the initial covariance estimate for A is given by

$$\hat{\Gamma}_n = \frac{1}{n} X^T X - \frac{\hat{\mathrm{tr}}(B)}{n} I_m \tag{2.18}$$

as shown in *Rudelson and Zhou* (2017). Note that, in this model, $\hat{\Gamma}_n$ is guaranteed to not be positive semidefinite when $m > n$, as $X^T X$ will have zero eigenvalues.

Covariance models. Let $\Omega = A^{-1} = (\omega_{ij})$. We consider simulation settings using the following covariance models for A, which are also used in *Zhou* (2014).

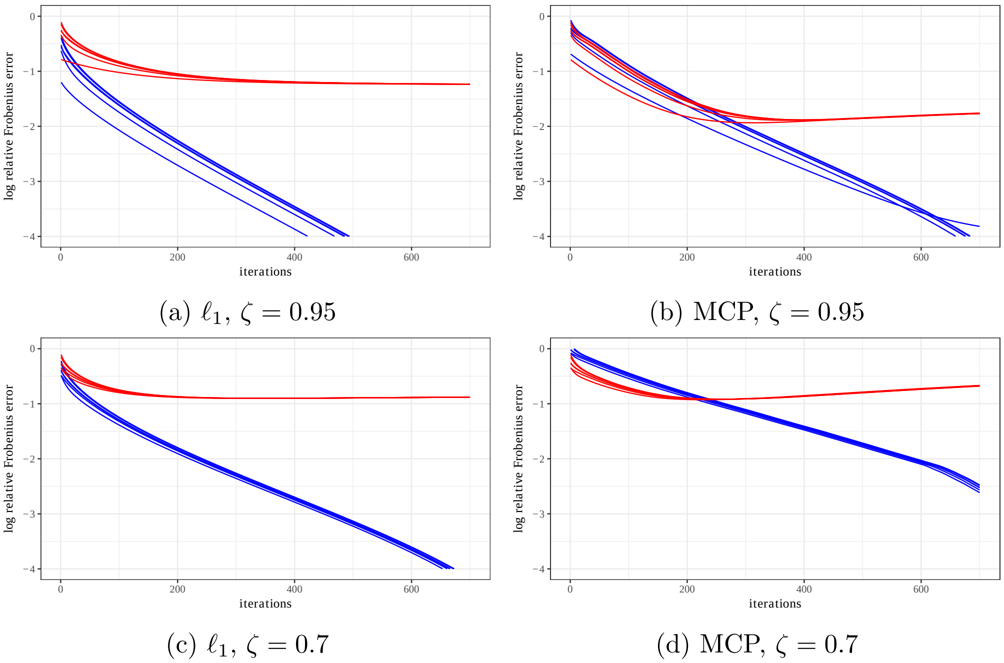
- **AR1(r)**: The covariance matrix is of the form $A = (r^{|i-j|})_{ij}$.
- **Star-Block (SB)**: Here the covariance matrix is block-diagonal, where each block's precision matrix corresponds to a star-structured graph with $A_{ii} = 1$. For the corre-

sponding edge set E, then $A_{ij} = r$ if $(i,j) \in E$ and $A_{ij} = r^2$ otherwise.

- **Erdos-Renyi random graph (ER)**: We initialize $\Omega = 0.25I$ then randomly select d edges. For each selected edge (i,j), we randomly choose $w \in [0.6, 0.8]$ and update $\omega_{ij} = \omega_{ji} \to \omega_{ij} - w$ and $\omega_{ii} \to \omega_{ii} + w$, $\omega_{jj} \to \omega_{jj} + w$.

2.3.3 Simulation results

Optimization performance. Figure 2.1 shows the optimization performance of Algorithm 1 using nonprojected input matrices from the missing data model with both ℓ_1 and nonconvex penalties (MCP). The top two panels present an "easy" scenario with a higher sampling rate, while the bottom two have a more challenging scenario with significant missing data. Blue lines report the optimization error while red lines are the statistical error.



(a) ℓ_1, $\zeta = 0.95$

(b) MCP, $\zeta = 0.95$

(c) ℓ_1, $\zeta = 0.7$

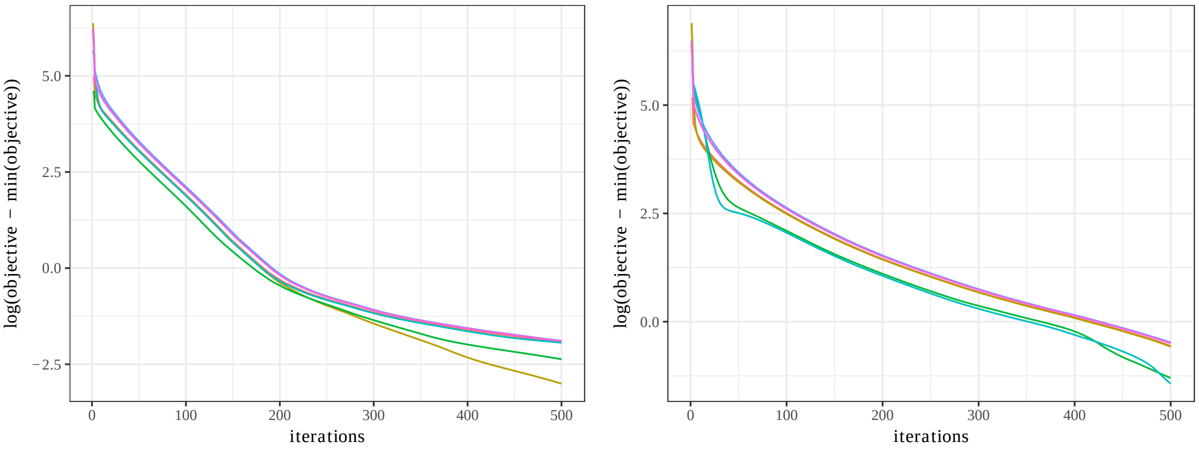
(d) MCP, $\zeta = 0.7$

Figure 2.1: Convergence of the ADMM algorithm for several initializations. Blue lines show the relative optimization error ($\|\Theta^k - \hat{\Theta}\|_F / \|\Theta^*\|_F$, where $\hat{\Theta}$ is the result of running our algorithm to convergence) while red lines show the statistical error ($\|\Theta^k - \Theta^*\|_F / \|\Theta^*\|_F$). All panels use an AR1(0.7) covariance with $m = 300$ and $n = 125$ and set $\rho = 12$. The left panels use an ℓ_1 penalty, while the right panels use MCP with $a = 2.5$. R is set to be three times the oracle spectral norm.

All the plots in Figure 2.1 have their optimization error quickly converge to below the statistical error. These plots also suggest that our algorithm can attain linear convergence rates. We find that the algorithm consistently converges well over a range of tested scenarios.

Comparing the statistical error of the top two plots, we see that MCP achieves significantly lower error for the easier scenario. But in the bottom two plots, where there is more missing data, it struggles relative to the ℓ_1 penalty. This is a common trend through our simulations, as the performance of estimators using MCP degrades as missingness increases while the ℓ_1-penalized versions are more robust.

Figure 2.2 shows the convergence behavior for several initializations in terms of objective value. Our algorithm seems to attain a linear convergence rate in terms of the objective values even with a nonconvex penalty regardless of the initialization. We find that the algorithm consistently converges well over a range of tested scenarios.



(a) KS, AR

(b) MD, ER

Figure 2.2: Convergence behavior of the ADMM algorithm for two objectives. Panel a shows the optimization convergence under the Kronecker sum model with $A = \text{AR1}(0.6)$, $B = \text{ER}$, $m = 300$, $n = 140$, $\tau_B = 0.3$, and $\lambda = 0.2$, while Panel b is for the missing data model with $A = ER$, $m = 400$, $n = 140$, $\zeta = 0.7$, and $\lambda = 0.2$. We choose $\rho = 12$ and the SCAD penalty is used with $a = 2.1$.

Comparison to gradient descent. Figure 2.3 compares the optimization performance of our ADMM algorithm to gradient descent. Note that since proximal gradient descent is difficult to do in this setting, requiring an interior optimization step, we use a heuristic version similar to that suggested by *Agarwal et al.* (2012) that does the proximal gradient step ignoring the side-constraint then projects back to the side-constraint space. Note that since ρ in ADMM is roughly equivalent to the inverse step size in gradient descent, we compare for difference values of ρ. These methods also take roughly the same computational time per iteration, as they are both dominated by either an eigenvalue decomposition or matrix inversion.

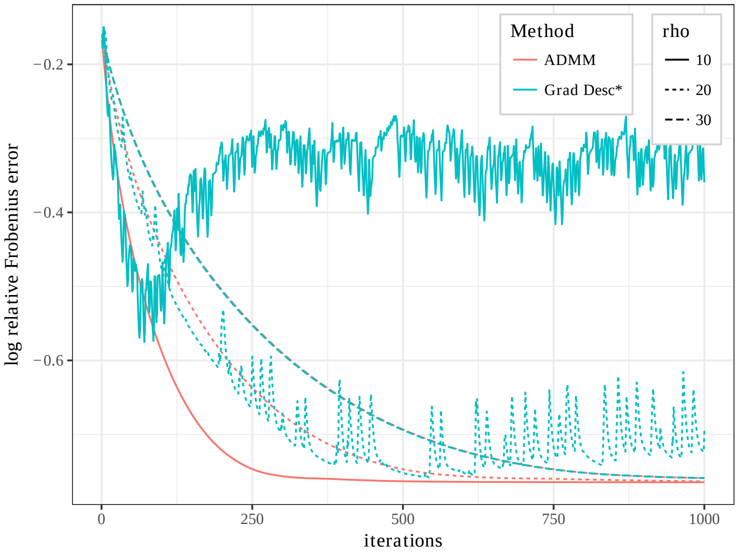


Figure 2.3: Comparing the convergence behavior of ADMM to gradient descent. Here we use an AR1(0.8) model with $m = 200$, $n = 150$, $\zeta = 0.6$, and use an ℓ_1 penalty with $\lambda = 0.11$. For gradient descent, ρ is the inverse of the step size. Note that since proximal gradient descent is difficult to do in this problem, this version performs the proximal gradient step without the side-constraint then projects back to the space.

We can see that for large enough values of ρ, these methods are nearly identical. Although there is no known theoretical guarantee of convergence, it seems that this heuristic gradient descent still convergence well for small enough step sizes.

But for smaller values, i.e. larger step sizes, ADMM still performs well and obtains faster convergence rates while gradient descent is unstable and inconsistent. This combined with the convergence guarantee of ADMM leads us to recommend this algorithm.

Method comparisons. Figure 2.4 demonstrates the statistical performance along the full regularization path. Across the panels from left to right, the sampling rate decreases and therefore the magnitude of the most negative eigenvalue increases.

In terms of Frobenius error, both projected methods and the nonprojected estimator with

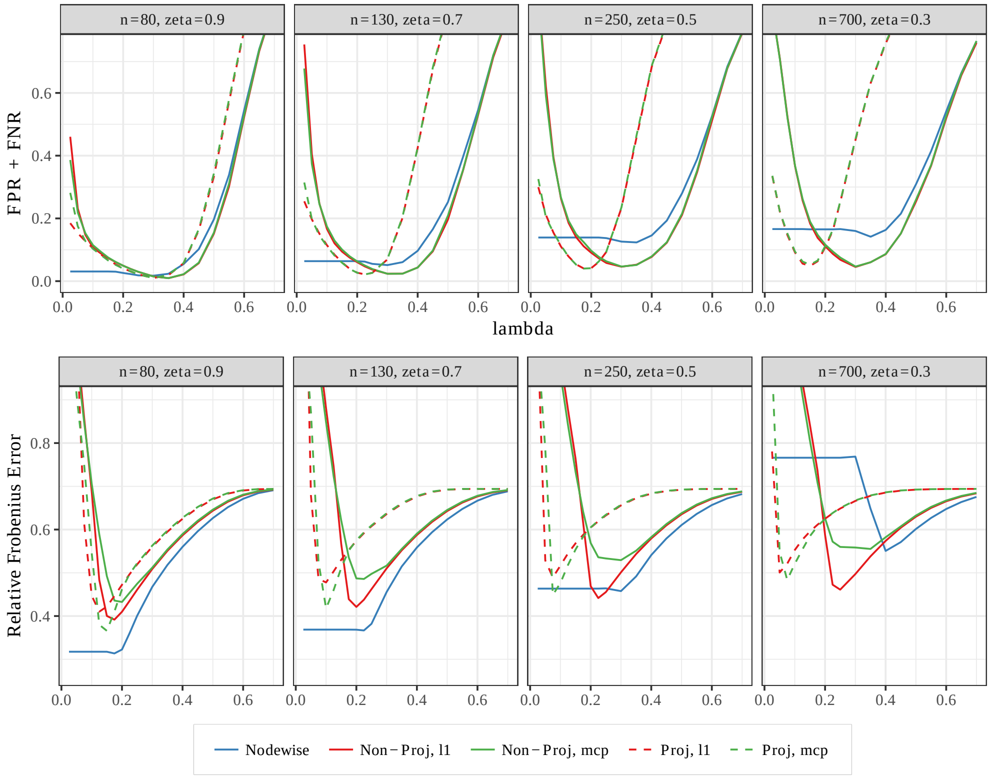


Figure 2.4: The performance of the various estimators for the missing data model in terms of relative Frobenius error ($\|\hat{\Theta} - \Theta^*\|_F / \|\Theta^*\|_F$) and model selection as measured by FPR + FNR. We use an AR(0.6) covariance and set $m = 1200$. Settings are chosen so that the effective sample size ($n\zeta^2$) is roughly equivalent. The MCP penalty uses $a = 2.5$. We set R to be 1.5 times the oracle value for each method and set $\rho = 24$. Our convergence criteria is $\|\Theta^{k+1} - \Theta^k\|_F / \|\Theta^k\|_F < 5\text{e}{-5}$.

the ℓ_1 penalty get slightly worse across panels, but the nodewise regression and the nonprojected MCP estimator react much more negatively to more indefinite input. The nodewise regression in particular goes from being among the best to among the worst estimators as the sampling rate decreases.

Comparing the projected and nonprojected curves in Figure 2.4, we see that the optimal value of λ, as well as the range of optimal values, shrinks for the projected method as the sampling rate decreases. This pattern is consistently repeated across models and scenarios, likely because the ℓ_∞ projection is shrinking the off-diagonal entries of the input matrix. We find that the nonprojected graphical Lasso performs slightly better than the projected version when used with the ℓ_1 penalty, likely due to the information lost in this shrinkage.

Figure 2.4 also shows how these methods perform in terms of model selection. We can see that the nonconvex penalties perform essentially identically to their ℓ_1 penalized counterparts. In particular, the degradation of the nonprojected MCP estimator in terms of norm error does not seem to affect its model selection performance. The nodewise regression, however, still demonstrates this pattern, as its model selection performance degrades across the panels. For scenarios with more missing data, the nonprojected estimators seem to be easier to tune, maintaining a wider range of λ values where they perform near-optimally. In Section A.3 of the supplement we perform similar experiments in a variety of different noise and model settings.

Sensitivity to *R*. Figure 2.5 demonstrates the sensitivity of the nonprojected estimators to the choice of R, the size of the side constraint. We can see that all these methods are sensitive to the choice of R for small values of λ in terms of norm error. None of the methods are sensitive in terms of model selection.

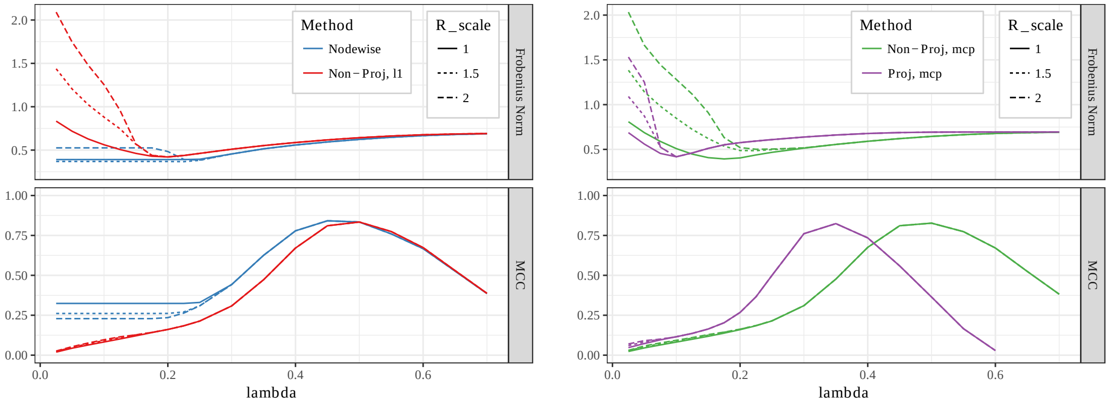


Figure 2.5: The performance of missing data estimators over different choices of R. The non-nodewise estimators set $R = \text{R\_scale} \times \|A\|_2$, while each node's regression in the nodewise estimator sets R to be R_scale times that node's oracle ℓ_1 value. We use an AR(0.6) covariance, set $m = 1200$, $n = 130$, and choose a sampling rate of $\zeta = 0.7$. The MCP penalty is chosen with $a = 2.5$.

The nonprojected graphical Lasso with MCP is the most sensitive to R and is also sensitive for larger choices of λ, which is important since it never reaches its oracle minimum norm errors when R is chosen to be larger than the oracle. The nonprojected graphical Lasso with ℓ_1 and the projected graphical Lasso with MCP both still achieve the same best-case performance when R is misspecified, though tuning λ becomes more difficult.

The nodewise regression results are also plotted here. Here R is the ℓ_1 side constraint level in (2.16). For smaller values of λ the nodewise estimator levels off, corresponding to when the side constraint becomes active over the penalty. Different values of R change when this occurs and, if R is chosen large enough, do not significantly affect ideal performance. Note that these use a stronger oracle that knows each column-wise ℓ_1 norm, but do show that this method can be improved with careful tuning.

Tuning parameter selection. Note that in practice tuning parameters must be selected for all these methods. In particular, we must tune λ and possibly the side-constraint R. Note that one often has a reasonable prior for the magnitude of the spectral norm of the true precision matrix, so if that is the case a multiple of that can often be used to choose R. Also, as noted in Section 2.3.3, when using the ℓ_1 penalty the choice of R primarily affects how difficult tuning λ will be. Though it is important to tune correctly when using nonconvex penalties, we do not recommend those methods when there is significant missing data. Therefore we will focus on tuning λ here, though the same methods can be used to choose R as well.

Two possible methods are to use cross-validation or a modified BIC criterion. Note that the particular implementation of both of these will depend on the data model that is being used, as these methods can be applied to any method that generates an indefinite initial estimate of the covariance, but we will show an example using the simple missing data case.

For the missing data case we can follow *Städler and Bühlmann* (2012), which uses the same data model. Recall the notation in Section 2.3.2, where X_{ij} denotes the ith value of variable j and U_{ij} tracks if that value is observed. Here, we define the observed log-likelihood of an observation X_i given a precision matrix estimate $\hat{\Theta}$ as

$$\ell(X_i, U_i; \hat{\Sigma}) = \log \phi(X_{i,U_i}; \hat{\Sigma}_{U_i,U_i})$$

where X_{i,U_i} is the vector of values that are observed for observation i, $\hat{\Sigma} = \hat{\Theta}^{-1}$, and ϕ is the multivariate normal density. The BIC criterion, which we minimize, is therefore

$$\text{BIC}(\lambda) = -2\sum_{i} \ell(X_i, U_i; \hat{\Sigma}) + \log(n)\sum_{j \leq j'} \mathbb{1}\{\hat{\Theta}_{jj'\}\neq 0}$$

To cross-validate, we can divide the data into V folds, where the vth fold contains indices N_v. The cross-validation score, which we maximize, is therefore

$$\text{CV}(\lambda) = \sum_{v}\sum_{i \in N_v} \ell(X_i, U_i; \hat{\Sigma}_{-v})$$

where $\hat{\Sigma}_{-v} = \hat{\Theta}_{-v}^{-1}$ and $\hat{\Theta}_{-v}$ is the estimate based on the sample omitting the observations in N_v.

Figure 2.6 presents an example of parameter tuning on a simulated scenario. We see that both BIC and CV select slightly higher-than-optimal levels of penalization in terms of model selection, but that selected model still achieves fairly good model selection.

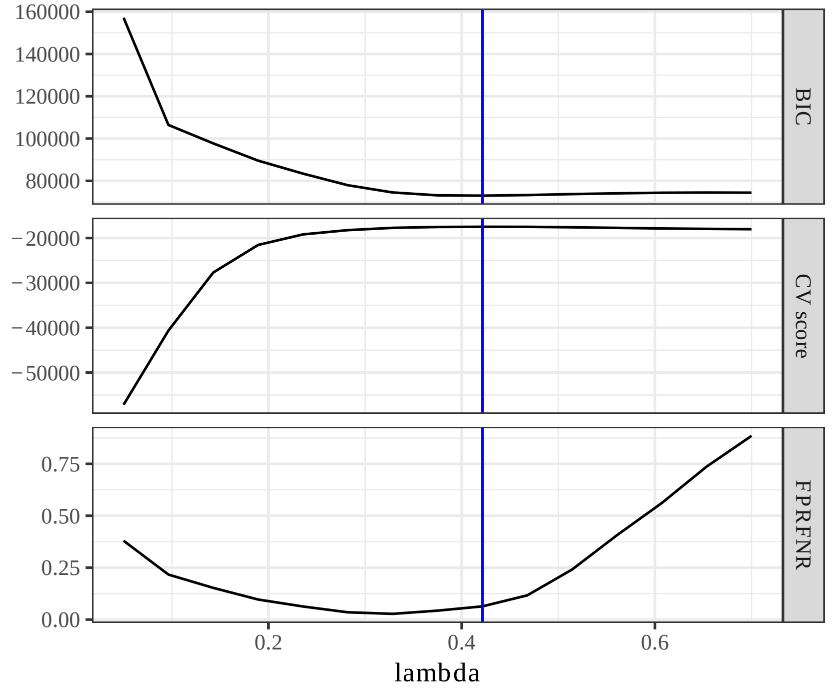


Figure 2.6: Example parameter tuning using BIC and CV. We additionally present the FPR+FNR rate of the estimate. The vertical lines show the optimal λ values for BIC and CV, which here happen to be identical. We set $m = 400$ and $n = 80$, the sampling rate to $\zeta = 0.8$, and let A be from an AR(0.6) model.

2.4 Conclusion

We study the estimation of sparse precision matrices from noisy and missing data. To close an existing algorithmic gap, we propose an ADMM algorithm that allows for fast optimization of the side-constrained graphical Lasso, which is needed to implement the graphical Lasso with either indefinite input and/or nonconvex penalties. We investigate its convergence properties and compare its performance with other methods that handle the indefinite sample covariance matrices that arise with dirty data.

We find that methods with nonconvex penalties are quite sensitive to the indefiniteness of the input covariance estimate, and are particularly sensitive to the magnitude of its negative eigenvalues. They may have better existing theoretical guarantees, but in practice we find that with nontrivial missingness or noise they perform worst than or, at best, recover the performance of their ℓ_1-normalized counterparts. The nonconvex methods can outperform the ℓ_1-penalized ones when there is a small amount of missingness or noise, but in these cases we often find the nodewise estimator to perform best.

In difficult settings with significant noise or missingness, the most robust and efficient method seems to be using the graphical Lasso with nonprojected input and an ℓ_1 penalty. As the application becomes easier – with more observations or less missing data – the nodewise estimator becomes more competitive, just as it is understood to be with fully observed data.

The projected graphical Lasso estimator with an ℓ_1 penalty seems to be slightly worse than its nonprojected counterpart. Projection does, however, allow for the use of nonconvex penalties in more difficult settings without the large degradation in performance. This may be desireable in some scenarios but in practice seems to simply add noise.

CHAPTER III

Covariance Estimation for Matrix-Variate Data with Missing Values and Mean Structure

The matrix-variate model, which allows for dependence along both axes of the data matrix, is an increasingly popular way to handle the complex and dependent data in modern data analysis. Its ability to model relationships between observations as well as covariates makes it useful for analyzing data with temporal, geographical, or other network relationships between them. Thus, applications for matrix-variate models often arise in biology, genetics, economics, climate science, and many other fields, where it is important to use methods that can at least account for these types of dependencies.

In this chapter, we propose methods for estimating both the row- and column-wise precision and covariance matrices in a matrix-variate data setting with missing data. In particular, we incorporate missing values with varying sample rates by column as well as two-group mean structure. These are based on the graphical Lasso estimator and assume sparsity in the inverse covariance (or precision) matrix and therefore also in the undirected graphs that they encode (in the Gaussian case). We establish the conditions required for consistency and present the convergence rates of our estimators, which attain the same rates as in the fully observed setting when the missing rates are fixed, but also allow for decreasing sampling rates as the data size increases.

In particular, existing graphical Lasso models generally assume matrices with zero means for simplicity, even though few practical applications meet this assumption. Rather than only focusing on mean-zero random matrices, we propose estimators that can jointly estimate a two-group mean structure with known groups. *Hornstein et al.* (2019) studied this problem in the fully observed case, but the we show how to prove similar results with missing values. We present theoretical results in this case and show how the demeaning process affects our estimation performance.

The remainder of this chapter is organized as follows. In Section 3.1 we present the data model we consider, which includes mean structure, matrix-variate dependence, and missing data. We also develop and present estimators for the sparse precision matrices along both data axes, including the intermediate covariance estimates and mask estimates required for fully automated estimation. Section 3.2 presents the theoretical consistency and convergence results for our methods. The proofs of these results are deferred until Chapter V. In Section 3.3, we present numerical examples and simulated comparisons testing the performance of our methods. Section 3.4 concludes.

3.1 Model and methods

We first present a model for matrix-variate data with two-group mean structure and missing values. We then present a present methodologies for estimating the sparse row-wise precision matrix both when the mean matrix is known and can be perfectly removed and when the mean is unknown.

Our (unobserved) full data follows the same basic model as in *Hornstein et al.* (2019). Consider a data matrix in $\mathbb{R}^{n \times m}$ composed of a mean matrix $\mathbb{M}$ and an error term $\mathbb{X}$.

$$X = \mathbb{M} + \mathbb{X} \tag{3.1}$$

Instead of this full matrix, we will observe a version with missing completely-at-random entries. For column-wise sampling rates $\zeta_1, \ldots, \zeta_m$, we then observe

$$\mathcal{X} = \mathbb{U} \circ (\mathbb{M} + \mathbb{X}) \quad \text{for } \mathbb{U} \in \{0,1\}^{n \times m}, \quad \mathbb{U}_{ij} \overset{\text{iid}}{\sim} \text{Bernoulli}(\zeta_j), \tag{3.2}$$

For now we assume the mean matrix has a known two-group structure, though this can be extended to k groups. This is a setting where the individual group labels are known, such as a genomics study with experimental and control groups (e.g. presence or non-presence of a disease, such as in *Hornstein et al.*, 2019) or the voting data we study in this paper with known party membership. In future work we hope to extend these methods to unlabelled data, where low-rank or clustering methods can be used to demean. So let $n = n_1 + n_2$ and assume that our data are sorted by the group label. Then we have $\mathbb{M} = D\mu$, where $D \in \mathbb{R}^{n \times 2}$ is the design matrix

$$D = \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 \\ \underbrace{0 \quad \cdots \quad 0}_{n_1} & & & \underbrace{1 \quad \cdots \quad 1}_{n_2} \end{pmatrix}^T$$

and $\mu = (\mu^{(1)}, \mu^{(2)})^T \in \mathbb{R}^{2 \times m}$ is a matrix of means for each variable and group.

By removing the group means, our estimates capture links between observations through the error term, or how they deviate from their means. This ensures group membership will not obfuscate these connections, allowing us to discover interesting connections and patterns not necessarily observable in the mean structure.

We assume that $\mathbb{X}$ has mean-zero subgaussian entries and a separable covariance structure, i.e. $\text{Cov}(\text{vec}(\mathbb{X})) = A_0 \otimes B_0$, where $A_0 \in \mathbb{R}^{m \times m}$ and $B_0 \in \mathbb{R}^{n \times n}$ are positive-definite covariance matrices and $\otimes$ denotes the Kronecker product. If we additionally assume $\mathbb{X}$ is distributed matrix-variate normal, then A_0^{-1} and B_0^{-1} encode the conditional independence relationships between columns and rows, respectively. In the subgaussian case this does not hold in general, but we can still estimate these precision matrices and obtain sparse partial correlation estimates using them.

Due to the structure of our covariance, A_0 and B_0 are only identifiable up to a constant factor, as for any constant c, $A_0 \otimes B_0 = (cA_0) \otimes (B_0/c)$. For this work we will focus on estimating the correlation matrices $\rho(A_0)$ and $\rho(B_0)$, as their sparse inverses encode the same graphical information as in A_0^{-1} and B_0^{-1}. We denote the true (normalized) precision matrices as $\Theta_0 = \rho(A_0)^{-1}$ and $\Phi_0 = \rho(B_0)^{-1}$.

Covariance estimation. Our estimator is then formed by first constructing appropriate estimates of the covariance matrices for each axis of the data matrix. Then, like in the traditional graphical lasso, these covariance estimates are plugged into penalized optimization programs to obtain the sparse estimates of the inverse covariance that we desire.

In our case, however, the presence of mean structure and missing data means that we cannot use the basic Gram matrices that the basic graphical lasso uses. For the mean structure, we show that even with missing data we can perform group-wise demeaing on each variable and still obtain convergence. And to account for the missing data we apply an element-wise adjustment to the Gram matrices that we call a mask matrix. In the following sections, we first present estimators assuming knowledge of the true masks, as well as demonstrate how mask estimators can be plugged in.

Let P be the matrix that performs by-group column sums:

$$P = \begin{bmatrix} \vec{1}_{n_1}\vec{1}_{n_1}^T & 0 \\ 0 & \vec{1}_{n_2}\vec{1}_{n_2}^T \end{bmatrix}$$

Then our centered data is

$$\tilde{\mathcal{X}} = \mathbb{U} \circ (\mathcal{X} - (P\mathcal{X}) \oslash (P\mathbb{U})) \tag{3.3}$$

Note that we are simply demeaning each group and variable with the observed means (contained in $P\mathcal{X} \oslash P\mathbb{U}$), but the notation is complicated by the missing values in our dataset.

Using this demeaned matrix, we form Gram matrices

$$\hat{S}(A) = \tilde{\mathcal{X}}^T \tilde{\mathcal{X}} \qquad \hat{S}(B) = \tilde{\mathcal{X}} \tilde{\mathcal{X}}^T \tag{3.4}$$

Estimators with known masks. We first estimate the column covariance A_0. Since $v^i \sim \mathbf{v}$ for $i = 1, \cdots, n$ are independent, we define the mask matrix

$$M := \mathbb{E} v^i \otimes v^i = \begin{bmatrix} \zeta_1 & \zeta_1\zeta_2 & \zeta_1\zeta_3 & \cdots & \zeta_1\zeta_m \\ \zeta_2\zeta_1 & \zeta_2 & \zeta_2\zeta_3 & \cdots & \zeta_2\zeta_m \\ \zeta_3\zeta_1 & \zeta_3\zeta_2 & \zeta_3 & \cdots & \zeta_3\zeta_m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \zeta_m\zeta_1 & \zeta_m\zeta_2 & \zeta_m\zeta_3 & \cdots & \zeta_m \end{bmatrix} \tag{3.5}$$

Assuming knowledge of ζ, we can create an estimator for A_0,

$$\widetilde{A} = \tilde{\mathcal{X}}^T \tilde{\mathcal{X}} \oslash (\mathrm{tr}(B_0) M), \tag{3.6}$$

and a corresponding estimator for the correlation $\rho_{ij}(A)$,

$$\widetilde{\Gamma}_{ij}(A) = \frac{\widetilde{A}_{ij}}{\sqrt{\widetilde{A}_{ii}\widetilde{A}_{jj}}} = \frac{\hat{S}(A)_{ij}}{\sqrt{\hat{S}(A)_{ii}\hat{S}(A)_{jj}}} \frac{1}{\sqrt{\zeta_i \zeta_j}}. \tag{3.7}$$

For estimating the row covariance B_0, consider the mask matrix $\mathcal{M} \in \mathbb{R}^{n \times n}$, where

$$\mathcal{M}_{k\ell} = \begin{cases} \sum_{j=1}^m a_{jj} \zeta_j & \text{if } k = \ell \\ \sum_{j=1}^m a_{jj} \zeta_j^2 & \text{if } k \neq \ell \end{cases}. \tag{3.8}$$

We then get estimators

$$\widetilde{B} = \tilde{\mathcal{X}} \tilde{\mathcal{X}}^T \oslash \mathcal{M} \tag{3.9}$$

and

$$\widetilde{\Gamma}_{ij}(B) = \frac{\widetilde{B}_{ij}}{\sqrt{\widetilde{B}_{ii}\widetilde{B}_{jj}}} = \frac{\hat{S}(B)_{ij}}{\sqrt{\hat{S}(B)_{ii}\hat{S}(B)_{jj}}} \frac{\sum_{k=1}^{m} a_{kk}\zeta_k}{\sum_{k=1}^{m} a_{kk}\zeta_k^2} \tag{3.10}$$

Versions of these oracles estimators designed for the mean-zero case were first proposed by *Zhou* (2019), which we adapt to our current model with group means and including demeaning.

Given penalties λ_B, λ_A and constants $R_A > \|\rho(A_0)^{-1}\|_2, R_B > \|\rho(B_0)^{-1}\|_2$, our inverse correlation estimates are the modified graphical Lasso estimators studied in *Loh and Wainwright* (2017) and *Fan et al.* (2019) using these inputs:

$$\begin{aligned} \widetilde{A}_\rho &= \underset{A_\rho \succ 0, \|A_\rho^{-1}\|_2 \le R_A}{\arg\min} \quad \mathrm{tr}(\widetilde{\Gamma}(A)A_\rho^{-1}) - \log|A_\rho| + \lambda|A_\rho^{-1}|_{1,\mathrm{off}} \\ \widetilde{B}_\rho &= \underset{B_\rho \succ 0, \|B_\rho^{-1}\|_2 \le R_B}{\arg\min} \quad \mathrm{tr}(\widetilde{\Gamma}(B)B_\rho^{-1}) - \log|B_\rho| + \lambda|B_\rho^{-1}|_{1,\mathrm{off}}. \end{aligned} \tag{3.11}$$

Mask estimation. To obtain fully automated estimators of these precision matrices, we can finally plug-in mask estimates for M and $\mathcal{M}$ into the above. Here we present flexible mask estimates that converge fast enough to have a minimal impact on the overall convergence rate.

To estimate M, we use

$$\widehat{M}_{ij} = \begin{cases} \hat{\zeta}_i & i = j \\ \hat{\zeta}_i\hat{\zeta}_j & i \ne j \end{cases} \tag{3.12}$$

where $\hat{\zeta}_j$ is the average number of observed entries in column j.

For estimating $\mathcal{M}$, we propose the estimator

$$\widehat{\mathcal{M}}_{ij} = \begin{cases} \mathrm{tr}(\tilde{\mathcal{X}}\tilde{\mathcal{X}}^T) & i = j \\ \frac{n}{n-1}\,\mathrm{tr}(\tilde{\mathcal{X}}^T\tilde{\mathcal{X}} \circ \widehat{M}) - \frac{1}{n-1}\,\mathrm{tr}(\tilde{\mathcal{X}}^T\tilde{\mathcal{X}}) & i \ne j \end{cases} \tag{3.13}$$

where we adopt the estimator originally designed for the corresponding zero-mean case

(*Zhou*, 2020), as discussed in Section 3.2. When using the demeaned data, this mask estimator still proves effective enough to obtain convergence guarantees for the final estimator.

Plugging these estimates into the above results in correlation estimators

$$\widehat{\Gamma}_{ij}(A_0) = \frac{\widehat{A}_{ij}}{\sqrt{\widehat{A}_{ii}\widehat{A}_{jj}}} = \frac{\hat{S}(A)_{ij}}{\sqrt{\hat{S}(A)_{ii}\hat{S}(A)_{jj}}}\frac{1}{\sqrt{\hat{\zeta}_i\hat{\zeta}_j}} \tag{3.14}$$

$$\widehat{\Gamma}_{ij}(B_0) = \frac{\widehat{B}_{ij}}{\sqrt{\widehat{B}_{ii}\widehat{B}_{jj}}} = \frac{\hat{S}(B)_{ij}}{\sqrt{\hat{S}(B)_{ii}\hat{S}(B)_{jj}}}\frac{\mathrm{tr}(\tilde{\mathcal{X}}\tilde{\mathcal{X}}^T)}{\widehat{\mathcal{M}}_{ij}} \tag{3.15}$$

Which we then plug into the minimization programs as (3.11) to get estimators $\widehat{A}_\rho, \widehat{B}_\rho$.

3.1.1 Flexible mean estimation

We extend our methodology to allow for the flexible estimation of low-rank mean matrices $\mathbb{M}$ following the low-rank matrix factorization literature. Here, we will assume that for some $r < n, m$, we have matrices $E \in \mathbb{R}^{n \times r}$, $F \in \mathbb{R}^{m \times r}$ such that $\mathbb{M} = EF^T$. Note that the known two-group case above also falls under this model, where E is known and must have a two-group structure.

To estimate $\mathbb{M}$ we optimize

$$\hat{E}, \hat{F} = \underset{E,F}{\arg\min} \|\mathcal{X} - \mathbb{U} \circ (EF^T)\|_F^2 \tag{3.16}$$

to get $\hat{\mathbb{M}} = \hat{E}\hat{F}^T$. We solve this optimization program with an alternating least squares estimator as first proposed by (*Wiberg*, 1976) and detailed in *Buchanan and Fitzgibbon* (2005). The estimators then proceed along the same lines as above plugging in the new demeaned estimator

$$\tilde{\mathcal{X}}_{\mathrm{lr}} = \mathbb{U} \circ (\mathcal{X} - \hat{E}\hat{F}^T) \tag{3.17}$$

We do not provide theoretical results for this flexible method, as in particular theoretical results for low-rank matrix factorization with dependent and missing data are not available to the best of our knowledge. We do, however, show through testing and simulation that this method performs very similarly to the two-group estimator with known groups presented above, and in practice has similar convergence behavior.

3.1.2 Related work

Our sparse precision matrix estimators extend the matrix-variate methods in *Zhou* (2014), which also uses separate estimators for A_0 and B_0 based on the graphical Lasso estimator in the fully observed setting. For proving the similar concentration equalities needed, we often rely on the sparse Hanson-Wright inequalities developed in *Zhou* (2019) for use with our sparsified data.

Missing data is often handled in practice through imputation. However, in a matrix-variate setting with dependence we do not have independent observations, so existing methods for imputation are generally not appropriate. These methods include k-nearest neighbors, random forest-based methods, or multiple imputation by chained equations (*Van Buuren and Oudshoorn*, 1999; *Troyanskaya et al.*, 2001; *Stekhoven and Bühlmann*, 2011). *Jamshidian and Bentler* (1999), *Städler and Bühlmann* (2012), and *Städler et al.* (2014) develop EM algorithm-based estimators, but these all depend on independent observations and do not extend to the matrix-variate case.

Glanz and Carvalho (2018) develop a method for EM-style estimation of Kronecker product covariances with missing data, but they rely on many samples of the matrix-variate data matrix rather than a single instance. *Allen and Tibshirani* (2010) presents the only method we are aware of for estimation and imputation with a single instance of matrix-variate data, but the EM-style estimation method proposed is not computationally feasible, especially in the high-dimensional setting. They instead focus on approximate algorithms designed for imputation rather than estimation of the covariances.

Instead of relying on imputation of EM-style methods, our work instead corrects the input covariance estimators to account for the multiplicative errors. This approach was first applied to correct covariance estimates in regression problems by *Hwang* (1986). *Loh and Wainwright* (2013, 2015) suggested the use of the same methods for graphical model estimation with missing data in the independent observations setting, but we develop new estimators for the matrix-variate case. We also modify the initial correlation estimates that are plugged into the graphical Lasso procedure to account for the missing values, though our data setting is significantly complicated by the two-way dependence.

We additionally show that these estimators are consistent when there are unknown group means, similar to the results shown by *Hornstein et al.* (2019) in the fully-observed case. The random missing values here significantly complicate the analysis, as we have to carefully account for how their effects propogate across the row and column dependencies to affect both the mean and subsequent covariance estimates.

The graphical Lasso estimator and its variants are well-studied methods for estimating sparse precision matrices (*Banerjee et al.*, 2008; *Yuan and Lin*, 2007; *Friedman et al.*, 2008; *Rothman et al.*, 2008; *Ravikumar et al.*, 2011; *Zhou et al.*, 2010, 2011, and others). A key difference, however, is that our input correlation estimates are not positive definite, as they are in *Zhou* (2014) and in the standard graphical Lasso literature, since the mask shrinks the diagonal entries more than it does the off-diagonals.

This results in a potential unbounded objective problem, which we solve by imposing an additional spectral norm constraint as in *Loh and Wainwright* (2015, 2017). For more details, see Chapter II, which includes *Fan et al.* (2019) and studies this problem in detail and develops an alternating direction method of multipliers (ADMM) algorithm to optimize objectives of the form in (3.11) with non-positive definite input matrices, which we use to implement our estimator.

3.2 Theoretical results

We make the standard assumption that our covariance matrices have bounded eigenvalues.

Assumption 1. *There exist some constants* $0 < \underline{\phi}, \overline{\phi} < \infty$ *such that* $\underline{\phi} < \phi_{\min}(A_0), \phi_{\min}(B_0)$ *and* $\phi_{\max}(A_0), \phi_{\max}(B_0) < \overline{\phi}$*.*

We also need to assume that the group sizes grow at the same rate and lower bound how small the sampling rate can be.

Assumption 2. *Define* $\zeta_{\min} = \min_{j=1,\dots,m} \zeta_j$*. Assume* $n_1, n_2 \approx n$ *and* $\zeta_{\min} \gtrsim \sqrt{\frac{\log(m \vee n)}{n}}$*.*

To obtain convergence, we bound the number of nonzero off-diagonal entries in our precision matrices. Note that, when account for differeing sampling rates, using the estimated mask requires a slightly stricter assumption.

Assumption 3. *For convergence in the known-mask case, we need to assume*

$$|A_0^{-1}|_{0,\text{off}} = o\left(\frac{n\zeta_{\min}^5}{\log(m \vee n)} \wedge \frac{n^2\zeta_{\min}^2}{\|B\|_1^2}\right) \qquad |B_0^{-1}|_{0,\text{off}} = o\left(\frac{m\zeta_{\min}^6}{\log(m \vee n)} \wedge \frac{n^2\zeta_{\min}^4}{\|B\|_1^2} \wedge n^2\zeta_{\min}^6\right)$$

When using the estimated masks, we require the stronger assumption of

$$|B_0^{-1}|_{0,\text{off}} = o\left(\frac{m\zeta_{\min}^8}{\log(m \vee n)} \wedge \frac{n^2\zeta_{\min}^6}{\|B\|_1^2} \wedge n^2\zeta_{\min}^8\right)$$

We assume our rows are sorted by group, so we denote

$$\mathbb{U} = \begin{bmatrix} \mathbb{U}^1 \\ \mathbb{U}^2 \end{bmatrix}, \quad \mathbb{X} = \begin{bmatrix} \mathbb{X}^1 \\ \mathbb{X}^2 \end{bmatrix}, \quad \text{where } \mathbb{U}^1, \mathbb{X}^1 \in \mathbb{R}^{n_1 \times m} \text{ and } \mathbb{U}^2, \mathbb{X}^2 \in \mathbb{R}^{n_2 \times m}$$

We similarly decompose the row-wise covariance B into

$$B = \begin{bmatrix} B^1 & B^{12} \\ B^{12T} & B^2 \end{bmatrix}$$

Let C be some absolute constant, and define

$$\alpha = CK^2 \frac{\log m \|A_0\|_2}{\sum_{k=1}^m a_{kk}\zeta_k^2} + CK^2 \frac{\log^{1/2}(m \vee n)}{\sqrt{m}} \frac{1}{\|\zeta\|_2/\sqrt{m}} \frac{\sqrt{a_\infty \|A_0\|_2}}{a_{\min}} \tag{3.18}$$

$$\beta = CK^2 \frac{\log n \|B_0\|_2}{\zeta_{\min} \operatorname{tr}(B_0)} + CK^2 \frac{\log^{1/2}(m \vee n)}{\sqrt{n}} \frac{1}{\zeta_{\min}} \frac{\sqrt{n}\|B_0\|_F}{\operatorname{tr}(B_0)} \tag{3.19}$$

These are the rates of convergence when the mean structure is known and therefore can be perfectly removed, which we show in Theorem 5.3.3. Note that under Assumption 1 the terms $\sqrt{a_\infty \|A_0\|_2}/a_{\min}$ and $\sqrt{n}\|B_0\|_F/\operatorname{tr}(B_0)$ can both be upper bounded by constants.

We first present results assuming the masks are known and covariance estimators are formed as in (3.7) and (3.10).

Theorem 3.2.1. *Consider data generating random matrices as in* (3.1) *and* (3.2) *and suppose Assumptions 1 and 2 hold. Let* $m \vee n \geq 3$, *and for some absolute constants* C_1, C_2 *define*

$$\begin{aligned} \alpha_{\text{mean}} = {} & \frac{C_1}{\zeta_{\min}^2 n_{\min}} \left(\|B\|_1 + \frac{1}{\zeta_{\min}} \frac{\operatorname{tr}(B)}{n_{\min}} \right) + C_1 K^2 \frac{\log m}{\zeta_{\min}^3 \operatorname{tr}(A)} \|A\|_2 \\ & + C_1 K^2 \log^{1/2}(m \vee n) \frac{\|A\|_F}{\zeta_{\min}^3 \operatorname{tr}(A)} + \alpha \end{aligned} \tag{3.20}$$

$$\begin{aligned} \beta_{\text{mean}} = {} & C_2 \frac{1}{\zeta_{\min} \operatorname{tr}(B_0)} \left(\frac{\operatorname{tr}(B^1)}{n_1} + \frac{\operatorname{tr}(B^2)}{n_2} \right) + C_2 \frac{1}{\zeta_{\min} \operatorname{tr}(B_0)} \left(\frac{1}{n_1} \left| \sum_{k \neq \ell} B^1_{k\ell} \right| + \frac{1}{n_2} \left| \sum_{k \neq \ell} B^2_{k\ell} \right| \right) \\ & + C_2 K^2 \frac{\log n}{\zeta_{\min}^2} \frac{\|B^1\|_2 + \|B^2\|_2}{\operatorname{tr}(B_0)} + C_2 K^2 \frac{\log^{1/2}(m \vee n)}{\zeta_{\min}^{5/2}} \frac{\|B^1\|_F + \|B^2\|_F}{\operatorname{tr}(B_0)} + \beta \end{aligned} \tag{3.21}$$

Let $\widetilde{B}_\rho, \widetilde{A}_\rho$ *be the unique minimizers defined by* (3.11) *with the input correlation matrices* $\widetilde{\Gamma}(B_0), \widetilde{\Gamma}(A_0)$. *Penalties are chosen as*

$$\lambda_B = \frac{1}{\epsilon}(3\alpha_{\text{mean}}) \qquad \lambda_A = \frac{1}{\varepsilon}(3\beta_{\text{mean}}) \tag{3.22}$$

for some $0 < \varepsilon, \epsilon < 1$, *and* $R_A > \|\rho(A_0)^{-1}\|_2$, $R_B > \|\rho(B_0)^{-1}\|_2$*.j Let* $\alpha_{\text{mean}}, \beta_{\text{mean}} < 1/3$ *and* *let C_3, C_4 be some absolute constants.*

Then, with probability at least $1 - 25/(m \vee n)^2$, we get that

$$\|\widetilde{B}_\rho - \rho(B_0)\|_2 \leq \|\widetilde{B}_\rho - \rho(B_0)\|_F \leq C_3 \kappa(\rho(B_0))^2 \lambda_B \sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1}$$

$$\|\widetilde{B}_\rho^{-1} - \rho(B_0)^{-1}\|_2 \leq \|\widetilde{B}_\rho^{-1} - \rho(B_0)^{-1}\|_F \leq \frac{C_3 \lambda_B \sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1}}{2\varphi^2_{\min}(\rho(B_0))} \tag{3.23}$$

Similarly, with probability at least $1 - 26/(m \vee n)^2$,

$$\|\widetilde{A}_\rho - \rho(A_0)\|_2 \leq \|\widetilde{A}_\rho - \rho(A_0)\|_F \leq C_4 \kappa(\rho(A_0))^2 \lambda_A \sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1}$$

$$\|\widetilde{A}_\rho^{-1} - \rho(A_0)^{-1}\|_2 \leq \|\widetilde{A}_\rho^{-1} - \rho(A_0)^{-1}\|_F \leq \frac{C_4 \lambda_A \sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1}}{2\varphi^2_{\min}(\rho(A_0))} \tag{3.24}$$

These all converge under Assumption 3.

In addition, we can prove similar results for our automated estimators (3.14) and (3.15) using mask estimates $\widehat{M}$ and $\widehat{\mathcal{M}}$. Theorem 3.2.2 shows these results, which have similar rates but pay some additional costs to account for the mask estimation.

Theorem 3.2.2. *Under the same conditions as in Theorem 3.2.1, let $\widehat{A}_\rho, \widehat{B}_\rho$ be the results of plugging $\widehat{\Gamma}(A), \widehat{\Gamma}(B)$ as defined in* (3.14) *and* (3.15) *into the optimization program* (3.11) *with penalties*

$$\lambda'_B = \frac{1}{\epsilon} 9 \frac{\alpha_{\text{mean}}}{\zeta_{\min}} \qquad \lambda'_A = \frac{1}{\varepsilon}\left(12\beta + 3\beta_{\text{mean}}\right) \tag{3.25}$$

Then we get convergence of $\widehat{B}_\rho$ and $\widehat{A}_\rho$ in the same sense as in Theorem 3.2.1, where the bounds on $\widetilde{B}_\rho, \widetilde{B}_\rho^{-1}$ hold with probability at least $1 - 25/(m \vee n)^2$ and the bounds on $\widetilde{A}_\rho, \widetilde{A}_\rho^{-1}$ hold with probability at least $1 - 26/(m \vee n)^2$, where we require Assumption 3 using the stricter assumption on $|B_0^{-1}|_{0,\text{off}}$.

Remark. To prove Theorem 3.2.1, we first prove similar results for when the group means are known. In this case, we have the data

$$\bar{\mathcal{X}} = \mathbb{U} \circ \mathbb{X} \quad \text{for } \mathbb{U} \in \{0,1\}^{n \times m}, \quad \mathbb{U}_{ij} \overset{\text{iid}}{\sim} \text{Bernoulli}(\zeta_j), \tag{3.26}$$

and define $\bar{\Gamma}_A, \bar{\Gamma}_B$ to be analogues of the estimators defined in (3.14) and (3.15), replacing $\widetilde{\mathcal{X}}$ with $\bar{\mathcal{X}}$. Section 5.3 proves the convergence of the zero- or known-mean covariance and correlation estimates at the rates as defined in (3.18) and (3.19) for B and A, respectively. The known-mask results were first presented in *Zhou* (2019), but we include the full proofs here for completeness and completed the full rates using the mask estimators.

The proof of Theorem 3.2.1 then bounds how far the mean estimation causes our estimates to deviate from this zero-mean baseline. Section 5.4.1 does this for when the masks are known, while Section 5.4.3 bounds the difference in the mask estimates used to prove Proposition 3.2.2.

Our analysis of the demeaned estimator decomposes the overall error into the error of the zero-mean estimator and the additional error from demeaning. Similar to *Hornstein et al.* (2019), these error terms manifest as a bias contribution, e.g. the first term in (3.20), and variance contributions, the second and third terms. However, the presence of random missing values makes controlling the error terms much more difficult, and we pay some additional factors of $1/\zeta_{\min}$ relative to the zero-mean rate. Simulations suggest that these factors may not be necessary with a tighter theoretical analysis.

Once we have ℓ_∞ rates on our correlation estimates, we apply Theorem 5.2.1, which then establishes the general convergence result for the graphical Lasso given convergence of the input correlation estimates. This is a standard result similar to those in *Rothman et al.* (2008) and *Zhou et al.* (2010). For completeness, we include the proof in Section 5.2, which is only slightly modified from the above work to allow for our side constraint.

3.3 Simulations

We use simulated data to demonstrate the performance of our methods under a variety of data scenarios. We generate matrix-variate normal data using three covariance models:

- AR(ρ): an autoregressive model with a single lag. This model sets $A = (\rho^{|i-j|})_{ij}$.
- SB(ρ, r): a star-block model. This covariance matrix is block-diagonal with blocks of size r, where blocks have a star-structure. In each block, a hub node is chosen, and the covariance is set to ρ for each nodes connection to the hub and ρ^2 otherwise.
- *ER*: an Erdős-Rényi random graph model, as described in *Zhou* (2014). We set the precision matrix $\Phi = 0.25 I_{n\times n}$. Then for each of n randomly selected edges (i_k, j_k), we choose a weight w uniformly from $[0.2, 0.4]$ and update $\Phi_{ij} = \Phi_{ji} \leftarrow \Phi_{ij} - w$, $\Phi_{ii} \leftarrow \Phi_{ii} + w$, and $\Phi_{jj} \leftarrow \Phi_{jj} + w$.

In these simulations we consider three estimation scenarios:

- The demeaned estimator is the baseline two-group estimator with missing data that demeans each group using known group labels constructs out estimates with this demeaned matrix.
- The low-rank estimator uses the low-rank mean estimation methodology in Section 3.1.1 to demean a two-group data matrix without knowledge of the group labels.
- The no-mean estimator assumes oracle knowledge of the mean matrix, so the data matrix can be perfectly demeaned and then treated as a zero-mean matrix. So here no mean estimation is needed and the perfectly demeaned data matrix is used to construct estimates.

All of our estimators use the automated mask estimators, and therefore correspond to the estimators in (3.14) and (3.15).

We evaluate convergence using both the relative Frobenius and spectral norms, and evaluate model selection using the sum of the false positive and false negative rates (FPR+FNR) and the Matthews correlation coefficient (MCC), which can be interpreted as a correlation coefficient between the predicted and observed edges.[1]

Figures 3.1 and 3.2 show how the the performance of our estimators changes over the full regularization paths as we vary λ under three different topologies. The These show that our estimator can approximately recover the graphical structure in the data as well as attain reasonably low relative errors in terms of both relative Frobenius and spectral norms. We see that this performance improves for the relevant ranges of regularization as we increase the sampling rate, and therefore increase the effective sample size.

We can also see the effect that needing to demean the data matrix has on estimator performance. There is a small but consistent gap in performance between the no-mean and demeaned estimators when estimating the row precision matrix $\rho(B_0)^{-1}$, but when estimating the colun precision matrix $\rho(B_0)^{-1}$ the two estimators performn nearly identically, so no cost is paid from having to also estimate the means.

Figure 3.3 shows how demeaning affects the input correlation estimates $\widehat{\Gamma}_{ij}(A_0)$ and $\widehat{\Gamma}_{ij}(B_0)$. Recall that our convergence rate depends on the ℓ_∞ error bounds of these terms. We can see how the demeaning causes very little difference in the distribution of errors for estimating the A-side correlation, which is consistent with the lack of performance gap exhibited in Figure 3.2. On the B-side, for small sample sizes the demeaned and low-rank demeaned estimates exhibit both bias and additional variation compared to the no-mean estimator, but these quickly decrease as we increase the sample size and the mean is estimated more accurately.

Figure 3.4 shows how these three estimators perform as we vary the sample size, holding the ratio $m = 4n$ constant. On the B-side the gap between the demeaned and oracle estimators decreases as we increase the sample size, matching the results of Theorem 3.2.2.

[1] MCC is defined as $\mathrm{MCC} = \mathrm{TP} \times \mathrm{TN}/\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{TN})}$.

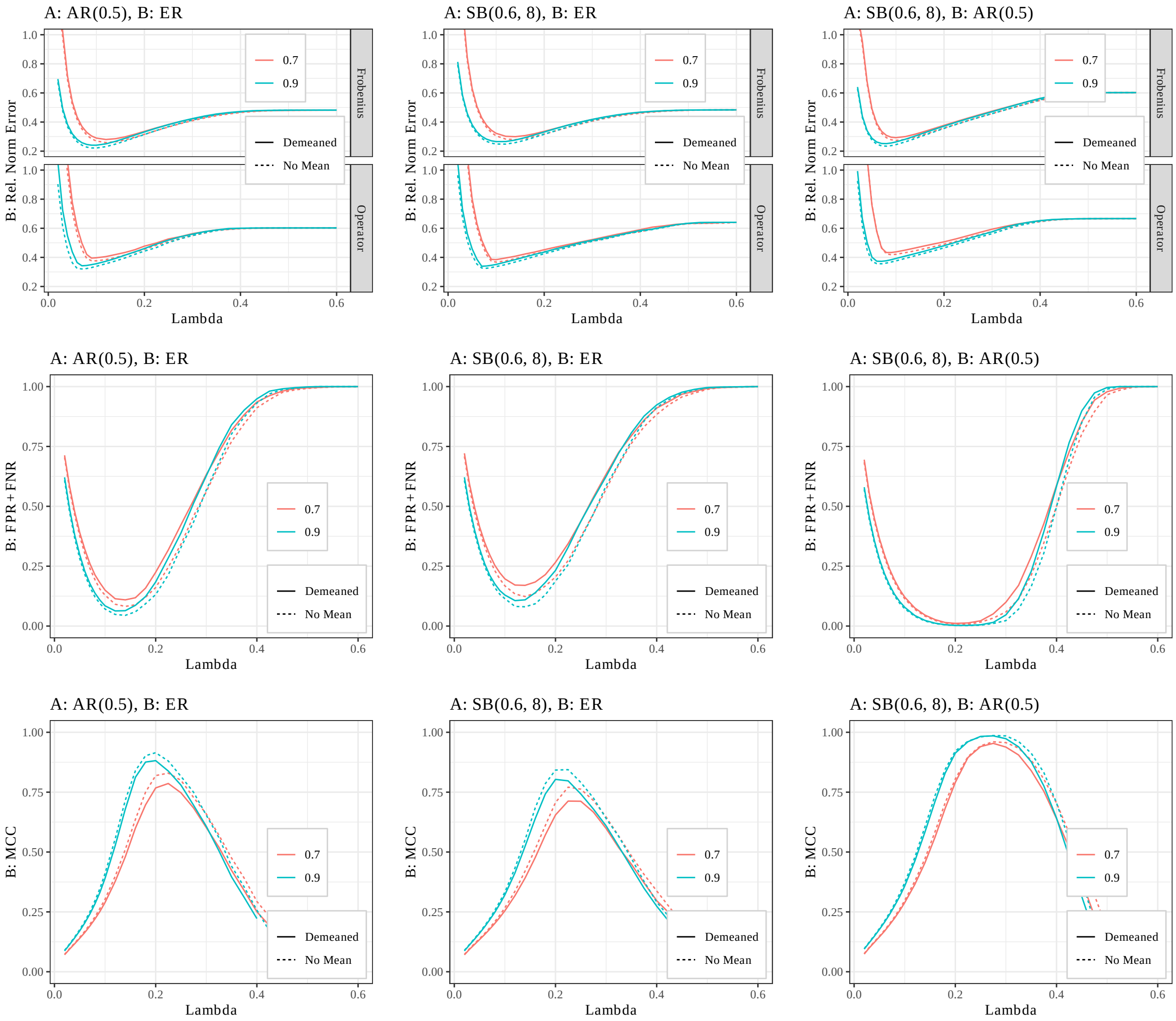


Figure 3.1: Performance of our penalized precision matrix estimators for $\rho(B_0)^{-1}$ for homogenous sampling rates over the regularization path. We set $m = 400$ and $n = 160$. The top row shows relative Frobenius and spectral norm errors, while the middle and bottom rows show selection performance in terms of FPR+FNR and MCC. We let $\zeta_1 = \cdots = \zeta_m = \zeta$, where $\zeta = 0.7, 0.9$.

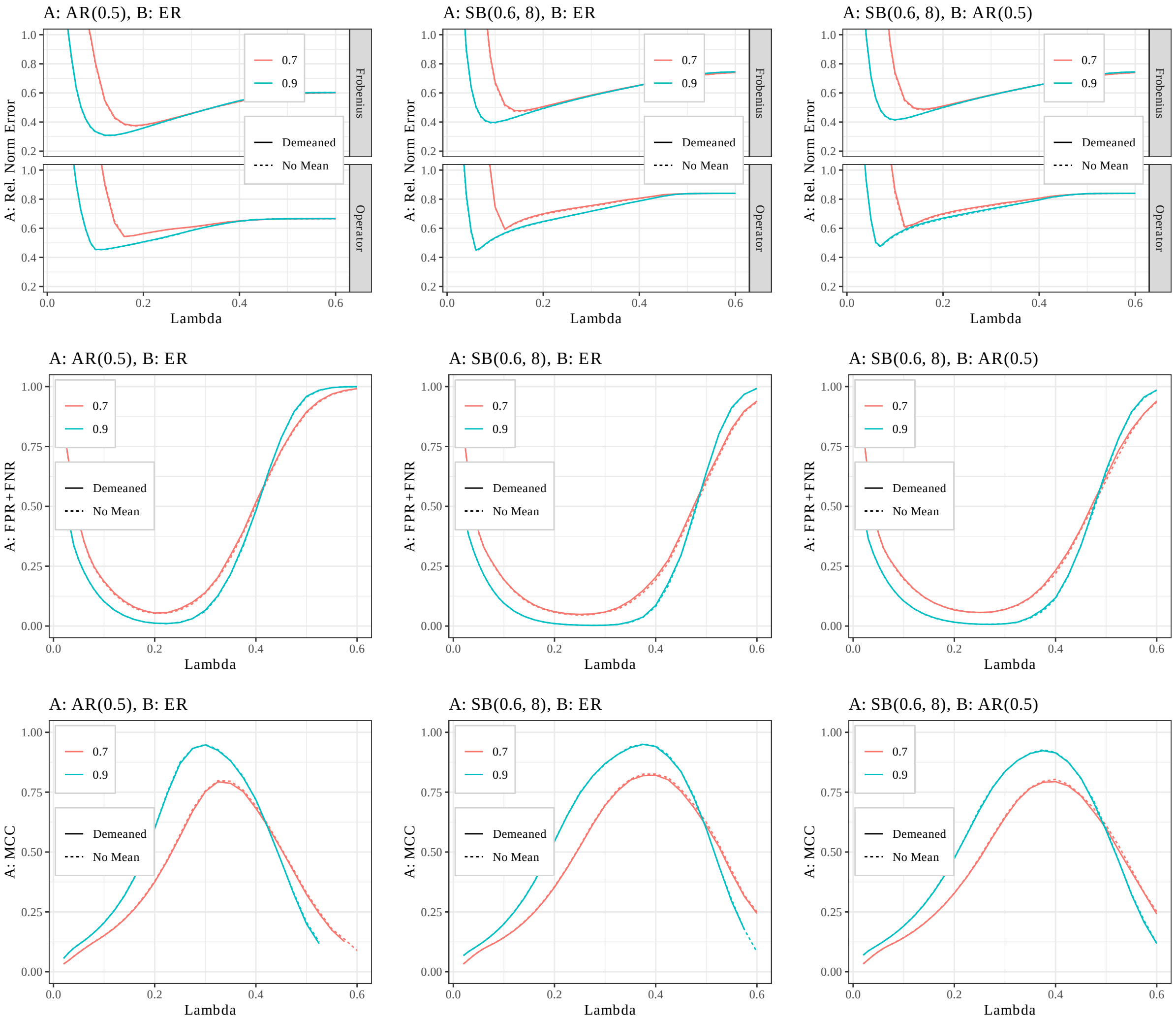
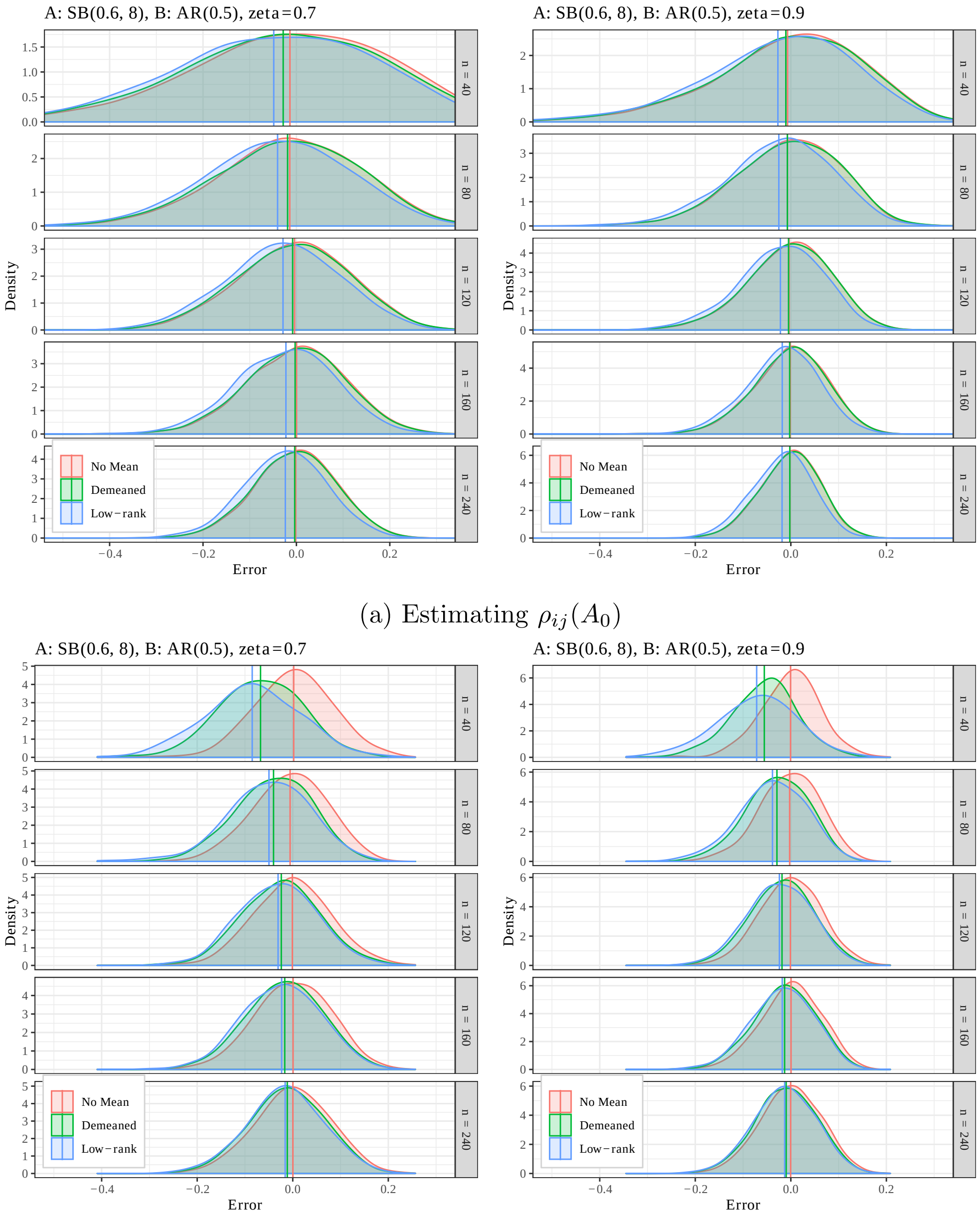


Figure 3.2: Performance of our penalized precision matrix estimators for $\rho(A_0)^{-1}$ for homogenous sampling rates over the regularization path. We set $m = 400$ and $n = 160$. The top row shows relative Frobenius and operator norm errors, while the bottom row shows selection performance in terms of FPR+FNR and MCC. We let $\zeta_1 = \cdots = \zeta_m = \zeta$, where $\zeta = 0.7, 0.9$.



(a) Estimating $\rho_{ij}(A_0)$

(b) Estimating $\rho_{ij}(B_0)$

Figure 3.3: Comparison of our initial correlation estimator $\widehat{\Gamma}$ on the demeaned data with the no-mean estimator using mean zero data. We set $m = 400$. For A, we present the errors for estimating the "spoke" connections in each star-block that connect nodes to the center node. For B we present the density of estimated errors for estimating $\rho_{i,i+1}(A_0)$, or the 1-off diagonal terms.

On the A-side these estimators track very closely in performance for the full range. The no-demean estimator, which applies the zero-mean methodology to a matrix with two-group means, shows the importance of correctly accounting for the mean in these scenarios. This estimator that ignores the mean never converges.

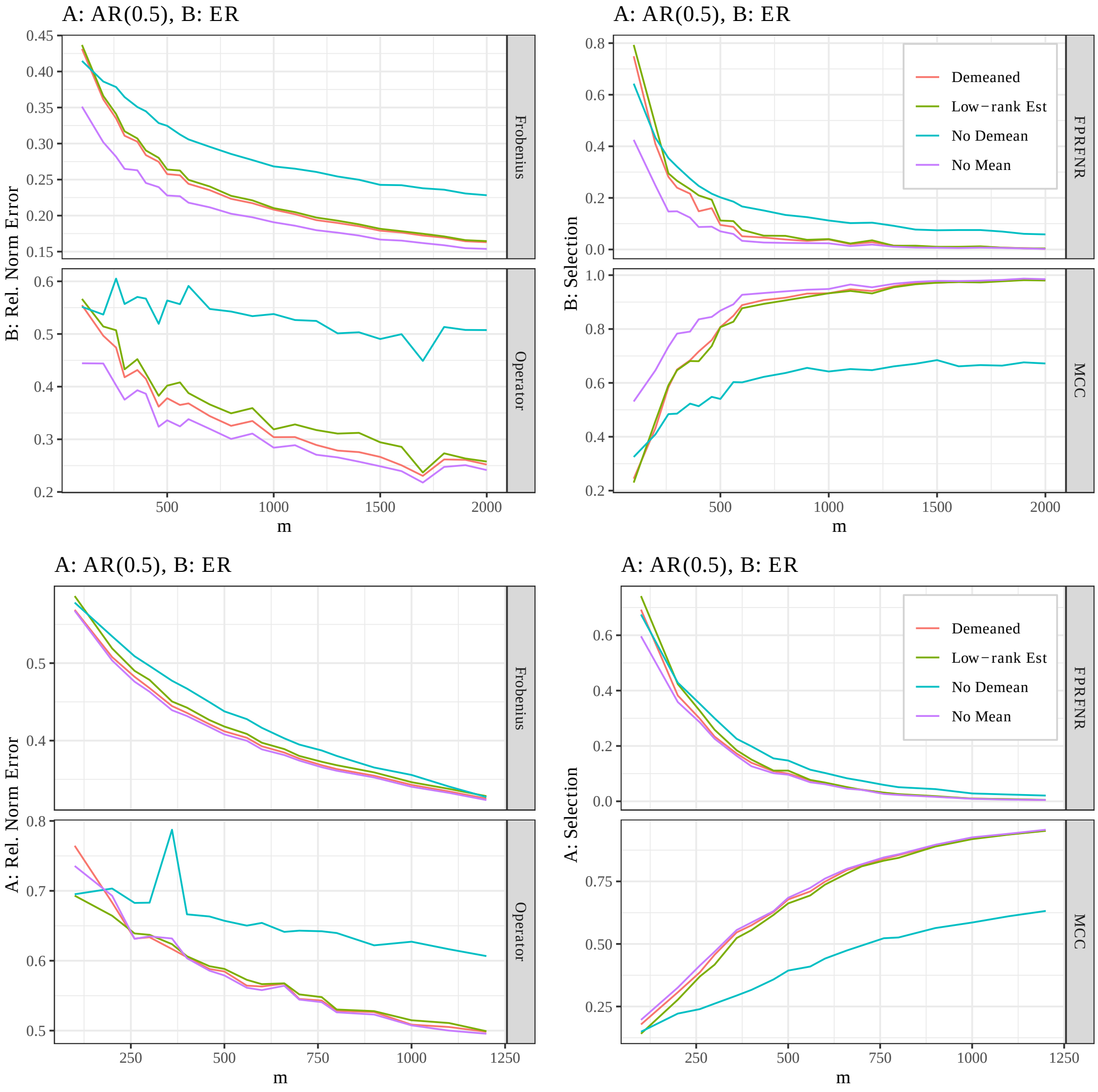


Figure 3.4: Performance of our precision matrix estimators in the presence of two-group mean structure. Here we set $m = 4n$ and fix $\zeta = 0.7$. We choose groups means to have a difference of $|\mu^{(1)} - \mu^{(2)}| = 1$. Each point shows the optimal λ over a range of values. The left panel shows performance in terms of relative Frobenius and operator norm error, while the right shows selection performance.

Figure 3.5 shows the behavior as we vary the sampling rate. As expected our estimators perform worse as the sampling rate, and therefore effective sample size, decreases. The demeaned and no-mean estimators, however, react almost identically as we vary the sampling rate. This suggests that the additional factors of $1/\zeta_{\min}$ that we pay in our theoretical rate when demaning are likely not binding in practice. In some scenarios it may even be possible to show tighter rates with respect to ζ.

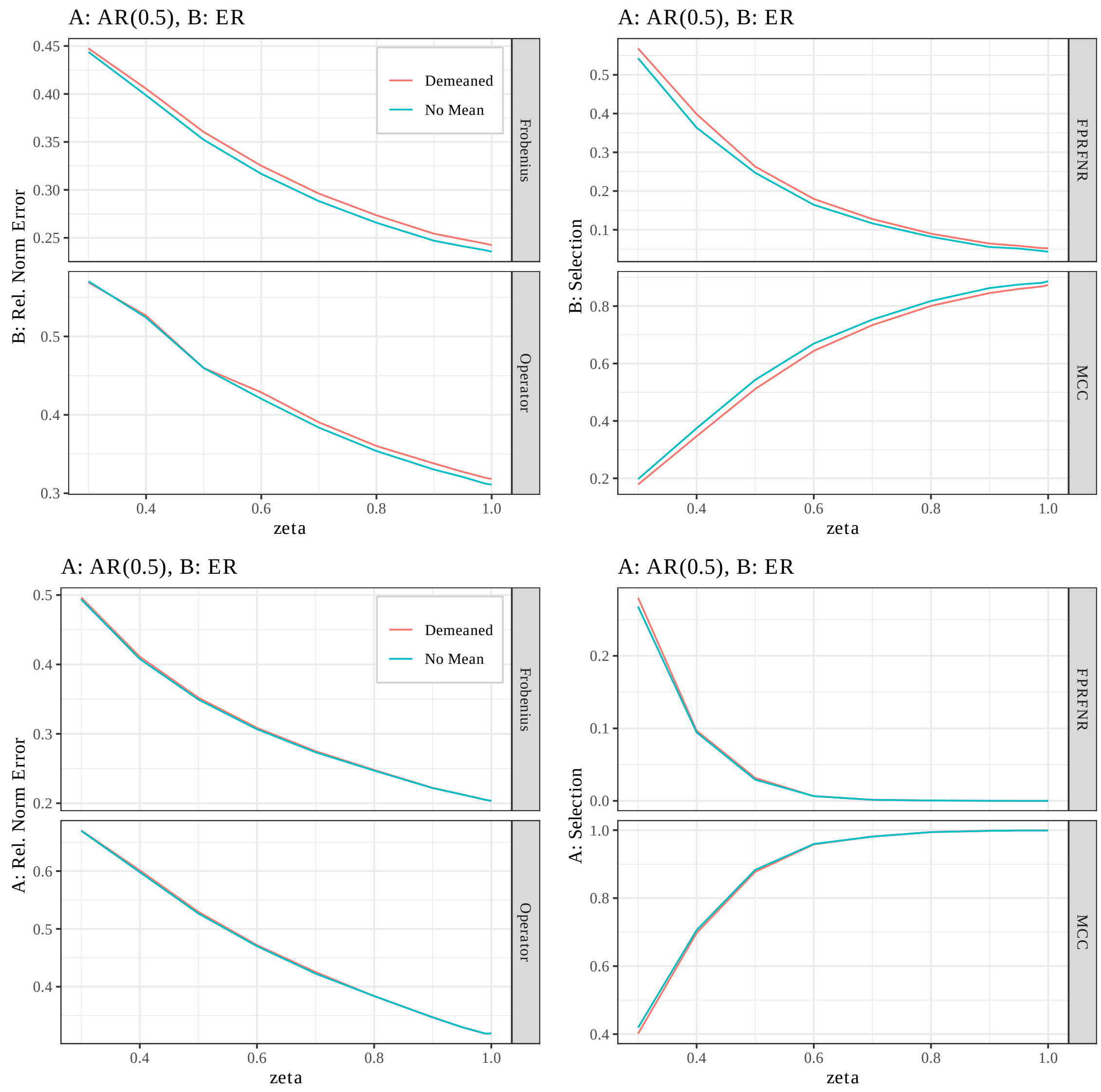


Figure 3.5: Performance of our precision matrix estimators in the presence of two-group mean structure. Here we set $m = n = 400$ and vary the sampling rate ζ. We choose groups means to have a difference of $|\mu^{(1)} - \mu^{(2)}| = 1$. Each point shows the optimal λ over a range of values. The left panels show performance in terms of relative Frobenius and operator norm error, while the right show selection performance.

Figure 3.6 explores how our estimators react to heterogenous sampling rates. In this experiment we only vary the sampling rate of 5% of columns. On the B-side we see that this has very little impact on the estimation performance. This matches our theoretical results, as the B-side convergence rate in (3.18) depends on an average sampling rate $\|\zeta\|$, rather than the minimum. Although $\zeta_{\min}$ terms are introduced in (3.20), as discussed above in practice these are likely on non-dominant terms.

Considering the A-side convergence in (3.19), it is then no surprise that the A-side estimation is sensitive to varying a small number of sampling rates, since it is more dependent on $\zeta_{\min}$. So varying the sampling rate of a small number of columns makes a significant impact on the overall estimation.

Figure 3.7 compares our method to using the standard graphical Lasso with naive mean imputation. The left panel holds the size of the data matrix constant while we vary the sampling rate. As expected, we see that the methods are identical when there are no missing values, but that a significant gap in the relative norm performance quickly appears as we introduce missing values.

The right panel demonstrates the convergence behavior of the two methods as we hold the sampling rate constant and increase both n and m, fixing $m = 4n$. We see that our proposed method quickly converges to low relative norm errors, while the imputed version converges at a much slower rate. Theoretically, we do not expect the version with mean imputation to be able to achieve convergence to zero, since its input covariance estimates will always be biased by the lack of adjustment.

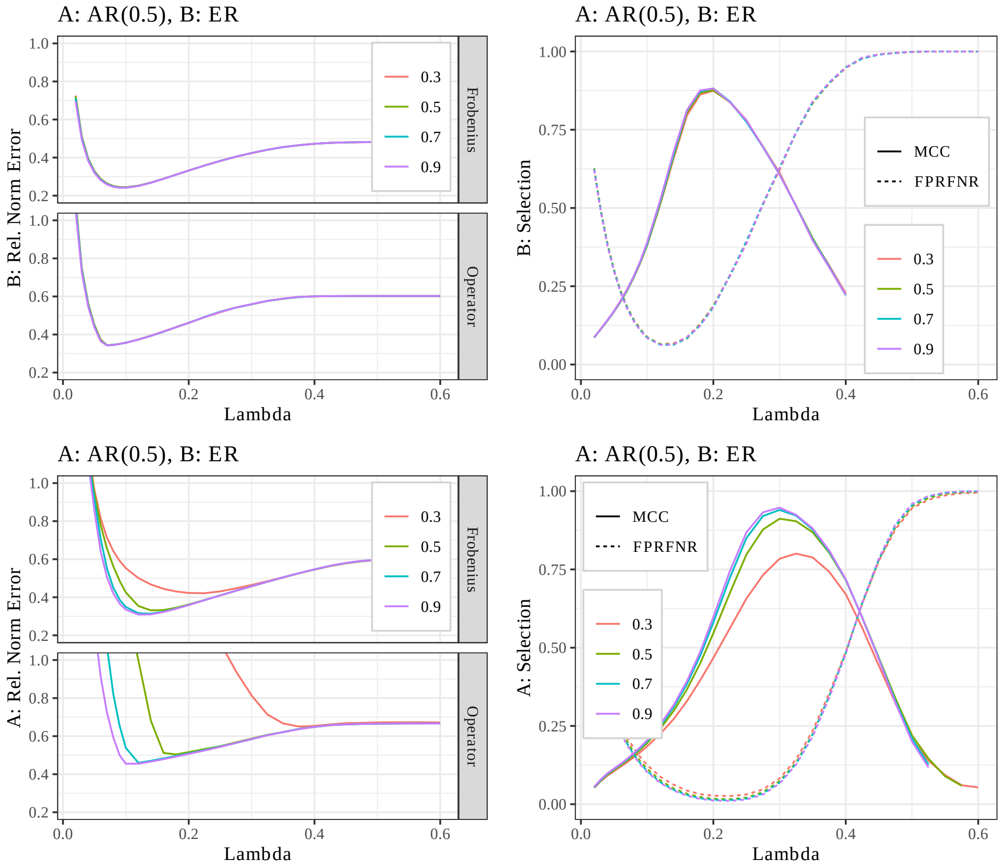
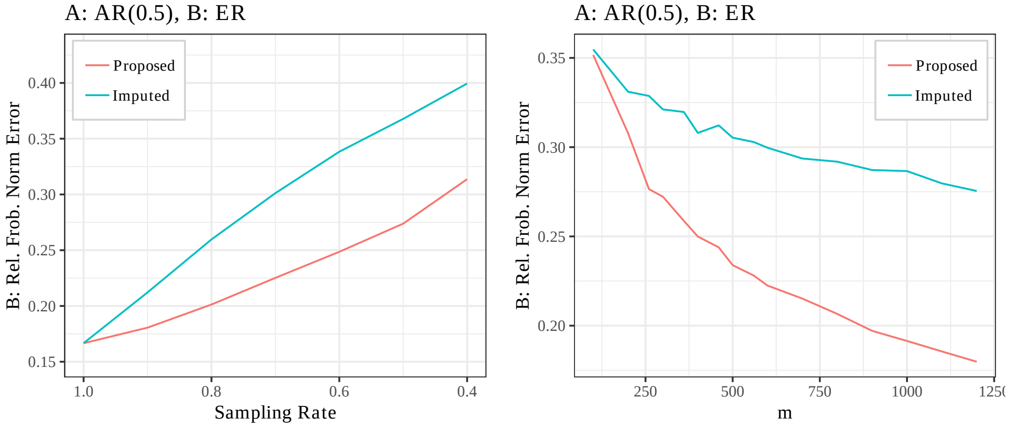


Figure 3.6: Performance of our penalized precision matrix estimators for heteogenous sampling rates. We set $m = 400$ and $n = 160$. For 95% of columns we set $\zeta_j = 0.9$ and for the remaining 5% we set $\zeta_j = 0.3, 0.5, 0.7, 0.9$. The left column show relative Frobenius and operator norm errors, while the right column shows selection performance in terms of FPR+FNR and MCC. All panels in this figure display results for the demeaned estimator.



(a) $n = 100$, $m = 600$

(b) $\zeta = 0.7$, $m = 4n$

Figure 3.7: Comparison of our proposed estimator for $\rho(B_0)^{-1}$ to using the standard graphical Lasso with naive mean imputation. In the left panel, we fix n and m and vary the sampling rate. In the right panel we fix the sampling rate and vary n and m. The minimum relative Frobenius error achieved over a range of penalization values λ is plotted at each point.

3.4 Conclusion

We have developed a method for estimating the sparse row precision matrix of a matrix-variate data matrix when there is missing data with varying missing rates by column. We have shown that this method is effective through both theoretical and empirical investigation, where our theoretical results rely on recently developed concentration inequalities under masks. We also present a statistical methodology for performing this estimation in the presence of two-group mean structure and show the presence of mean structure affects the convergence rates.

We also note that there are many applications with data that may exhibit these types of observation-observation dependencies. These methods show promise for applications such as flexibly correcting for dependent experimental design, estimating the connectivity structure across both space and time in medical imagining data, or estimating observation-observation networks in data collected from social or physical networks. In future work we hope to make this method more flexible, especially in allowing for more complex missing data structures, to enable more of these types of analyses.

CHAPTER IV

U.S. Senate Data Analysis

We explore a dataset of U.S. Senate voting records and apply the proposed methodology.[1] We will focus on two recent Congresses, the 114th (2015-2016) and 115th (2017-2018). These took place during the last two years of President Obama's second term and the first two years of President Trump's, and marked a significant shift in the political climate. We will also compare these results to those from the 106th Congress (1999-2000), during the final two years of President Clinton's second term.

In this data, we believe that there should be natural correlations between both senators and bills. Senators will naturally be correlated with other like-minded senators, not just due to party membership but also because of factors like the states and geographical areas they represent, basic ideological and philosophical beliefs, and political considerations like forming voting blocs. We also think that bills or votes may be correlated. For instance, multiple bills on similar issues may induce similar voting patterns, or there may be multiple votes with similar purposes, like rejecting poison-pill amendments to the same bill.

All three of these Congresses exhibit Republican majorities in the Senate. The 106th and 114th Congresses are in particular very similar, both in the last years of Democratic presidents and with similar Republican majorities (54 to 46 or 55 to 45). The 115th Congress has a smaller Republican majority, where they have 50-52 of the votes over the course of the

[1] This data is part of the public record and open-source code to download and process the data can be found at `https://github.com/unitedstates/congress`.

time period.

We remove senators who did not serve full terms as well as unanimous or unanimous-by-party votes, leaving us with datasets of 96-100 senators and 414-549 votes per Congress. Missing values in the dataset come from “Not Voting” or “Present” votes, consisting of roughly 2-3% of all votes. Individual bills have vote rates ranging from 70% to 100%, with most falling between 90% and 100%. Previous analyses of this data have generally imputed missing votes, often with “Nay” votes (*Banerjee et al.*, 2008; *Guo et al.*, 2015a). This likely does not significantly bias results in this dataset given the small missing rates, but other roll call datasets of interest (such as the European Parliament data studied in *Han*, 2007) have significantly higher missing rates and imputation may not be appropriate.

4.1 Related work

Note that the voting data here is binary, while most of the methodology we have developed in Chapter III is designed for graph estimation of continuous, and especially multivariate normal data. While we do not present theoretical results for use with binary data in this work, we do hope to partially bridge this gap with some proposed methodology and simulation results.

There have been several proposed methods for estimating graphical models with discrete data in the independent case. Many of the most popular models are based on Markov Random Fields and in particular the Ising model. *Ravikumar et al.* (2010) develop a neighborhood selection-based estimator for this model based on running individual logistic regressions, similar in spirit to the method in *Meinshausen and Bühlmann* (2006). Using logistic regression-based estimators, *Guo et al.* (2015a) develop an Ising model that separates observations into categories and allows for differences in the graphs between categories. They use this model to explore Senate roll call data, assuming senator networks are per-category but that otherwise votes are independent. Similarly *Kolar et al.* (2010) estimate a time-varying model on Senate voting records assuming graph parameters change smoothly over time using total variation-penalized logistic regressions.

Banerjee et al. (2008); *Kolar and Xing* (2008) develop global (rather than neighborhood) estimators for Ising models based on approximate likelihood upper bounds. *Banerjee et al.* (2008) in particular apply their method to Senate voting records data, assuming bills are independent.

Note that all of the voting recrods applications here use naive imputation techniques, generally replacing missing votes with either "No" votes or party majorities, and, although some allow the senator-senator graph to change for different bills, all assume independence on the bill-side of the data.

A separate line of models involve latent continuous variables, usually multivariate normal, that are discretized to form observed binary or ordinal data. When the latent variables are multivariate normal and directly discretized, this is sometimes known as the multivariate probit model (*Ashford and Sowden*, 1970). *Chib and Greenberg* (1998); *Guo et al.* (2015b) develop EM-based estimators for this model, while *Suggala et al.* (2017); *Fan et al.* (2017); *Feng and Ning* (2019) instead propose direct estimators of the latent correlation matrix. These all allow for estimation of the latent precision matrix and graphical model from discretized but otherwise fully observed data, but we leave adaptation of these methods to our setting with dependent and missing data to future work.

4.2 Data model

To obtain binary data with two-way dependence, we consider an Ising model of the following form.

$$\begin{aligned} P(X;\mu,A,B) = \exp\Big\{ & \sum_{i=1}^{n}\sum_{j=1}^{m}(\mathbb{1}_{1\le i\le n_1}\mu_j^1 + \mathbb{1}_{n_1<i\le n}\mu_j^2)X_{ij} \\ & - \sum_{i=1}^{n}\sum_{k=i}^{n}\sum_{j=1}^{m}\sum_{\ell=j}^{m} I_{(i,k)\neq(j,\ell)} B_{ik}^{-1}A_{j\ell}^{-1}X_{ij}X_{k\ell} - \theta(\mu,A,B)\Big\} \end{aligned} \tag{4.1}$$

where $\theta(\mu, A, B)$ is the normalizing constant. This essentially means the vectorized data $\text{vec}(X)$ is a single draw from a standard Ising model with interaction parameters defined by the off-diagonal elements of $(-A^{-1} \otimes B^{-1})$. As usual, we then observe $\mathbb{X} = \mathbb{U} \circ X$ for $\mathbb{U}_{ij} \overset{\text{i.i.d}}{\sim} \text{Bernoulli}(\rho_j)$.

Note that the fully observed data here reduces to the standard independent Ising model in *Banerjee et al.* (2008) when $B = I$ and $\mu_j^1 = \mu_j^2$. In this i.i.d. case, they develop a Gaussian approximation to the Ising log-likelihood based on replacing $\theta()$ with an upper bound. This is estimated using a graphical Lasso-type estimator based on the standard sample covariance

$$S_{ij} = \frac{1}{n} \sum_{k=1}^{n} (X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j).$$

In particular, *Banerjee et al.* (2008) use $S + (1/3)I$ in their estimator due to the nature of their upper bound.

Viallon et al. (2014) shows that this approximation is competitive with or outperforms other approximations and exact methods, including the logistic regression-based methods of *Ravikumar et al.* (2010). In particular, they use a modification that replaces $S + (1/3)I$ with the sample correlation matrix and show that this version of the estimator is competitive at recovering the structure of the Ising model while being significantly computationally cheaper than the alternatives.

We therefore adopt this methodology to our more complex case, with dependence from B, μ exhibiting two-group mean structure, and missing data, by replacing the simple sample correlation with the correlation estimators we develop in Chapter III:

$$\widehat{\Gamma}_{ij}(A_0) = \frac{\widehat{A}_{ij}}{\sqrt{\widehat{A}_{ii}\widehat{A}_{jj}}} = \frac{\hat{S}(A)_{ij}}{\sqrt{\hat{S}(A)_{ii}\hat{S}(A)_{jj}}} \frac{1}{\sqrt{\hat{\zeta}_i \hat{\zeta}_j}} \tag{3.14}$$

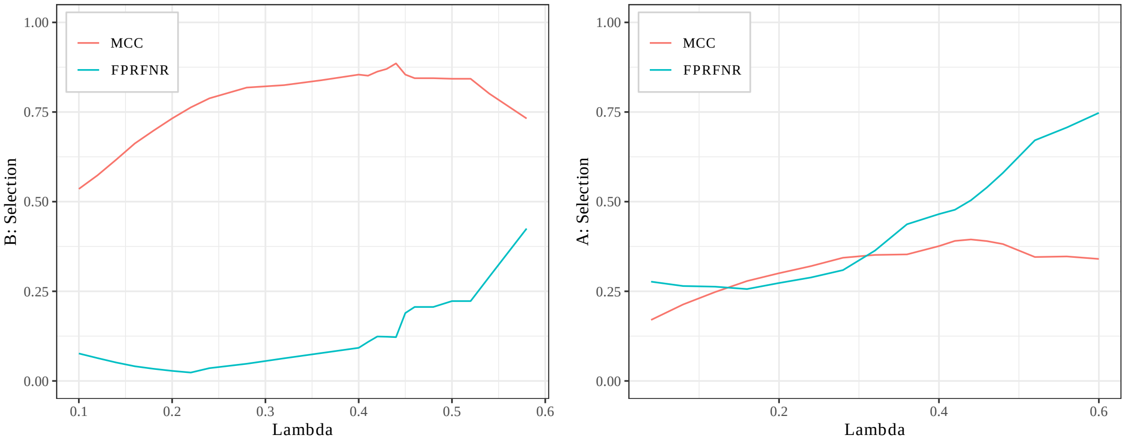
$$\widehat{\Gamma}_{ij}(B_0) = \frac{\widehat{B}_{ij}}{\sqrt{\widehat{B}_{ii}\widehat{B}_{jj}}} = \frac{\hat{S}(B)_{ij}}{\sqrt{\hat{S}(B)_{ii}\hat{S}(B)_{jj}}} \frac{\text{tr}(\tilde{\mathcal{X}}\tilde{\mathcal{X}}^T)}{\widehat{\mathcal{M}}_{ij}} \tag{3.15}$$

We show in that chapter that these are consistent estimators for the column- and row-wise correlations, respectively. Following *Viallon et al.* (2014), we then use these in the graphical Lasso objective.

$$\begin{aligned}\widehat{A}_\rho^{-1} &= \underset{A_\rho^{-1}\succ 0, \|A_\rho^{-1}\|_2 \le R_A}{\arg\min} \operatorname{tr}(\widehat{\Gamma}(A)A_\rho^{-1}) - \log|A_\rho| + \lambda |A_\rho^{-1}|_{1,\text{off}} \\ \widehat{B}_\rho^{-1} &= \underset{B_\rho^{-1}\succ 0, \|B_\rho^{-1}\|_2 \le R_B}{\arg\min} \operatorname{tr}(\widehat{\Gamma}(B)B_\rho^{-1}) - \log|B_\rho| + \lambda |B_\rho^{-1}|_{1,\text{off}}.\end{aligned} \tag{4.2}$$

Here we simple aim to provide a basic methodology for use with this data, we leave refinement and a theoretical understanding of these estimators and the Kronecker-product Ising model presented here to future work.

Figure 4.1 shows how well our estimators recover the Ising structure defined in A^{-1} and B^{-1} for a dataset without two-group mean structure. Due to the challenging nature of simulating data from the Kronecker Ising model described above, we limit the size of our simulated data for these exercises. We can see that our model recovery on the smaller B-side is quite good, while the recovery on the larger A side is more challenging here. The A-side estimator manages to capture the block structure of the graph, but there is not enough data to accurately capture individual edges.

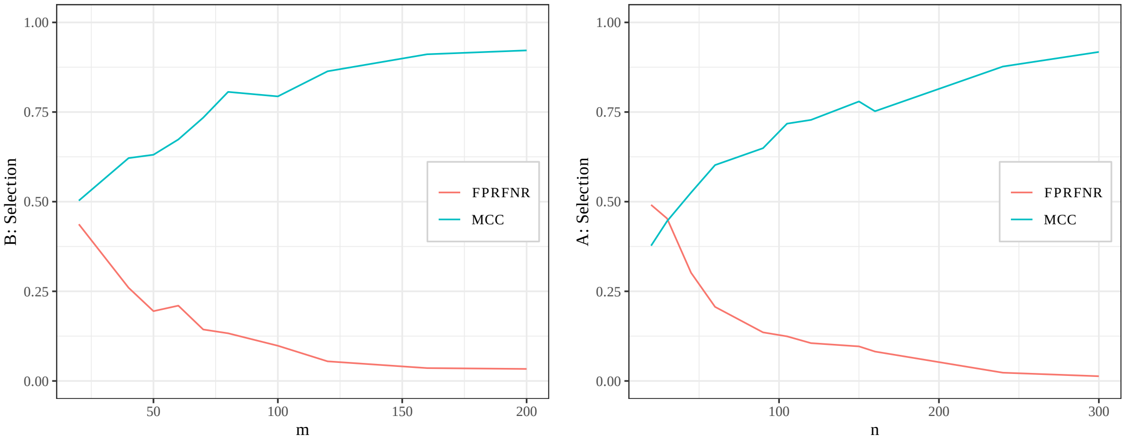


(a) B Selection

(b) A Selection

Figure 4.1: Estimator performance on simulated Ising data. We set $n = 60$, $m = 240$, and the sampling rate $\zeta = 0.95$. $A = \text{SB}(0.5, 10)$ model and $B = \text{AR}(0.7)$.

In Figure 4.2 we test the convergence behavior in terms of model selection as we fix each dimension and allow the other dimension to increase. In these smaller examples, we see that we are able to recover the graph defined by the Ising parameters quite well along either dimension once we have sufficient data.



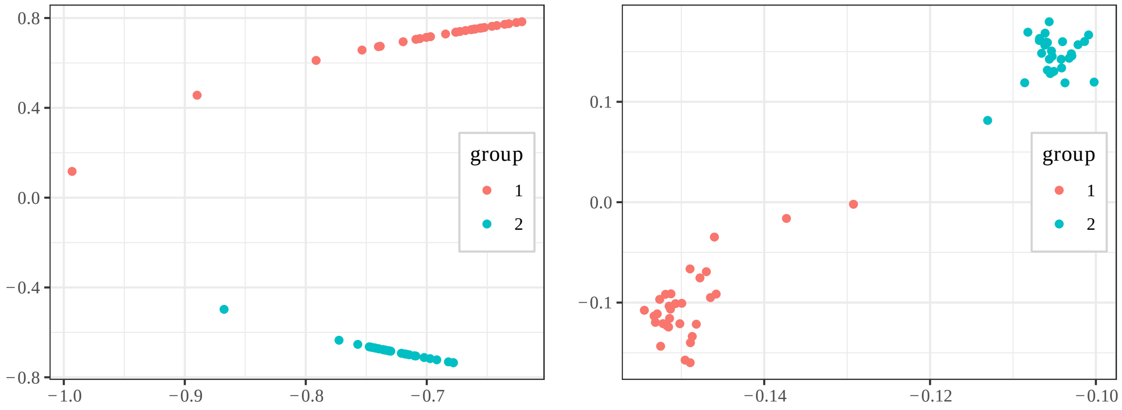
(a) B Selection, $n = 20$

(b) A Selection, $m = 30$

Figure 4.2: Estimator performance on simulated Ising data. We set the sampling rate to $\zeta = 0.95$ and let $A = \text{SB}(0.5, 10)$ model and $B = \text{AR}(0.5)$ modified to be a full loop graph. Each point is optimized over a range of penalties and represents the mean of 5 replications.

We now move on to simulations with two-group mean structure. We first validate that, even if group labels are unknown, clustering algorithms can still accurately recover the group structure. We consider two spectral clustering algorithms, the “Classify” algorithm from *Blum et al.* (2007), which is specialized for separating two populations, and the clustering algorithm from *Ng et al.* (2002) using the simple matching coefficient (SMC) as an affinity measure, which we denote as “Spectral”. Figure 4.3 shows the components and classifications from each algorithm, and both are able to perfectly recover the group mean structure.

Figure 4.4 shows example graph estimates from the two-group and globally demeaned estimators on Ising data that is simulated with two-group mean structure. The importance of two-group demeaning is evident, as global centering makes it impossible to estimate negative within-group connections due to the mean effect swamping them out.



(a) Spectral

(b) Classify

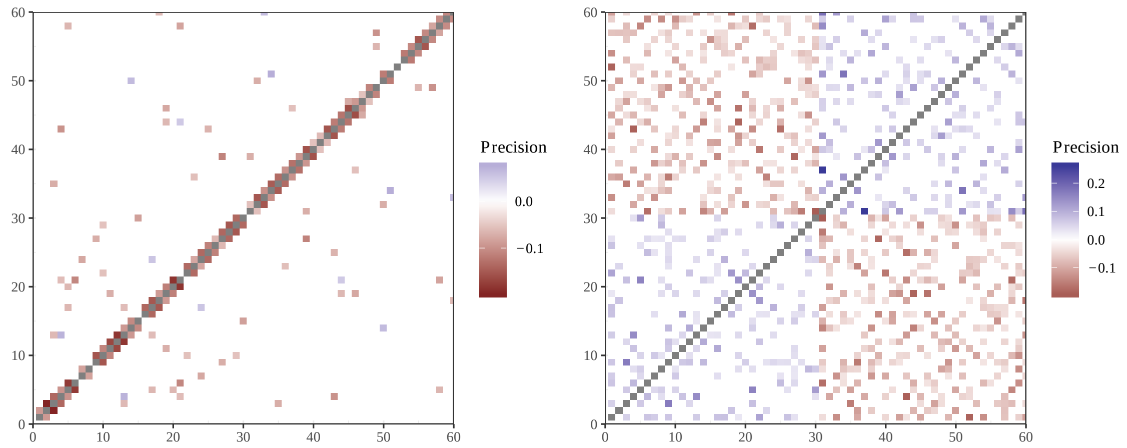
Figure 4.3: Spectral clustering components for simulated Ising data with two-group mean structure. We set $n = 60$, $m = 240$, and the sampling rate $\zeta = 0.95$. $A = \mathrm{SB}(0.5, 10)$ model and $B = \mathrm{AR}(0.7)$. μ^1, μ^2 are set so each column has group means of approximately 0.2 and 0.8.

4.2.1 Demeaning

Before we apply our graphical Lasso estimators, the two-party structure of our data suggests that the two-group methods developed in Chapter III will be applicable in this setting. Given this data, the natural mean structure to assume is separate means for each bill and party. Figure 4.5 shows the differences between the raw Democrat and Republican means by bill. The significant amount of weight away from the center for these differences in means supports the use of the two parties as group labels.

Figure 4.6 shows the pairwise agreement percentages for senators in the 115th Congress. The bimodal nature of this distribution is also consistent with the presence of two distinct groups, likely resulting from within-group and between-group pairs of senators.

To validate that the two parties (including independents as Democrats) are indeed reasonable labels to use for two-group demeaning, we compare several alternative methods for estimating the mean matrix. None of these methods are specialized for estimating the mean matrix in the presence of dependent errors $\mathbb{X}$, as we expect to have in our model, but in the independent setting these are all standard methods for estimating mean matrices $\mathbb{M}$ with two-group structure.



(a) Two-group Demeaning

(b) Global Demeaning

Figure 4.4: Performance of the B-side precision matrix estimator for simulated Ising data with two-group mean structure. We set $n = 60$, $m = 240$, and the sampling rate $\zeta = 0.95$. $A = \mathrm{SB}(0.5, 10)$ model and $B = \mathrm{AR}(-0.7)$. μ^1, μ^2 are set so each column has group means of approximately 20% and 80%. The penalty is set at 0.1 and 0.34 for the left and right panels, respectively.

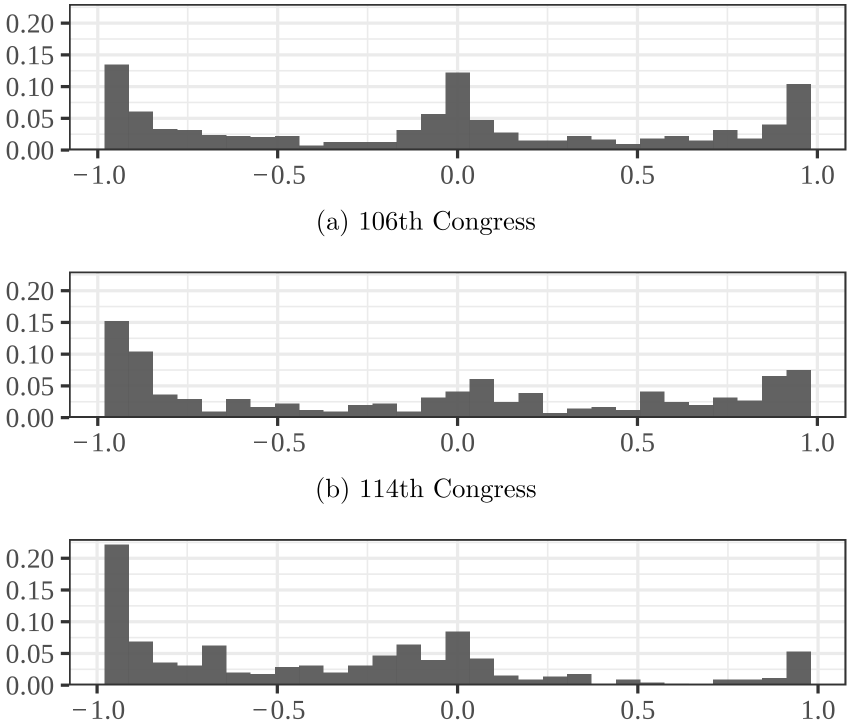
We estimate clusters using the "Classify" and "Spectral" algorithms as described in Section 4.2. Once the clusters are estimated using these methods, we use those to estimate two-group means. We also implement and test the low-rank mean estimator proposed in Section 3.1.1 for rank-2 means, denoted as "Low-rank".

Table 4.1 compares the mean estimates of these methods to the mean estimate using the original party labels. The Classify method perfectly recovers the party labels for all the Congresses, while the Spectral method does for two of the three. As expected, the Low-rank method deviates somewhat farther, since it does not make the same explicit two-group assumption on the mean, but still estimates a similar overall mean matrix.

Table 4.1: The difference in mean estimation of our alternative methods compared to using the true party labels. We present the relative Frobenius differnce $\|\widehat{\mathbb{M}} - \mathbb{M}^*\|_F / \|\mathbb{M}^*\|_F$, where $\mathbb{M}^*$ is the estimated two-group mean matrix using the true party labels.

Method	106th	114th	115th
Classify	0.00	0.00	0.00
Spectral	0.00	0.00	0.09
Low-rank	0.10	0.13	0.15

Figure 4.7 displays the eigenvalue plots from the Spectral method, confirming that using



(a) 106th Congress

(b) 114th Congress

(c) 115th Congress

Figure 4.5: Histogram of raw group mean difference by bill, plotted as Dem% - Rep%.

a two-group mean structure is appropriate.

These results show that using the two-group mean structure corresponding to the two-party structure and labels available in the data is a reasonable method of handling the mean $\mathbb{M}$, and therefore that the demeaning strategies used to form our precision matrix estimators in Chapter III are appropriate to apply here.

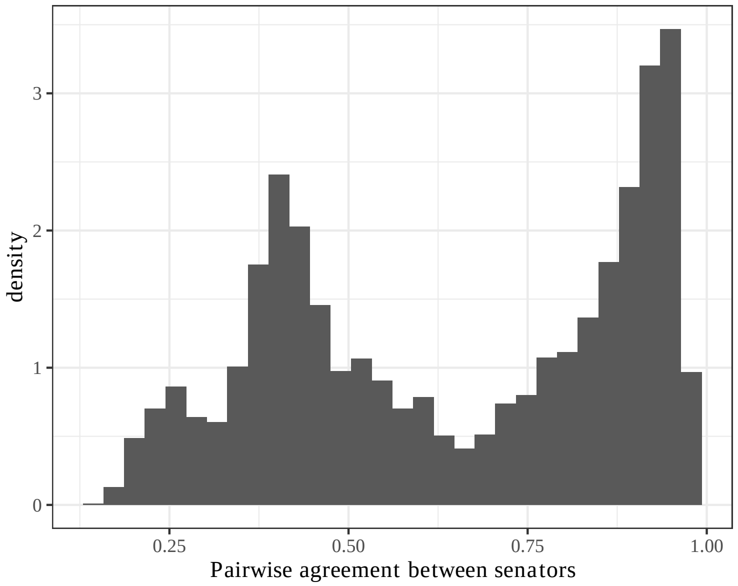
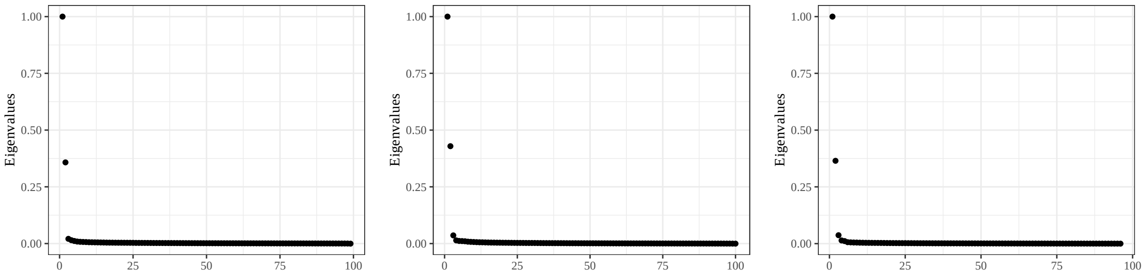


Figure 4.6: Histogram of pairwise agreements between senators for the 115th Congress.



(a) 106th

(b) 114th

(c) 115th

Figure 4.7: Eigenvalue plots of the spectral clustering algorithm “Spectral”.

4.3 Senator-side analysis

Our demeaned estimator first removes the effect of these by-party means and uses the demeaned matrix to estimate the precision matrix. By doing so, we attempt to isolate how senators are connected in terms of how they deviate from their respective party means, rather than the overall mean vote. So two senators being positively linked might mean they tend to break with their respective parties at the same time, and in the same direction. If this two-group demeaning is not performed, cross-party connections are much more difficult to discover due to the dominating effect of party means. Figure 4.8 demonstrates this phenomenon for the senator-side, comparing the results of our demeaned precision matrix estimator on the left to the non-demeaned estimator on the right. We see how the right panel is dominated by the mean structure, with separated groups and dense positive connectivity within each group, while the left panel has a wide variety of more subtle relationships. After removing the mean component as discussed above, we are now able to examine the relationships contained in the covariance of the error term without the mean effect's contamination.

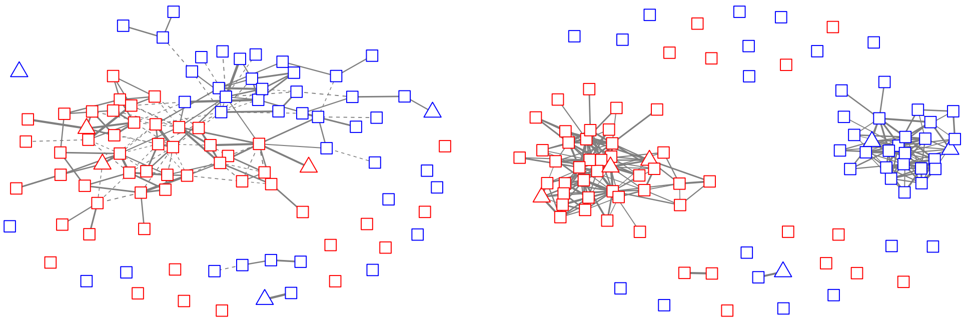


Figure 4.8: Estimated links between senators for the 106th Congress. The left panel presents our two-group demeaned estimator, while the right panel presents the estimator without demeaning. Solid lines denote positive partial correlations and dashed lines are negative. Red nodes are Republicans while Blue nodes are Democrats and Independents (caucasing with Democrats).

Figure 4.9 shows estimated graphs from the precision matrices of each of the Congresses we consider. Our data is not Gaussian, so we cannot interpret our precision matrix estimates as conditional dependency graphs, but we can still use the estimated partial correlations to understand how pairs of senators are correlated after controlling for the behavior of the rest of the Senate.

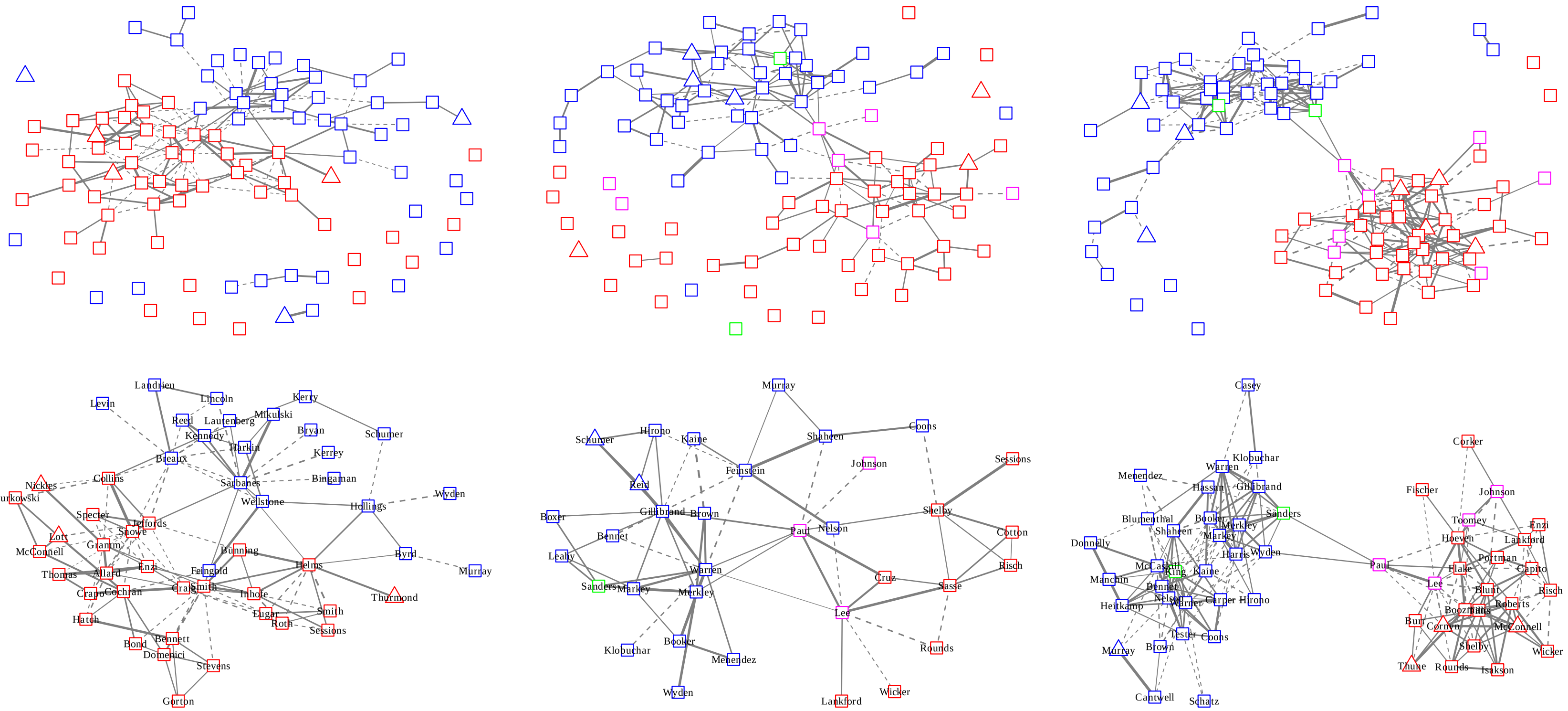
In particular, we are interested in comparing the structure of the estimated graphs from recent years (2015-2018) to a Congress from the past (here we are using the 106th, 1999-2000). It seems that the number and variety of cross-party connections has decreased significantly over time. The connections that have survived to the recent past mostly consist of connections through the "extremes" of each party, where the most liberal Democrats are connected positively to the most conservative Republicans. In the 114th and 115th Congress, these cross-party links tend to connect Tea Party-associated Republicans, some of the most conservative senators, positively to the liberal wing of the Democratic party.

The 115th Congress exhibits even fewer connections than the 114th, likely explained by the election of President Trump and the resulting political dynamic in the Senate, often described in the media as "dysfunctional" or "broken." In fact, the graph estimated at the same tuning levels does not exhibit any cross-party links (Figure 4.10), and when the penalization is relaxed until cross-party links appear they only consist of links between Rand Paul and the liberal wing of the Democratic party.

In the 106th Congress you still observe connections between the extremes of the party, such as Feingold-Smith. But we also see a significantly higher number of non-extreme-to-extreme connections in the 106th Congress. For instance, Sarbanes and Kennedy, two of the most liberal senators, are connected positively to a group of the most liberal Republicans: Collins, Jeffords, and Snowe. But Breaux, a very conservative Democrat, is negatively correlated with this same group. Or the Hollings-Byrd-Helms triangle that connects three socially conservative senators from South Carolina, West Virginia, and North Carolina, respectively, all southern east-coast states, despite Hollings and Byrd being Democrats and Helms being a Republican.

This observed change over time matches previous observations on the increasing polarization and partisanship in American politics (*Abramowitz and Saunders*, 2008; *Hare and Poole*, 2014; *Iyengar et al.*, 2019). As the parties have moved away from each other ideologically and strategically, fewer cross-party connections are observed. The remaining links likely are connecting extreme-to-extreme because of an "ends vs. the middle" phenomenon, where the extreme wings of both parties deviate away from their party consensus to vote against moderate bills for opposite reasons.

Figure 4.11 shows the estimated subgraphs for each party during these time periods. Looking first at the top row, the Democrats during the 106th and 114th Congresses feature a mostly-connected core of senators, though the 114th Congress Democrats are significantly more interconnected. But in the 115th they separate into two distinct factions, as a highly



(a) 106th Congress

(b) 114th Congress

(c) 115th Congress

Figure 4.9: Estimated links between senators for the 106th (1999-2000), 114th (2015-2016), and 115th (2017-2018) Congresses. Penalties are tuned so that the graphs have approximately the same number of edges, except for the 115th Congress where lambda is decreased until cross-party links exist. Solid lines denote positive estimated partial correlations, while dashed lines denote negative. Red nodes are Republicans while Blue nodes are Democrats and Independents (caucasing with Democrats). The top panels show the full precision matrix estimates, while the bottom are close-ups on the cross-party connections.

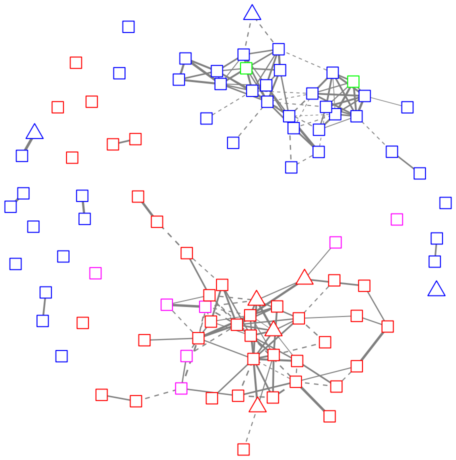
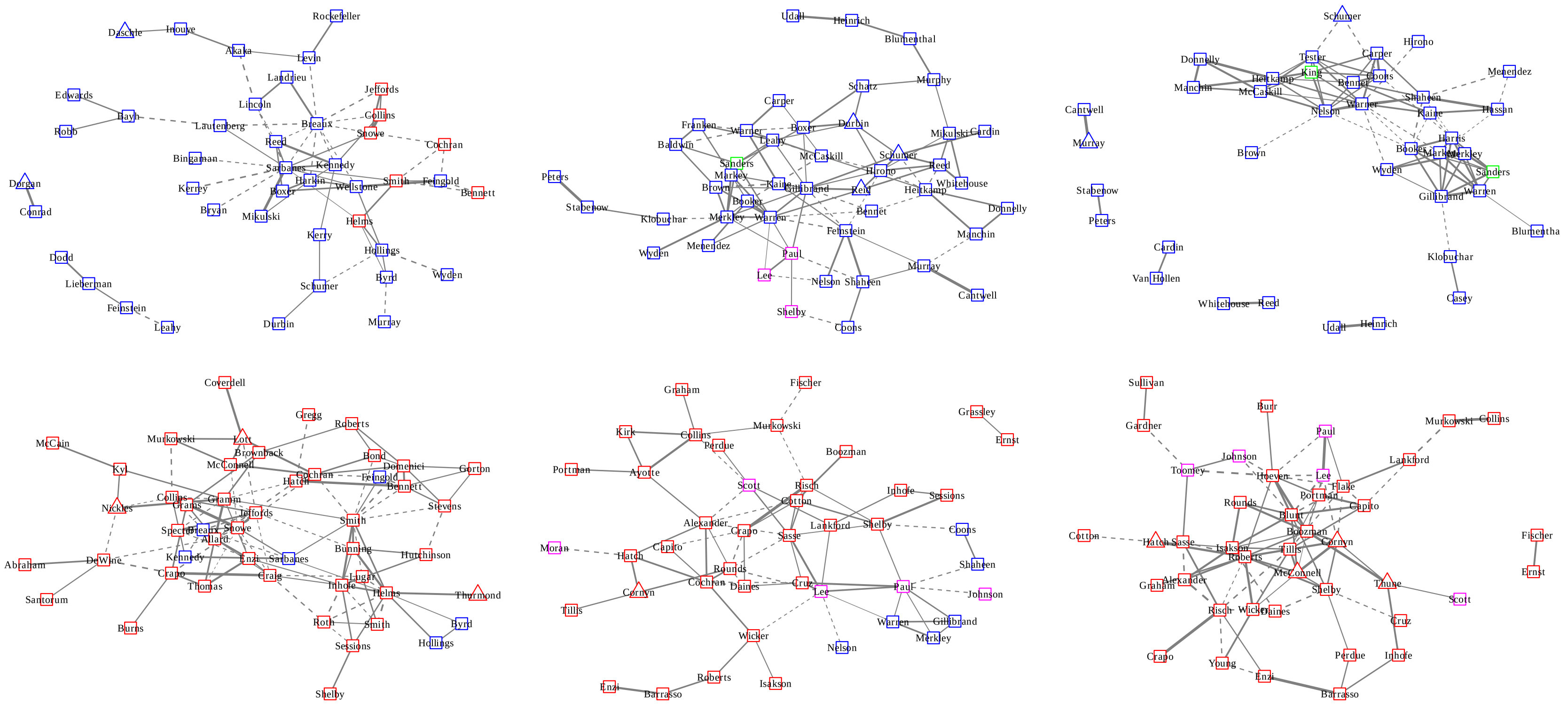


Figure 4.10: Estimated links between senators for the 115th (2017-2018) Congress. This tuning corresponds to the tuning in Figure 4.9 for the 106th and 114th Congresses, which here results in no cross-party links.

connected liberal wing of the party splits from the rest. Interestingly, this group contains all of the Democratic senators who had serious interest in presidential campaigns for 2020, such as Sanders, Warren, Gillibrand, Harris, and Booker. Perhaps all of these senators voted similarly during this time in order to match trends in public opinion and get publicity to setup their hopeful campaigns.

Looking at the Republican party, we see an opposite trend over these three Congresses. In the 106th Congress, there are several mostly-separated factions within the Republican party. In the lower-right portion of the graph is a group containing much of the conservative wing of the party, including senators like Sessions, Inhofe, Shelby, and Helms. There is also a smaller separated clique of Jeffords, Snowe, Collins and Specter (more easily visible in Figure 4.9, who are all generally known as the four most moderate Republicna senators of this time. The rest of the party is mostly in a third, larger connected component.

In the 114th Congress we see the same conservative faction, with previous senators like Session, Inhofe, and Shelby joined now by Tea Party senators like Paul and Lee, as well as other conservative-wing Republicans like Sasse and Cruz. The rest of the party is still mostly in another connected component, where the more moderate senators like Collins,



(a) 106th Congress

(b) 114th Congress

(c) 115th Congress

Figure 4.11: Estimated links between senators for the 106th (1999-2000), 114th (2015-2016), and 115th (2017-2018) Congresses by party. These show the same results as in Figure 4.9 (except the third column, which corresponds to Figure 4.10), but as by-party close-ups. Blue and red nodes are Democrats and Republicans, while pink nodes are senators associated with the Tea Party movement of 2009-2010 and green nodes denote independent senators, who during these time periods all caucused with the Democrats.

Murkowski, Kirk, and Ayotte are still identifiably grouped together.

Note that it is this conservative wing that contains all of the connections to the Democratic party, positively to liberals like Warren and Merkley, but negatively to middle or moderate senators like Coons and Nelson. This suggests that in this more partisan time, it is more likely to extremists like Paul and Lee to "cross the aisle" to vote against their party and with the Democrats, not because they agree with the Democratic position but because they think the bills proposed are not extreme enough.

In the 115th Congress, however, these Republican factions largely merge into a single connected component. Note that the Republican majority was significantly smaller during this time, as they controlled only 50-52 votes compared to the 54-55 votes they had in the 106th and 114th Congresses. This likely necessitated tighter control over the party by leadership to get their desired bills passed.

4.3.1 Stability

To test the robustness of our estimated connections to deviations in the data, we conduct a stability exercise similar to that done in *Banerjee et al.* (2008). To do this, we divide our data into 10 folds and, for each fold, we estimate the graph leaving out that fold. We then compute the average number of entries that are classified differently from the estimator using the full dataset. This provides an empirical estimate of how unstable our estimator is in terms of edge classification.

Figure 4.12 shows the estimated instability for the 106th Congress as we vary lambda. As expected, for small lambda values this instability can be quite high, as the penalization is not enough to control noise in the data. But for the penalization level we present in Figures 4.9 and 4.11, the estimated instability is quite low at 0.008, so as we perturb the data less than 1% of links disagree with the base estimate.

For the bill-side estimates, presented below, we find that the estimated graph in Figure 4.13 has an estimated instability of 0.001.

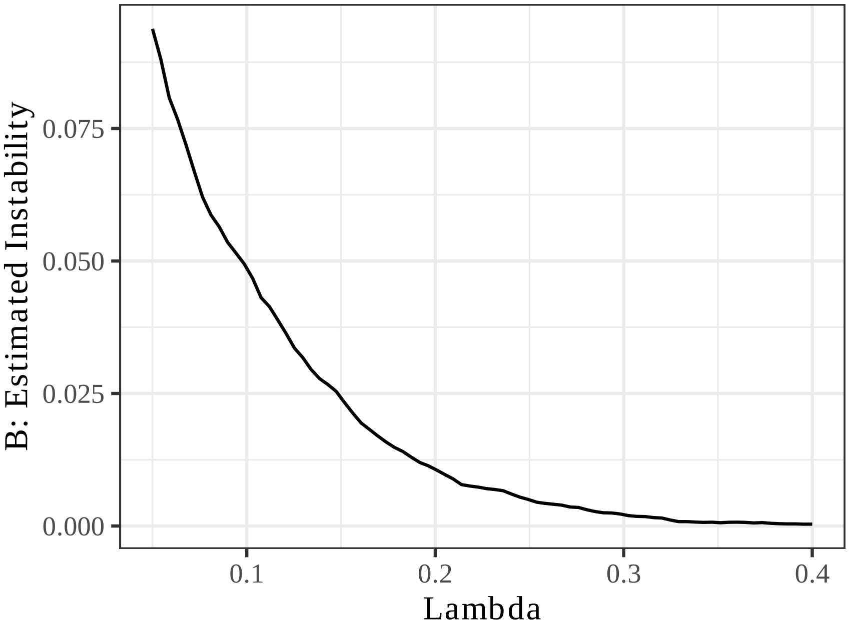


Figure 4.12: Estimated instability in the Senator-Senator graph for the 106th Congress.

4.4 Bill-side analysis

Figures 4.13 and 4.14 show the estimated vote graphs for the 106th (1999-2000) and 114th (2015-2016) Congresses. In each of the graphs, seveal of the more common vote/bill topics are highlighed.

One of the major acts of Congress during the 106th Congress was H.R. 4444, which extended permanent normal trade relations (PNTR) to China. In Figure 4.13 votes related to this bill are tightly grouped into two main groupings, one large one in the top-center of the panel and another smaller grouping in the middle-left. The middle-left group contains both the cloture motion and final passage of the bill, both of which had broad non-partisan support. This likely identifies the other two votes in this group, which were voted down similarly, as poison pill votes whose passage would've likely been tantamount to killing the bill. The unlabelled vote connected to the middle-left grouping is also interesting, as it is an unrelated vote from a year earlier regarding trade with Vietnam that was voted down in a similar matter, another similar hard-line action against trade with an east Asian country.

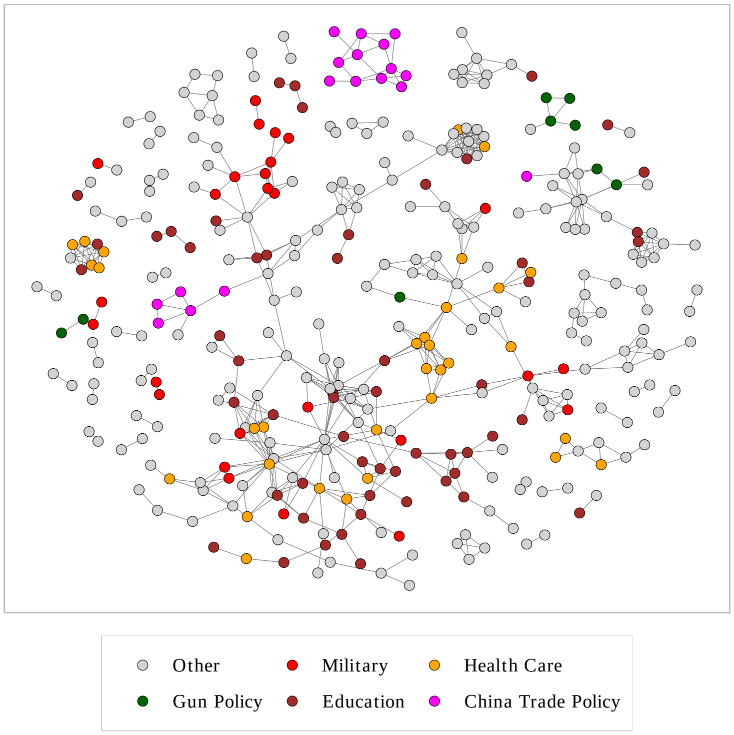


Figure 4.13: Estimated links between votes for the 106th (1999-2000) Congress.

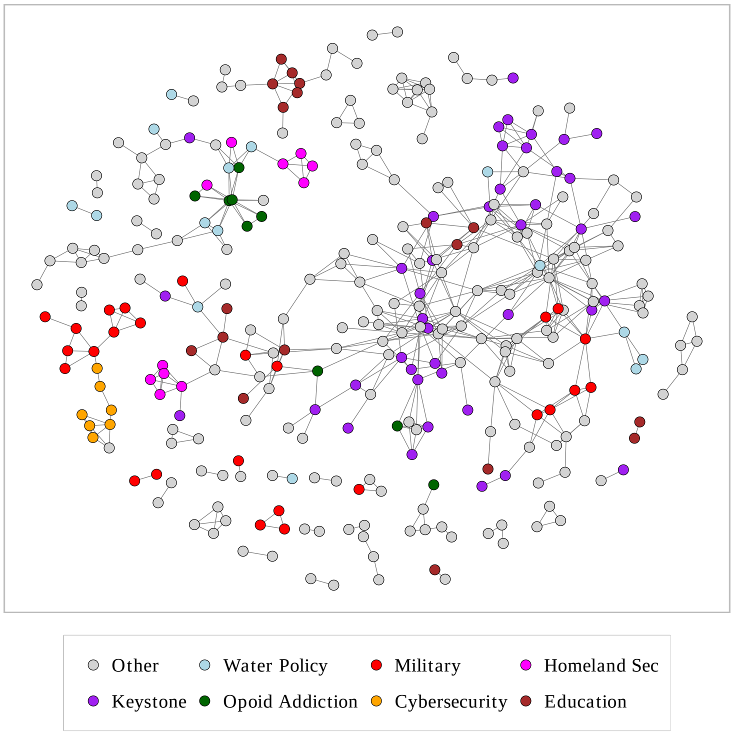
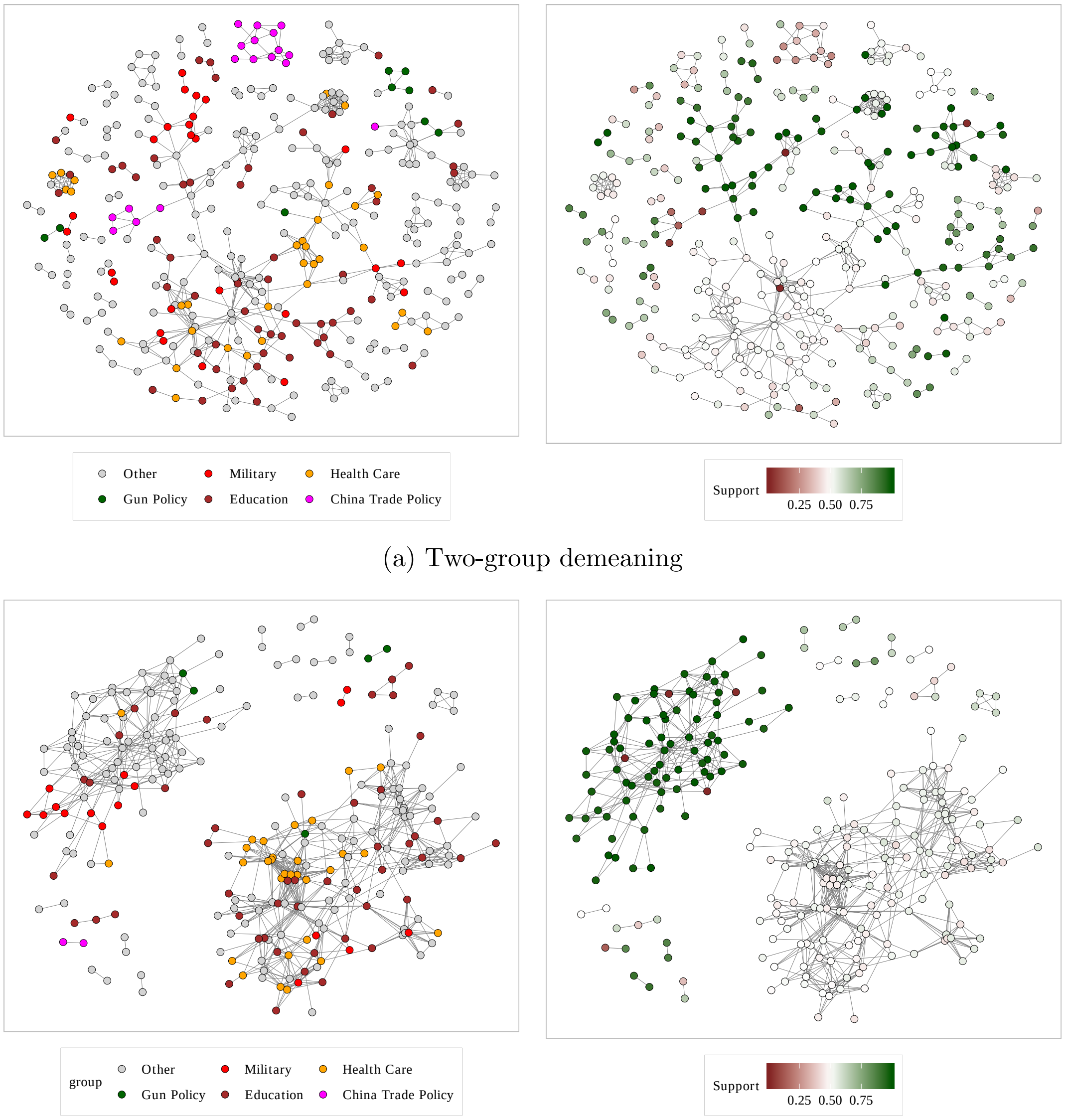


Figure 4.14: Estimated links between votes for the 114th (2015-2016) Congress.

The top-center grouping consists of votes on several proposed amendments, mostly adding smaller restrictions or conditions to the PNTR status, none of which passed but which generally induced similar sets of additional senators to vote against their party consensus on these bills.

Similar patterns and information can be found by examining other clusters as well. Looking to health care, the tighter grouping just southeast of the middle of the panel is largely regarding imposing regulations on and adding protections to health care plans, which notably don't require significant federal funding or financial commitment. But the health care bills in the bottom-middle intermixed with education and unlabelled votes involve increasing federal health care spending and moving budget resources, and therefore are closely linked to many votes involving funding for and investment in education.

On the bill-side the two-group demeaning means that bills are linked because they have similar patterns of senators that deviate from their party consensus together. Per the simulation results in Figure 3.4, we expect less issues to be caused by not demeaning here than on the senator-side, but if two-group mean structure is present not demeaning still causes a significant reduction in model selection performance. Figure 4.15 again shows the differences between the two-group demeaned and no-demeaning estimators. We see that the latter exhibits a strong separation between bills with high amounts of overall agreement and more contentious bills. Some of the structures within these groups are similar to those discussed above using the demeaned estimator, but the tight within-group connectivity and inability to connect between groups makes it much more difficult to explore more nuanced relationships.



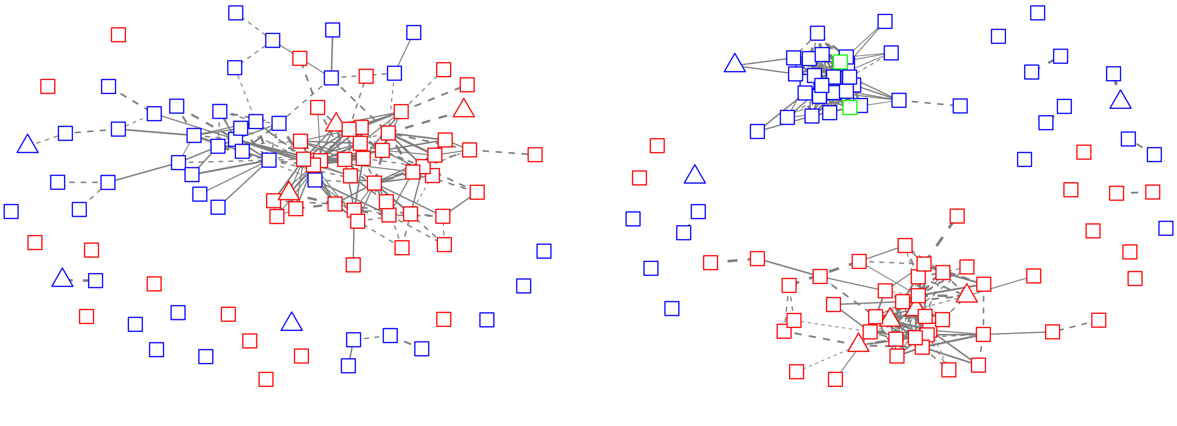
(a) Two-group demeaning

(b) No demeaning

Figure 4.15: Estimated links between votes for the 106th (1999-2000) Congress.

4.5 Covariance thresholding

Note that in addition to presenting precision matrix estimators, our work in Section 3.1 also develops correlation estimates for this two-group setting with missing data, including theoretical rates of concentration for both correlation matrices (Theorems 5.4.1 and **??**). Using these estimators, we also explore applying the covariance thresholding methods studied in *Bickel and Levina* (2008) and *Cai and Liu* (2011). Figure 4.16 shows the results of applying covariance thresholding to the senator matrix of the 106th and 115th Congresses.



(a) 106th Congress

(b) 115th Congress

Figure 4.16: Plots of thresholded correlation estimates using our demeaned input correlation estimators. These correlation estimates are then soft-thresholded by 0.248 for the left panel and 0.337 for the right.

We observe a similar pattern to the one detailed in Section 4.3, where there is significantly less between-party connectivity in the more recent Congress. In fact, we observe the same phenomenon detailed in *Mazumder and Hastie* (2012), where the pattern of connected components in the left and right panels closely matches the corresponding precision matrix estimates show in the top-left panel of Figure 4.9 and Figure 4.10, respectively. Though the correlation structure is much more tightly bound within-groups than the sparser precision matrix.

CHAPTER V

Theoretical Results for Covariance Estimation for Matrix-Variate Data with Missing Values and Mean Structure

In this chapter we prove the theoretical results presented in Chapter III. Section 5.2 first shows the convergence rate of the graphical Lasso estimator given an input correlation estimate. This is a standard result (see *Rothman et al.*, 2008; *Zhou et al.*, 2010) that we only slightly modify to account for our additional side constraint.

Section 5.3 proves consistency and convergence results for the zero-mean estimator, or when we assume oracle knowledge of the mean matrix. For this section we assume the mean matrix is zero, so $\mathbb{M} = 0$, and therefore have the data

$$\bar{\mathcal{X}} = \mathbb{U} \circ \mathbb{X} \quad \text{for } \mathbb{U} \in \{0,1\}^{n \times m}, \quad \mathbb{U}_{ij} \overset{\text{iid}}{\sim} \text{Bernoulli}(\zeta_j), \tag{3.26}$$

and define $\bar{\Gamma}_A, \bar{\Gamma}_B$ to be analogues of the estimators defined in (3.14) and (3.15), replacing $\widetilde{\mathcal{X}}$ with $\bar{\mathcal{X}}$.

The proof of Theorem 3.2.1 then bounds how far the mean estimation causes our estimates to deviate from this zero-mean baseline. Section 5.4.1 does this for the B-side estimators for when the masks are known, while Section 5.4.3 bounds the difference in the mask estimates used to prove Proposition 3.2.2. Section 5.5 proves the same bounds in the error caused by the mean estimation for the A-side.

5.1 Preliminaries

Recall that our data model is

$$
\begin{aligned}
X &= \mathbb{M} + \mathbb{X} \qquad (3.1)\\
\mathbb{X} &= B^{1/2} Z A^{1/2}
\end{aligned}
$$

For $Z \in \mathbb{R}^{m \times n}$ is a mean-zero random matrix with independent subgaussian entries with $\|Z_{ij}\|_{\psi_2} \leq K$. Then, for column-wise sampling rates $\zeta_1, \ldots, \zeta_m$, we observe

$$
\mathcal{X} = \mathbb{U} \circ (\mathbb{M} + \mathbb{X}) \quad \text{for } \mathbb{U} \in \{0,1\}^{n \times m}, \quad \mathbb{U}_{ij} \overset{\text{iid}}{\sim} \text{Bernoulli}(\zeta_j). \qquad (3.2)
$$

For some known design matrix $D \in \mathbb{R}^{n \times 2}$ of the form

$$
D = \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 \\ \underbrace{0 \quad \cdots \quad 0}_{n_1} & & & \underbrace{1 \quad \cdots \quad 1}_{n_2} & & \end{pmatrix}^T,
$$

we have $\mathbb{M} = D\mu$, where $\mu = (\mu^{(1)}, \mu^{(2)})^T \in \mathbb{R}^{2 \times m}$ is a matrix of means for each variable and group.

Let us denote

$$
\begin{aligned}
\mathbb{X} &= [x^1, x^2, \cdots, x^m] = [y^1, y^2, \cdots, y^n]^T \\
\mathbb{U} &= [u^1, u^2, \cdots, u^m] = [v^1, v^2, \cdots, v^n]^T
\end{aligned}
$$

for column vectors $x^j, u^j \in \mathbb{R}^n$ and row vectors $y^i, v^i \in \mathbb{R}^m$.

Recall that

$$\widehat{\mathcal{M}}_{k\ell} = \begin{cases} \mathrm{tr}(\mathcal{X}\mathcal{X}^T) & k = \ell \\ S_c & k \neq \ell \end{cases}. \tag{3.13}$$

where

$$S_c = \frac{n}{n-1}\,\mathrm{tr}(\mathcal{X}^T\mathcal{X} \circ \widehat{M}) - \frac{1}{n-1}\,\mathrm{tr}(\mathcal{X}^T\mathcal{X}) \tag{5.1}$$

Recall that $\phi_i(A)$ are the eigenvalues of A, where $\phi_{\min}(A)$ is the minimum eigenvalue. $\kappa(A)$ is the condition number. $a_\infty = \max_{i,j} a_{ij}$, and $a_{\min} = \min_i a_{ii}$. $\rho(A)$ is the correlation matrix corresponding to the covariance matrix A.

We also denote the sampling rates as $\zeta = (\zeta_1, \ldots, \zeta_m)$. Let $\zeta_{\min} = \min_i \zeta_i$. Let $C_1, C_2, \ldots$ be some absolute constants that may differ from line to line.

5.2 Graphical Lasso consistency

Theorem 5.2.1. *Suppose that Assumption 1 holds. Let us say that the event $\mathcal{T}(B)$ holds for an input correlation matrix $\widehat{\Gamma}(B)$ for some parameter $\delta^B_{n,m}$ if $\widehat{\Gamma}_{jj}(B) = \rho_{jj}(B_0) = 1$ for all j*

$$\max_{j,k,j\neq k} |\widehat{\Gamma}_{jk}(B) - \rho_{jk}(B)| \leq \delta_{n,m} \tag{5.2}$$

Let $\widehat{B}_\rho$ be the unique minimizer defined by (3.11) *with the input correlation matrix $\widehat{\Gamma}(B_0)$. Suppose that event $\mathcal{T}(B_0)$ holds for $\widehat{\Gamma}(B_0)$ for some $\delta^B_{n,m}$ such that*

$$\delta^B_{n,m}\sqrt{|B_0^{-1}|_{0,\mathrm{off}} \vee 1} = o(1) \tag{5.3}$$

and that for some $0 < \epsilon$ and $\varepsilon < 1$ we choose $\lambda = \delta^B_{n,m}/\epsilon$ and $R > \|\rho(B_0)^{-1}\|_2$. Then, on *event $\mathcal{T}(B_0)$, we have that for some constant C*

$$\|\widehat{B}_\rho - \rho(B_0)\|_2 \leq \|\widehat{B}_\rho - \rho(B_0)\|_F \leq C\kappa(\rho(B_0))^2\lambda\sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1}$$
$$\|\widehat{B}_\rho^{-1} - \rho(B_0)^{-1}\|_2 \leq \|\widehat{B}_\rho^{-1} - \rho(B_0)^{-1}\|_F \leq \frac{C\lambda\sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1}}{2\phi^2_{\min}(\rho(B_0))} \tag{5.4}$$

Here we focus on showing Frobenius norm consistency with an ℓ_1 penalty, but note that we can easily replace the ℓ_1 penalty in (3.11) with a nonconvex penalty such as SCAD or MCP (*Fan and Li*, 2001; *Zhang*, 2010). Using the same framework as *Loh and Wainwright* (2017), we can then show the same model selection consistency without incoherence results that they attain for the graphical Lasso. In practice, however, we find that when used with non-trivial amounts of missing data the nonconvex penalties generally perform similar to or worse than the ℓ_1 penalty, since they interact poorly with the already nonconvex objective. See *Fan et al.* (2019) for more details and performance comparisons.

Proof of Theorem 5.2.1. This proof closely follows the consistency proofs of *Rothman et al.* (2008) and *Zhou et al.* (2010), with small changes to account for our additional constraint. The lemmas and propositions within are proved in those works, so we omit their proofs here.

We first need the following lemma, which is stated and proved in *Rothman et al.* (2008) and *Zhou* (2014). Let S be an index set and $W = (w_{ij})$ be some matrix. Then we write $W_S = (w_{ij}I_{(i,j)\in S})$.

Lemma 5.2.2. *Let $\Phi_0 \succ 0$. Let $S = \{(i,j) : \ \Phi_{0ij} \neq 0, \ i \neq j\}$ and $S^c = \{(i,j) : \ \Phi_{0ij} = 0, \ i \neq j\}$. Then for all $\Delta \in \mathbb{R}^{m\times m}$, we have*

$$|\Phi_0 + \Delta|_{1,\text{off}} - |\Phi_0|_{1,\text{off}} \geq |\Delta_{S^c}|_1 - |\Delta_S|_1 \tag{5.5}$$

Moreover, we have on event $\mathcal{T}(B_0)$, for all $\Delta \in \mathbb{R}^{m\times m}$,

$$|\text{tr}\left(\Delta(\hat{\Gamma}(B_0) - \rho(B_0))\right)| \leq \delta_n|\Delta|_{1,\text{off}} = \delta_n\left(|\Delta_{S^c}|_1 + |\Delta_S|_1\right). \tag{5.6}$$

Proposition 5.2.3 is a standard result; see *Zhou et al.* (2010) for its proof.

Proposition 5.2.3. *Let B be a $p \times p$ matrix. If $B \succ 0$ and $B + D \succ 0$, then $B + vD \succ 0$ for all $v \in [0, 1]$.*

Let G_B be some convex set such that $\Phi_0 \in G_B$. Define the indicator function

$$\mathbb{1}_{G_B}(\Phi) = \begin{cases} 0 & \text{if } \Phi \in G_B \\ \infty & \text{otherwise} \end{cases}$$

So $\mathbb{1}_{G_B}(\Phi_0) = 0$ by assumption.

Let $\underline{0}$ be a matrix with all entries being zero. Let $\widehat{\Gamma}(B)$ be the input correlation matrix. Let $\lambda_n := \lambda$. For some $\Phi \succ 0$, let $\Delta := \Phi - \Phi_0$ and

$$\begin{aligned} Q(\Phi) &= \operatorname{tr}(\Phi\widehat{\Gamma}(B)) - \log|\Phi| + \lambda_n|\Phi|_{1,\text{off}} + \mathbb{1}_{G_B}(\Phi) \\ &\qquad - \operatorname{tr}(\Phi_0\widehat{\Gamma}(B)) + \log|\Phi_0| - \lambda_n|\Phi_0|_{1,\text{off}} - \mathbb{1}_{G_B}(\Phi_0) \\ &= \operatorname{tr}\left(\Delta(\widehat{\Gamma}(B) - \rho(B_0))\right) - (\log|\Phi| - \log|\Phi_0|) + \operatorname{tr}\left(\Delta\rho(B_0)\right) \\ &\qquad + \lambda_n(|\Phi|_{1,\text{off}} - |\Phi_0|_{1,\text{off}}) + \mathbb{1}_{G_B}(\Phi) \end{aligned}$$

Note that $\hat{\Phi}$ minimizes $Q(\Phi)$, or equivalently $\hat{\Delta} = \hat{\Phi} - \Phi_0$ minimizes $G(\Delta) := Q(\Phi_0 + \Delta)$. Also $G(\underline{0}) = 0$ and hence $G(\hat{\Delta}) \leq G(\underline{0}) = 0$ by definition.

Consider now the set

$$\mathcal{T}_n = \{\Delta : \Delta = \Phi - \Phi_0, \Phi, \Phi_0 \succ 0, \|\Delta\|_F = Mr_n\},$$

where by assumption

$$r_n = \delta_n\sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1} = o(1) \quad \text{and} \quad M = \frac{9}{2}\frac{1+\varepsilon}{\varepsilon}\frac{1}{\varphi^2_{\min}(\rho(B_0))}. \tag{5.7}$$

Proposition 5.2.4. *Under our assumptions, for all $\Delta \in \mathcal{T}_n$, we have by* (5.7)

$$\varphi_{\min}(\Phi_0) > 2Mr_n = o(1) \tag{5.8}$$

so that $\Phi_0 + v\Delta \succ 0, \forall v \in I \supset [0,1]$, where I is an open interval containing $[0,1]$.

Thus we have that $\log\det(\Phi_0 + v\Delta)$ is infinitely differentiable on the open interval $I \supset [0,1]$ of v. This allows us to use the Taylor's formula with an integral remainder to obtain the following lemma:

Lemma 5.2.5. *On event $\mathcal{T}(B_0)$, we have $G(\Delta) > 0$ for all $\Delta \in \mathcal{T}_n$.*

We state the following proposition, the proof of which is shown in *Zhou et al.* (2010).

Proposition 5.2.6. *If $G(\Delta) > 0, \forall \Delta \in \mathcal{T}_n$, then $G(\Delta) > 0$ for all Δ in $\mathcal{V}_n = \{\Delta : \Delta = D - \Phi_0, D \succ 0, \|\Delta\|_F > Mr_n \ \text{ for } \ r_n \ \text{ as in } (5.7)\}$. Hence if $G(\Delta) > 0, \forall \Delta \in \mathcal{T}_n$, then $G(\Delta) > 0$ for all $\Delta \in \mathcal{T}_n \cup \mathcal{V}_n$.*

By Proposition 5.2.6 and the fact that $G(\hat{\Delta}) \leq G(0) = 0$, we have the following: If $G(\Delta) > 0, \forall \Delta \in \mathcal{T}_n$, then $\hat{\Delta} \notin (\mathcal{T}_n \cup \mathcal{V}_n)$, that is, $\|\hat{\Delta}\|_F < Mr_n$, given that $\hat{\Delta} = \hat{\Phi}_n - \Phi_0$, where $\hat{\Phi}_n, \Phi_0 \succ 0$. We thus establish that $\|\Delta_{B_0}\|_F \leq Mr_n$ on the event $\mathcal{T}(B_0)$ by Lemma 5.2.5, and hence (5.4) holds on event $\mathcal{T}(B_0)$.

It remains to bound the last set of inequalities. Clearly, on event $\mathcal{T}(B_0)$, for the choice of M as in (5.7) and the bound on $\|\Delta_{B_0}\|_F < Mr_n$, we have by (5.8)

$$\begin{aligned} \|\Delta_{B_0}\|_2 \leq \|\Delta_{B_0}\|_F < Mr_n < \frac{1}{2}\phi_{\min}(\Phi_0) = \frac{1}{2}\phi_{\min}(\rho(B_0)^{-1}) \\ \text{and} \quad \phi_{\min}(\widehat{B}^{-1}) \geq \phi_{\min}(\rho(B_0)^{-1}) - \|\Delta_{B_0}\|_2 \geq \frac{1}{2}\phi_{\min}(\rho(B_0)^{-1}) \end{aligned}$$

Thus we have on $\mathcal{T}(B_0)$,

$$\|\hat{B} - \rho(B_0)\|_F \leq \frac{\|(\hat{B})^{-1} - \rho(B_0)^{-1}\|_F}{\phi_{\min}(\hat{B}^{-1})\phi_{\min}(\rho(B_0)^{-1})} \leq 9(1+\varepsilon)\kappa(\rho(B_0))^2\lambda_A\sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1}$$

Thus all statements in the theorem hold. □

5.3 Convergence for the zero-mean covariance estimates

In this section we first establish results for the zero-mean estimator, so throughout we assume $\mathbb{M} = 0$ and therefore recall that

$$\bar{\mathcal{X}} = \mathbb{U} \circ \mathbb{X} \quad \text{for } \mathbb{U} \in \{0,1\}^{n \times m}, \quad \mathbb{U}_{ij} \overset{\text{iid}}{\sim} \text{Bernoulli}(\zeta_j). \tag{3.26}$$

Let us denote

$$\begin{aligned} \mathbb{X} &= [x^1, x^2, \cdots, x^m] = [y^1, y^2, \cdots, y^n]^T \\ \mathbb{U} &= [u^1, u^2, \cdots, u^m] = [v^1, v^2, \cdots, v^n]^T \end{aligned}$$

for column vectors $x^j, u^j \in \mathbb{R}^n$ and row vectors $y^i, v^i \in \mathbb{R}^m$.

Recall that we set

$$\alpha = CK^2 \frac{\log m \|A_0\|_2}{\sum_{k=1}^m a_{kk}\zeta_k^2} + CK^2 \log^{1/2}(m \vee n) \frac{\sqrt{a_\infty}\|A_0\|_2}{a_{\min}\|\zeta\|_2} \tag{3.18}$$

$$\beta = CK^2 \frac{\log n \|B_0\|_2}{\zeta_{\min} \operatorname{tr}(B_0)} + CK^2 \log^{1/2}(m \vee n) \frac{\|B_0\|_F}{\zeta_{\min} \operatorname{tr}(B_0)} \tag{3.19}$$

for some absolute constant C.

We then prove the following bounds on our masked correlation estimates using this zero-mean data and estimator.

Theorem 5.3.1. *Consider the data generating random matrices as in* (3.26) *and suppose Assumptions 1 and 2 hold.*

Let $\widetilde{B}_\rho, \widetilde{A}_\rho$ *be the unique minimizers defined by* (3.11) *with the input correlation matrices*

$\widetilde{\Gamma}(B_0), \widetilde{\Gamma}(A_0)$ *(calculated with non-demeaned data). Penalties are chosen as*

$$\lambda_B = \frac{1}{\epsilon}(6 + 6\kappa(B_0)/\sqrt{n})\alpha \qquad \lambda_A = \frac{1}{\varepsilon}(15\beta)$$

for some $0 < \varepsilon, \epsilon < 1$*, and* $R_A > \|\rho(A_0)^{-1}\|_2$*,* $R_B > \|\rho(B_0)^{-1}\|_2$*.j Let* $\alpha, \beta < 1/3$ *and let* C_3, C_4 *be some absolute constants.*

Then, with probability at least $1 - 8/(m \vee n)^2$*,*

$$\|\widetilde{B}_\rho - \rho(B_0)\|_2 \leq \|\widetilde{B}_\rho - \rho(B_0)\|_F \leq C_3\kappa(\rho(B_0))^2\lambda_B\sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1}$$

$$\|\widetilde{B}_\rho^{-1} - \rho(B_0)^{-1}\|_2 \leq \|\widetilde{B}_\rho^{-1} - \rho(B_0)^{-1}\|_F \leq \frac{C_3\lambda_B\sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1}}{2\phi^2_{\min}(\rho(B_0))}$$

Similarly, with probability at least $1 - 8/(m \vee n)^2$*,*

$$\|\widetilde{A}_\rho - \rho(A_0)\|_2 \leq \|\widetilde{A}_\rho - \rho(A_0)\|_F \leq C_4\kappa(\rho(A_0))^2\lambda_A\sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1}$$

$$\|\widetilde{A}_\rho^{-1} - \rho(A_0)^{-1}\|_2 \leq \|\widetilde{A}_\rho^{-1} - \rho(A_0)^{-1}\|_F \leq \frac{C_4\lambda_A\sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1}}{2\phi^2_{\min}(\rho(A_0))}$$

Proof of Theorem 5.3.1. To prove this result, we simply apply Theorem 5.2.1 using the correlation convergence results in Theorem 5.3.3. □

To show this, entry-wise concentration results for the oracle correlation matrix are proved in Section 5.3.2, and the rates for the mask estimates are shown in Section 5.3.3.

5.3.1 Correlation convergence

Theorem 5.3.2. *Consider a data matrix as in* (3.26)*. Let* $m \vee n \geq 3$*. Then, on event* Λ_B *as defined in Lemma 5.3.7, for* $\alpha < 1/3$*, and* $\widetilde{\Gamma}(B_0)$ *as defined in* (3.10)

$$\forall i \neq j, \quad |\widetilde{\Gamma}_{ij}(B_0) - \rho_{ij}(B_0)| \leq \frac{\alpha}{1-\alpha} + |\rho_{ij}(B_0)|\frac{\alpha}{1-\alpha} \leq 3\alpha$$

Similarly, on event Λ_A as defined in Lemma 5.3.8, for $\beta < 1/3$, and $\widetilde{\Gamma}(A_0)$ as defined in (3.7)

$$\forall i \neq j, \quad |\widetilde{\Gamma}_{ij}(A_0) - \rho_{ij}(A_0)| \leq \frac{\beta}{1-\beta} + |\rho_{ij}(A_0)| \frac{\beta}{1-\beta} \leq 3\beta$$

Proof. Under Λ_B,

$$\begin{aligned}
|\widetilde{\Gamma}_{ij}(B_0) - \rho_{ij}(B_0)| &= \left| \frac{\langle v^i \circ y^i, v^j \circ y^j \rangle / \mathcal{M}_{ij}}{(\|v^i \circ y^i\|_2 / \sqrt{\mathcal{M}_{ii}})(\|v^j \circ y^j\|_2 / \sqrt{\mathcal{M}_{jj}})} - \rho_{ij}(B_0) \right| \\
&\leq \left| \frac{\langle v^i \circ y^i, v^j \circ y^j \rangle / (\sqrt{b_{ii} b_{jj}} \mathcal{M}_{ij}) - \rho_{ij}(B_0)}{(\|v^i \circ y^i\|_2 / \sqrt{b_{ii} \mathcal{M}_{ii}})(\|v^j \circ y^j\|_2 / \sqrt{b_{jj} \mathcal{M}_{jj}})} \right| \\
&\quad + |\rho_{ij}(B_0)| \left| \frac{1}{(\|v^i \circ y^i\|_2 / \sqrt{b_{ii} \mathcal{M}_{ii}})(\|v^j \circ y^j\|_2 / \sqrt{b_{jj} \mathcal{M}_{jj}})} - 1 \right| \\
&\leq \frac{\alpha}{1-\alpha} + |\rho_{ij}(B_0)| \frac{\alpha}{1-\alpha}
\end{aligned}$$

The proof of the second statement is identical. □

Theorem 5.3.3. *Consider a data matrix as in* (3.26)*. Let $m \vee n \geq 3$. Then, with probability at least $1 - 8/(n \vee m)^2$, for $\alpha < 1/3$, and $\widehat{\Gamma}(B_0)$ as defined in* (3.15)

$$\forall i \neq j, \quad |\widehat{\Gamma}_{ij}(B_0) - \rho_{ij}(B_0)| \leq 6\alpha + (6\kappa(B_0)/\sqrt{n})\alpha$$

Similarly, with probability at least $1 - 8/(n \vee m)^2$, for $\beta < 1/3$, and $\widehat{\Gamma}(A_0)$ as defined in (3.14)

$$\forall i \neq j, \quad |\widehat{\Gamma}_{ij}(A_0) - \rho_{ij}(A_0)| \leq 15\beta$$

Proof of Theorem 5.3.3. For the B-side, consider the events Λ_B as defined in Lemma 5.3.7 and Λ_M as defined in Proposition 5.3.11. Under Λ_M,

$$\left| \frac{S_c}{ES_c} - 1 \right| \leq \alpha, \qquad \left| \frac{\mathrm{tr}(\mathcal{X}\mathcal{X}^T)}{E\,\mathrm{tr}(\mathcal{X}\mathcal{X}^T)} - 1 \right| \leq \frac{\alpha}{\sqrt{n}} \kappa(B_0).$$

Also note that,

$$\widehat{\Gamma}_{ij}(B_0) = \frac{\langle v^i \circ y^i, v^j \circ y^j \rangle}{\|v^i \circ y^i\|_2 \|v^j \circ y^j\|_2} \frac{\operatorname{tr}(\mathcal{X}\mathcal{X}^T)}{S_c} = \widetilde{\Gamma}_{ij}(B_0) \frac{\operatorname{tr}(\mathcal{X}\mathcal{X}^T)}{E\operatorname{tr}(\mathcal{X}\mathcal{X}^T)} \frac{ES_c}{S_c}$$

So, using Theorem 5.3.2,

$$\begin{aligned}
|\widehat{\Gamma}_{ij}(B_0) - \widetilde{\Gamma}_{ij}(B_0)| &\le |\widetilde{\Gamma}_{ij}(B_0)| \left| \frac{\operatorname{tr}(\mathcal{X}\mathcal{X}^T)}{E\operatorname{tr}(\mathcal{X}\mathcal{X}^T)} \frac{ES_c}{S_c} - 1 \right| \\
&\le |\widetilde{\Gamma}_{ij}(B_0)| \max\left(\left| \frac{1 + \alpha\kappa(B_0)/\sqrt{n}}{1-\alpha} - 1 \right|, \left| \frac{1 - \alpha\kappa(B_0)/\sqrt{n}}{1+\alpha} - 1 \right| \right) \\
&\le |\widetilde{\Gamma}_{ij}(B_0)| \max\left(1 + \frac{3}{2}\alpha + \frac{3\kappa(B_0)}{\sqrt{n}}\alpha - 1, 1 - \left(1 - \alpha - \frac{\kappa(B_0)}{\sqrt{n}}\alpha \right) \right) \\
&\le (1+3\alpha)\left(\frac{3}{2} + \frac{3\kappa(B_0)}{\sqrt{n}} \right)\alpha = \left(\frac{3}{2} + \frac{3\kappa(B_0)}{\sqrt{n}} \right)\alpha + \left(\frac{9}{2} + \frac{9\kappa(B_0)}{\sqrt{n}} \right)\alpha^2
\end{aligned}$$

and therefore, when $\alpha < 1/3$,

$$\begin{aligned}
|\widehat{\Gamma}_{ij}(B_0) - \rho_{ij}(B_0)| &\le |\widehat{\Gamma}_{ij}(B_0) - \widetilde{\Gamma}_{ij}(B_0)| + |\widetilde{\Gamma}_{ij}(B_0) - \rho_{ij}(B_0)| \\
&\le 6\alpha + (6\kappa(B_0)/\sqrt{n})\alpha
\end{aligned}$$

For the A-side, note that

$$\begin{aligned}
|\widehat{\Gamma}_{ij}(A_0) - \widetilde{\Gamma}_{ij}(A_0)| &= \left| \widetilde{\Gamma}_{ij}(A_0) \frac{\sqrt{\zeta_i \zeta_j}}{\sqrt{\hat{\zeta}_i \hat{\zeta}_j}} - \widetilde{\Gamma}_{ij}(A_0) \right| \\
&= \widetilde{\Gamma}_{ij}(A_0) \left| \frac{\sqrt{\zeta_i \zeta_j}}{\sqrt{\hat{\zeta}_i \hat{\zeta}_j}} - 1 \right|
\end{aligned}$$

Using Hoeffding's Inequality, we get that

$$\begin{aligned}
& P\left(|\hat{\zeta}_i - \zeta_i| \le \sqrt{3/2}\sqrt{\frac{\log m \vee n}{n}} \right) \ge 1 - 2/(m \vee n)^3 \\
\implies & P\left(|\hat{\zeta}_i - \zeta_i| \le \sqrt{3/2}\sqrt{\frac{\log m \vee n}{n}} \ \forall\ i \right) \ge 1 - 2m/(m \vee n)^3
\end{aligned}$$

So, with probability at least $1 - 2/(m \vee n)^2$,

$$|\hat{\zeta}_i - \zeta_i| \le \beta\zeta_{\min} \ \forall \ i = 1, \ldots, m$$

We assume that $\zeta_{\min} \gtrsim \sqrt{\frac{\log m \vee n}{n}}$. So, for appropriately chosen constants,

$$|\hat{\zeta}_i - \zeta_i| \le \sqrt{3/2}\sqrt{\frac{\log m \vee n}{n}} \implies \hat{\zeta}_i \ge \frac{\zeta_{\min}}{2}$$

Therefore

$$\left|\frac{\zeta_i}{\hat{\zeta}_i} - 1\right| = \frac{|\zeta_i - \hat{\zeta}_i|}{\hat{\zeta}_i} \le 2\beta$$

And, for $\beta < 1$,

$$\begin{aligned}\left|\frac{\zeta_i\zeta_j}{\hat{\zeta}_i\hat{\zeta}_j} - 1\right| &= \left|\left(\frac{\zeta_i}{\hat{\zeta}_i} - 1\right)\left(\frac{\zeta_j}{\hat{\zeta}_j} - 1\right) + \frac{\zeta_i}{\hat{\zeta}_i} - 1 + \frac{\zeta_j}{\hat{\zeta}_j} - 1\right| \\ &\le \left|\frac{\zeta_i}{\hat{\zeta}_i} - 1\right|\left|\frac{\zeta_j}{\hat{\zeta}_j} - 1\right| + \left|\frac{\zeta_i}{\hat{\zeta}_i} - 1\right| + \left|\frac{\zeta_j}{\hat{\zeta}_j} - 1\right| \le 6\beta\end{aligned}$$

Finally, we note that, for any a and $\hat{a}$ and $\varepsilon \in [0, 1]$,

$$\begin{aligned}\left|\frac{a}{\hat{a}} - 1\right| \le \varepsilon &\implies \sqrt{1-\varepsilon} \le \frac{\sqrt{a}}{\sqrt{\hat{a}}} \le \sqrt{1+\varepsilon} \\ &\implies \left|\frac{\sqrt{a}}{\sqrt{\hat{a}}} - 1\right| \le \max(1 - \sqrt{1-\varepsilon}, \sqrt{1+\varepsilon} - 1) \le \varepsilon\end{aligned}$$

So we get that

$$\left|\frac{\sqrt{\zeta_i\zeta_j}}{\sqrt{\hat{\zeta}_i\hat{\zeta}_j}} - 1\right| \le 6\beta \tag{5.9}$$

And therefore, using Theorem 5.3.2, we get that, for $\beta < 1/3$ and with probability at least $1 - 8/(n \vee m)^2$,

$$\begin{aligned}
|\widehat{\Gamma}_{ij}(A_0) - \rho_{ij}(A_0)| &\leq |\widehat{\Gamma}_{ij}(A_0) - \widetilde{\Gamma}_{ij}(A_0)| + |\widetilde{\Gamma}_{ij}(A_0) - \rho_{ij}(A_0)| \\
&\leq \widetilde{\Gamma}_{ij}(A_0) \left| \frac{\sqrt{\zeta_i \zeta_j}}{\sqrt{\hat{\zeta}_i \hat{\zeta}_j}} - 1 \right| + 3\beta \\
&\leq (1 + 3\beta) 6\beta + 3\beta \\
&= 15\beta
\end{aligned}$$

□

5.3.2 Concentration results

We use the following results from *Zhou* (2019),

Theorem 5.3.4 (Theorem 1.2 in *Zhou*, 2019)**.** *Let $X = (X_1, \ldots, X_m) \in \mathbb{R}^m$ be a random vector with independent components X_i satisfying $EX_i = 0$ and $\|X_i\|_{\psi_2} \leq K$. Let $\xi = (\xi_1, \ldots, \xi_m) \in \{0, 1\}^m$ be independent of X with independent Bernoulli components ξ_i such that $E\xi_i = p_i$. Let D_ξ be the diagonal matrix* diag(ξ)*. Let $D_0 \in \mathbb{R}^{m \times m}$ be a symmetric matrix, and let $A_0 = (a_{ij}) = D_0^2$. Define $Y = D_0 X$.*

Then for any $t > 0$,

$$P(|Y^T D_\xi Y - E Y^T D_\xi Y| > t) \leq 2 \exp\left(-c_2 \min\left(\frac{t^2}{K^4 (\sum_{i=1}^m p_i a_{ii}^2 + \sum_{i \neq j} a_{ij}^2 p_i p_j)}, \frac{t}{K^2 \|A_0\|_2} \right) \right)$$

for some absolute constant c_2.

Theorem 5.3.5 (Theorem 1.3 in *Zhou*, 2019)**.** *Consider the same setting as in Theorem 5.3.4. Additionally, let $X' \sim X$ be an independent but identically distributed copy of X that is also independent of ξ. Define $Y = D_0 X$ and $Y' = D_0 X'$.*

Then for any $t > 0$,

$$P(|Y^T D_\xi Y' - EY^T D_\xi Y'| > t) \le 2 \exp\left(-c_2 \min\left(\frac{t^2}{K^4(\sum_{i=1}^m p_i a_{ii}^2 + \sum_{i \neq j} a_{ij}^2 p_i p_j)}, \frac{t}{K^2 \|A_0\|_2}\right)\right).$$

Note that when we incorporate mean estimation we will also use variants of these theorems where D_0 is not necessarily symmetric and $A_0 = D_0^T D_0$. A close examination of the proofs in *Zhou* (2019) shows that these results still hold in this case.

These results allow us to obtain the following concentration bound results on the individual covariance entries, which we prove later in this section. Note that these results were first presented in *Zhou* (2019), but we include the full suite of results and proofs here for completeness.

Theorem 5.3.6. *Consider the generating model in* (3.26). *We first consider the B-side correlatione stimation. For $t > 0$, for each i,*

$$\begin{aligned} &P\left(\frac{1}{b_{ii}}\left|\langle v^i \circ y^i, v^i \circ y^i \rangle - \sum_{k=1}^m a_{kk}\zeta_k\right| > t\right) \\ &\quad \le 2\exp\left(-c_2 \min\left(\frac{t^2}{4K^4(\sum_{k=1}^m a_{kk}^2 \zeta_k + \sum_{k \neq \ell} a_{k\ell}^2 \zeta_k \zeta_\ell)}, \frac{t}{2K^2\|A_0\|_2}\right)\right) \end{aligned}$$

and for each $i \neq j$,

$$\begin{aligned} &P\left(\left|\frac{\langle v^i \circ y^i, v^j \circ y^j \rangle}{\sqrt{b_{ii} b_{jj}}} - \rho_{ij}(B_0)\sum_{k=1}^m a_{kk}\zeta_k^2\right| > t\right) \\ &\quad \le 6\exp\left(-c_2 \min\left(\frac{t^2}{4K^4(\sum_{k=1}^m a_{kk}^2 \zeta_k^2 + \sum_{k \neq \ell} a_{k\ell}^2 \zeta_k^2 \zeta_\ell^2)}, \frac{t}{2K^2\|A_0\|_2}\right)\right) \end{aligned}$$

Similarly, for estimating the A-side correlations, we get that

$$\begin{aligned} &P\left(\frac{1}{a_{ii}}\left|\langle u^i \circ x^i, u^i \circ x^i \rangle - \zeta_i \operatorname{tr}(B_0)\right| > t\right) \\ &\quad \le 2\exp\left(-c_2 \min\left(\frac{t^2}{4K^4 \zeta_i \|B_0\|_F}, \frac{t}{2K^2\|B_0\|_2}\right)\right) \end{aligned}$$

and for each $i \neq j$,

$$P\left(\left|\frac{\langle u^i \circ x^i, u^j \circ x^j\rangle}{\sqrt{a_{ii}a_{jj}}} - \rho_{ij}(A_0)\zeta_i\zeta_j \operatorname{tr}(B_0)\right| > t\right) \\ \leq 6\exp\left(-c_2 \min\left(\frac{t^2}{4K^4\zeta_i\zeta_j\|B_0\|_F}, \frac{t}{2K^2\|B_0\|_2}\right)\right)$$

The following lemma combines these concentration results into the appropriate ℓ_∞ norms that we need.

Lemma 5.3.7. *Consider data generating model as in* (3.26). *Let $\mathcal{M}_{ii} = \sum_{k=1}^m a_{kk}\zeta_k$ and $\mathcal{M}_{ij} = \sum_{k=1}^m a_{kk}\zeta_k^2$ as defined in* (3.9). *Then, denote by Λ_B the event that the following inequalities hold simultaneously for $j = 1, \ldots, n$,*

$$\begin{aligned}\left|\frac{\|v^j \circ y^j\|_2^2}{b_{jj}\mathcal{M}_{jj}} - 1\right| &\leq CK^2 \frac{\log m \|A_0\|_2}{\sum_{k=1}^m a_{kk}\zeta_k} + CK^2 \log^{1/2}(m \vee n) \frac{\sqrt{\sum_{k=1}^m a_{kk}^2\zeta_k + \sum_{k\neq\ell} a_{k\ell}^2\zeta_k\zeta_\ell}}{\sum_{k=1}^m a_{kk}\zeta_k} \\ &\leq CK^2 \frac{\log m \|A_0\|_2}{\sum_{k=1}^m a_{kk}\zeta_k} + CK^2 \log^{1/2}(m \vee n) \frac{\sqrt{\sum_{k=1}^m a_{kk}^2\zeta_k + \sum_{k\neq\ell} a_{k\ell}^2\sqrt{\zeta_k\zeta_\ell}}}{\sum_{k=1}^m a_{kk}\zeta_k} \\ &=: \alpha_{\text{diag}}\end{aligned} \tag{5.10}$$

and for each $i \neq j$,

$$\begin{aligned}&\left|\frac{\langle v^j \circ y^j, v^i \circ y^i\rangle}{\sqrt{b_{ii}b_{jj}}\mathcal{M}_{ij}} - \rho_{ij}(B_0)\right| \\ &\leq CK^2 \frac{\log m \|A_0\|_2}{\sum_{k=1}^m a_{kk}\zeta_k^2} + CK^2 \log^{1/2}(m \vee n) \frac{\sqrt{\sum_{k=1}^m a_{kk}^2\zeta_k^2 + \sum_{k\neq\ell} a_{k\ell}^2\zeta_k^2\zeta_\ell^2}}{\sum_{k=1}^m a_{kk}\zeta_k^2} \\ &\leq CK^2 \frac{\log m \|A_0\|_2}{\sum_{k=1}^m a_{kk}\zeta_k^2} + CK^2 \log^{1/2}(m \vee n) \frac{\sqrt{\sum_{k=1}^m a_{kk}^2\zeta_k^2 + \sum_{k\neq\ell} a_{k\ell}^2\zeta_k\zeta_\ell}}{\sum_{k=1}^m a_{kk}\zeta_k^2} =: \alpha_{\text{offd}} \quad (5.11) \\ &\leq CK^2 \frac{\log m \|A_0\|_2}{\sum_{k=1}^m a_{kk}\zeta_k^2} + CK^2 \log^{1/2}(m \vee n) \frac{\sqrt{a_\infty \|A_0\|_2}}{a_{\min}\sqrt{m}} \frac{\sqrt{m}}{\|\zeta\|_2}\end{aligned}$$

Then, we get that $P(\Lambda_B) \geq 1 - \frac{6}{(m\vee n)^2}$, where C are some absolute constants chosen so that the probability holds.

Lemma 5.3.8. *Consider data generating model as in (3.26). Let* $\mathcal{N}_{jj} = \operatorname{tr}(B_0)\zeta_j$ *for all* j *and* $\mathcal{N}_{ij} = \zeta_i\zeta_j \operatorname{tr}(B_0)$ *for all* $i \neq j$*. Denote by* Λ_A *the event that the following two inequalities hold simultaneously for* $j = 1, \dots, m$,

$$\left| \frac{\|u^j \circ x^j\|_2^2}{a_{jj}\mathcal{N}_{jj}} - 1 \right| \leq CK^2 \frac{\log n \|B_0\|_2}{\zeta_j \operatorname{tr}(B_0)} + CK^2 \log^{1/2}(m \vee n) \frac{\|B_0\|_F}{\operatorname{tr}(B_0)\zeta_j^{1/2}} =: \beta_j \tag{5.12}$$

and for all $i \neq j$,

$$\left| \frac{\langle u^i \circ x^i, u^j \circ x^j \rangle}{\sqrt{a_{ii}a_{jj}}\mathcal{N}_{ij}} - \rho_{ij}(A_0) \right| \leq CK^2 \frac{\log n \|B_0\|_2}{\zeta_i\zeta_j \operatorname{tr}(B_0)} + CK^2 \log^{1/2}(m \vee n) \frac{\|B_0\|_F}{\sqrt{\zeta_i\zeta_j} \operatorname{tr}(B_0)} =: \beta_{ij} \tag{5.13}$$

Then, we get that $\mathbb{P}(\Lambda_A) \geq 1 - \frac{6}{(m \vee n)^2}$*, where* C *is some absolute constant chosen so that the probability holds.*

Proof of Lemmas 5.3.7 and 5.3.8. We use the first result in Theorem 5.3.6 with

$$t = CK^2 \log m \|A_0\|_2 + CK^2 \log^{1/2}(m \vee n) \sqrt{\sum_{k=1}^m a_{kk}^2 \zeta_k + \sum_{k \neq \ell} a_{k\ell}\zeta_k\zeta_\ell}.$$

Using a union bound, we therefore get that

$$P\left(\frac{1}{b_{ii}} \left| \langle v^i \circ y^i, v^i \circ y^i \rangle - \sum_{k=1}^m a_{kk}\zeta_k \right| < t \ \forall\ i = 1, \dots, n \right) \geq 1 - 2\frac{n}{m^4}$$

Similarly, we use the second result with

$$t = CK^2 \log m \|A_0\|_2 + CK^2 \log^{1/2}(m \vee n) \sqrt{\sum_{k=1}^m a_{kk}^2 \zeta_k^2 + \sum_{k \neq \ell} a_{k\ell}\zeta_k^2\zeta_\ell^2}$$

and a union bound to get that

$$P\left(\left| \frac{\langle v^i \circ y^i, v^j \circ y^j \rangle}{\sqrt{b_{ii}b_{jj}}} - \rho_{ij}(B_0) \sum_{k=1}^m a_{kk}\zeta_k^2 \right| < t \ \forall\ i \neq j \right) \geq 1 - 6\frac{n(n-1)}{m^4}$$

The proof of Lemma 5.3.8 proceeds similarly with the third and fourth results in Theorem 5.3.6. □

Proof of Theorem 5.3.6. ***B*-side.** We first consider the diagonal case. Note that y^i has covariance $b_{ii}A_0$, so we can write $y^i = (b_{ii}A_0)^{1/2}(g_1, \dots, g_m)^T$ for $g_1, \dots, g_m \overset{\text{i.i.d}}{\sim} Z$, i.e. entries of g are subgaussian with a subgaussian constant K, mean zero, and unit variance.

We, for $D = \operatorname{diag}(v^i)$, can then show that

$$\begin{aligned} \langle v^i \circ y^i, v^i \circ y^i \rangle &= \sum_{k=1}^{m} \sum_{\ell=1}^{m} D_{k\ell} y^i_\ell y^i_k \\ &= g b_{ii} A_0^{1/2} D A_0^{1/2} g^T \end{aligned}$$

Noting that $ED = \operatorname{diag}(\zeta_1, \dots, \zeta_m)$, we can then apply Theorem 5.3.4 to the quadratic form

$$g A_0^{1/2} D A_0^{1/2} g^T - E g A_0^{1/2} D A_0^{1/2} g^T$$

to get our first inequality.

For the off-diagonal case, we first need the following proposition.

Proposition 5.3.9. *Let $B_{0,(i,j)} = (b_{ij})_{i,j=1}^2 \in \mathbb{R}^{2\times 2}$ be the positive definite submatrix of B_0 with rows and columns i, j. Denote it's unique symmetric square root as*

$$\begin{pmatrix} c_{ii} & c_{ij} \\ c_{ij} & c_{jj} \end{pmatrix}$$

Define

$$D'(i,j) = \begin{pmatrix} c_{ii}c_{ij} & c_{ii}c_{jj} \\ c_{ij}c_{ij} & c_{ij}c_{jj} \end{pmatrix} \otimes A_0^{1/2} D A_0^{1/2}$$

Then

$$\|D'(i,j)\|_2 \leq \sqrt{b_{ii}b_{jj}}\|A_0^{1/2}DA_0^{1/2}\|_2 \tag{5.14}$$

$$\|D'(i,j)\|_F \leq \sqrt{b_{ii}b_{jj}}\|A_0^{1/2}DA_0^{1/2}\|_F \tag{5.15}$$

$$\tag{5.16}$$

And, recalling that $\rho_{ij}(B_0) = b_{ij}/\sqrt{b_{ii}b_{jj}}$,

$$\left|\rho_{ij}(B_0)\frac{c_{ii}c_{jj}+c_{ij}^2}{b_{ij}}\right| < 1 \tag{5.17}$$

Without loss of generality, let $i = 1, j = 2$. Then we concatenate the vectors y^1, y^2 to form $(y^1, y^2) \in \mathbb{R}^{2m}$ with covariance matrix $B_{0,(1,2)} \otimes A_0$. Defining $g_1, \ldots, g_{2m} \overset{\text{i.i.d}}{\sim} Z$ as we did above, we get that

$$(y^1, y^2) = B_{0,(1,2)}^{1/2} \otimes A_0^{1/2} g^T = \begin{pmatrix} c_{11}A_0^{1/2} & c_{12}A_0^{1/2} \\ c_{12}A_0^{1/2} & c_{22}A_0^{1/2} \end{pmatrix} g^T$$

Then, for any given matrix D, we get that

$$\sum_{k=1}^{m}\sum_{\ell=1}^{m} D_{k\ell}y_k^1 y_\ell^2 = y^1 D y^{2T} = \sum_{k=1}^{2m}\sum_{\ell=1}^{2m} D'_{k\ell}(1,2)g_k g_\ell$$

where

$$E\sum_{k=1}^{2m}\sum_{\ell=1}^{2m} D'_{k\ell}(1,2)g_k g_\ell = \operatorname{tr}(D'(1,2)) = b_{12}\operatorname{tr}(A_0 D).$$

Applying this to our setting, let $D = \text{diag}(v^i \otimes v^j)$. Then

$$\begin{aligned}
\langle v^i \circ y^i, v^j \circ y^j \rangle &= \sum_{k=1}^{m} y_k^i v_k^i v_k^j y_k^j = y^i D y^{jT} \\
&= g(c_i \otimes A_0^{1/2}) D (c_j^T \otimes A_0^{1/2}) g^T \\
&= g(c_i c_j^T \otimes (A_0^{1/2} D A_0^{1/2})) g^T = g D'(i,j) g^T
\end{aligned}$$

where $c_i = (c_{ii}, c_{ij})^T$ and $c_j = (c_{ij}, c_{jj})^T$, and therefore

$$c_i c_j^T = \begin{pmatrix} c_{ii}c_{ij} & c_{ii}c_{jj} \\ c_{ij}c_{ij} & c_{ij}c_{jj} \end{pmatrix}$$

If we partition $g = (g_1, g_2)$, consider the following quadratic forms

$$\begin{aligned}
Z &= g_1(A_0^{1/2} D A_0^{1/2}) g_1^T - E g_1(A_0^{1/2} D A_0^{1/2}) g_1^T \\
Z' &= g_2(A_0^{1/2} D A_0^{1/2}) g_2^T - E g_2(A_0^{1/2} D A_0^{1/2}) g_2^T \\
U &= g_1(A_0^{1/2} D A_0^{1/2}) g_2^T - E g_1(A_0^{1/2} D A_0^{1/2}) g_2^T
\end{aligned}$$

For Z, Z' independent.

For $i \neq j$ we have that $ED = \text{diag}(\zeta_1^2, \ldots, \zeta_m^2)$ and

$$E\langle v^i \circ y^i, v^j \circ y^j \rangle = b_{ij} E\langle A_0, D \rangle = b_{ij} \sum_{k=1}^{m} a_{kk} \zeta_k^2.$$

So

$$\begin{aligned}
&\frac{1}{\sqrt{b_{ii}b_{jj}}}gD'(i,j)g^T - E\frac{1}{\sqrt{b_{ii}b_{jj}}}gD'(i,j)g^T \\
&\quad = \left|\frac{b_{ij}}{\sqrt{b_{ii}b_{jj}}}\left(\frac{1}{b_{ij}}(\langle v^i \circ y^i, v^j \circ y^j\rangle - E\langle v^i \circ y^i, v^j \circ y^j\rangle)\right)\right| \\
&\quad \leq |\rho_{ij}(B_0)|\left(\left|\frac{c_{ii}c_{ij}}{b_{ij}}Z + \frac{c_{ij}c_{jj}}{b_{ij}}Z'\right| + \left|\frac{c_{ii}c_{jj} + c_{ij}^2}{b_{ij}}U\right|\right) \\
&\quad = |\rho_{ij}(B_0)|\left(|s_1 Z + s_2 Z'| + \left|\frac{c_{ii}c_{jj} + c_{ij}^2}{b_{ij}}U\right|\right) \\
&\quad \leq |\rho_{ij}(B_0)||s_1 Z + s_2 Z'| + |U| \qquad (5.18)
\end{aligned}$$

where $s_1 = (c_{ii}c_{ij})/b_{ij}$, $s_2 = (c_{ij}c_{jj})/b_{ij}$, and $s_1, s_2 \in [0,1]$ with $s_1 + s_2 = 1$ since $B_{0,(i,j)}^{1/2}$ is positive definite and $c_{ii}c_{ij} + c_{ij}c_{jj} = \mathrm{tr}(c_i c_j^T) = b_{ij}$.

Therefore, we can use (5.18) to get that

$$\begin{aligned}
&P\left(\left|\frac{\langle v^i \circ y^i, v^j \circ y^j\rangle}{\sqrt{b_{ii}b_{jj}}} - E\frac{\langle v^i \circ y^i, v^j \circ y^j\rangle}{\sqrt{b_{ii}b_{jj}}}\right| > t\right) \\
&\quad = P\left(\left|\frac{gD'(i,j)g^T}{\sqrt{b_{ii}b_{jj}}} - E\frac{gD'(i,j)g^T}{\sqrt{b_{ii}b_{jj}}}\right| > t\right) \\
&\quad \leq P(|\rho_{ij}(B_0)||s_1 Z + s_2 Z'| > t/2) + P(|U| > t/2) \\
&\quad \overset{(a)}{\leq} P(|\rho_{ij}(B_0)||Z| > t/2) + P(|\rho_{ij}(B_0)||Z'| > t/2) + P(|U| > t/2) \\
&\quad =: p_1 + p_2 + p_3
\end{aligned}$$

where (a) is because $s_1, s_2 \in [0,1]$ and $s_1 + s_2 = 1$ means that $|s_1 Z + s_2 Z'| > k \implies |Z| > k$ and/or $|Z'| > k$.

We can apply Theorem 5.3.4 to p_1 and p_2 with Proposition 5.3.9, noting that $p_1 = P(|\rho_{ij}(B_0)||Z| > t/2) \leq P(|Z| > t/2)$, to get that

$$p_1, p_2 \leq 2\exp\left(-c_2 \min\left(\frac{t^2}{4K^4(\sum_{k=1}^m a_{kk}^2\zeta_k^2 + \sum_{k\neq\ell} a_{k\ell}^2\zeta_k^2\zeta_\ell^2)}, \frac{t}{2K^2\|A_0\|_2}\right)\right).$$

Using Theorem 5.3.5, we get that

$$p_3 \leq 2\exp\left(-c_2 \min\left(\frac{t^2}{4K^4(\sum_{k=1}^m a_{kk}^2\zeta_k^2 + \sum_{k\neq\ell} a_{k\ell}^2\zeta_k^2\zeta_\ell^2)}, \frac{t}{2K^2\|A_0\|_2}\right)\right).$$

Adding these probabilities gives us our desired result.

***A*-side.** The diagonal case proceeds similarly to above. We can write $x^i = (a_{ii}B_0)^{1/2}(g_1, \ldots, g_n)^T$ for $g_1, \ldots, g_n \overset{\text{i.i.d}}{\sim} Z$. So for $D = \operatorname{diag}(u^i)$, can then show that

$$\begin{aligned}\langle u^i \circ x^i, u^i \circ x^i\rangle &= \sum_{k=1}^n \sum_{\ell=1}^n D_{k\ell} x_\ell^i x_k^i \\ &= g a_{ii} B_0^{1/2} D B_0^{1/2} g^T\end{aligned}$$

Where $ED = \zeta_i I$, we can then apply Theorem 5.3.4 to the quadratic form

$$gB_0^{1/2}DB_0^{1/2}g^T - EgB_0^{1/2}DB_0^{1/2}g^T$$

to get that

$$\begin{aligned}&P\left(\frac{1}{a_{ii}}\left|\langle u^i \circ x^i, u^i \circ x^i\rangle - \zeta_i \operatorname{tr}(B_0)\right| > t\right) \\ &\qquad \leq 2\exp\left(-c_2 \min\left(\frac{t^2}{4K^4(\zeta_i \sum_{k=1}^n b_{kk}^2 + \zeta_i^2 \sum_{k\neq\ell} b_{k\ell}^2)}, \frac{t}{2K^2\|B_0\|_2}\right)\right) \\ &\qquad \leq 2\exp\left(-c_2 \min\left(\frac{t^2}{4K^4\zeta_i\|B_0\|_F}, \frac{t}{2K^2\|B_0\|_2}\right)\right)\end{aligned}$$

The off-diagonal case follows the same steps as the off-diagonal case for the B-side above, noting that Proposition 5.3.9 and the argument using it are symmetric in A and B. The primary difference being that $D = \text{diag}(u^i \circ u^j)$ with $ED = \zeta_i \zeta_j I$. This gives us that

$$
\begin{aligned}
&P\left(\left|\frac{\langle u^i \circ x^i, u^j \circ x^j\rangle}{\sqrt{a_{ii}a_{jj}}} - \rho_{ij}(A_0)\zeta_i\zeta_j \operatorname{tr}(B_0)\right| > t\right) \\
&\quad \le 6\exp\left(-c_2 \min\left(\frac{t^2}{4K^4(\zeta_i\zeta_j \sum_{k=1}^n b_{kk}^2 + \zeta_i^2\zeta_j^2 \sum_{k\neq\ell} b_{k\ell}^2)}, \frac{t}{2K^2\|B_0\|_2}\right)\right) \\
&\quad \le 6\exp\left(-c_2 \min\left(\frac{t^2}{4K^4\zeta_i\zeta_j\|B_0\|_F}, \frac{t}{2K^2\|B_0\|_2}\right)\right)
\end{aligned}
$$

□

Proof of Proposition 5.3.9. Without loss of generality, let $i = 1, j = 2$. So

$$
\begin{aligned}
\|c_1 c_2^T\|_F^2 = \left\|\begin{pmatrix} c_{11}c_{12} & c_{11}c_{22} \\ c_{12}c_{12} & c_{12}c_{22}\end{pmatrix}\right\|_F^2 &= \operatorname{tr}(c_2 c_1^T c_1 c_2^T) \\
&= \|c_1\|_2^2 \|c_2\|_2^2 = (c_{11}^2 + c_{12}^2)(c_{12}^2 + c_{22}^2) = b_{11}b_{22}
\end{aligned}
$$

So

$$
\|D'(1,2)\|_F^2 = \|c_1 c_2^T \otimes (A_0^{1/2} D A_0^{1/2})\|_F^2 = b_{11}b_{22}\|A_0^{1/2} D A_0^{1/2}\|_F
$$

Thus (5.15) holds. Using a similar argument we can show (5.14).

To show the last inequality,

$$
\begin{aligned}
(c_{11}c_{22} + c_{12}^2)^2 &= (c_{11}c_{22})^2 + c_{12}^4 + 2c_{11}c_{22}c_{12}^2 \\
&\le (c_{11}c_{22})^2 + c_{12}^4 + (c_{11}c_{12})^2 + (c_{22}c_{12})^2 \\
&= (c_{11}^2 + c_{12}^2)(c_{12}^2 + c_{22}^2) = b_{11}b_{22}.
\end{aligned}
$$

So

$$
\rho_{ij}(B_0)\frac{c_{ii}c_{jj} + c_{ij}^2}{b_{ij}} \le \frac{b_{ij}}{\sqrt{b_{ii}b_{jj}}}\frac{\sqrt{b_{ii}b_{jj}}}{b_{ij}} = 1.
$$

□

5.3.3 Mask estimation

Recall that

$$S_c = \frac{n}{n-1}\operatorname{tr}(\mathcal{X}^T\mathcal{X} \circ \widehat{M}) - \frac{1}{n-1}\operatorname{tr}(\mathcal{X}^T\mathcal{X}) \tag{5.1}$$

and that

$$\widehat{\mathcal{M}}_{k\ell} = \begin{cases} \operatorname{tr}(\mathcal{X}\mathcal{X}^T) & k = \ell \\ S_c & k \neq \ell \end{cases}. \tag{3.13}$$

Proposition 5.3.10. *$\widehat{\mathcal{M}}$ is an unbiased estimate of the mask $\mathcal{M}$ multiplied by $\operatorname{tr}(B_0)$. So*

$$\begin{aligned} E\operatorname{tr}(\mathcal{X}\mathcal{X}^T) &= E\operatorname{tr}(\sum_{i=1}^{n}(v^i \circ y^i) \otimes (v_i \circ y_i)) = \sum_{i=1}^{n}\operatorname{tr}(E(v^i \otimes v^i) \circ E(y^i \otimes y^i)) \\ &= \operatorname{tr}(M \circ A_0 \operatorname{tr}(B_0)) = \sum_{i=1}^{m}\zeta_i a_{ii}\operatorname{tr}(B_0) \end{aligned}$$

And

$$ES_c = E\left(\frac{n}{n-1}\operatorname{tr}(\mathcal{X}^T\mathcal{X} \circ \widehat{M}) - \frac{1}{n-1}\operatorname{tr}(\mathcal{X}^T\mathcal{X})\right) = \sum_{j=1}^{m} a_{jj}\zeta_j^2 \operatorname{tr}(B_0)$$

Proof. Let $M^i = v^i \otimes v^i$. Then

$$\begin{aligned} \operatorname{tr}(\mathcal{X}^T\mathcal{X} \circ \widehat{M}) &= \operatorname{tr}\left(\mathcal{X}^T\mathcal{X} \circ \Big(\frac{1}{n}\sum_{j=1}^{n} v^j \otimes v^j\Big)\right) = \frac{1}{n}\sum_{j=1}^{n}\operatorname{tr}(\mathcal{X}^T\mathcal{X} \circ (v^j \otimes v^j)) \\ &= \frac{1}{n}\sum_{j=1}^{n}\operatorname{tr}\left(\Big(\sum_{i=1}^{n}(v^i \otimes v^i) \circ (y^i \otimes y^i)\Big) \circ (v^j \otimes v^j)\right) \\ &= \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{n}\operatorname{tr}(M^i \circ (y^i \otimes y^i) \circ M^j) \\ &= \frac{1}{n}\sum_{i=1}^{n}\operatorname{tr}(M^i \circ (y^i \otimes y^i)) + \frac{1}{n}\sum_{i \neq j}^{n}\operatorname{tr}(M^i \circ (y^i \otimes y^i) \circ M^j) \\ &= \frac{1}{n}\operatorname{tr}(\mathcal{X}^T\mathcal{X}) + \frac{1}{n}\sum_{i \neq j}^{n}\operatorname{tr}(M^i \circ M^j \circ (y^i \otimes y^i)). \end{aligned}$$

Taking expectations and recalling that M^i, M^j, and X are independent,

$$\begin{aligned}
E\operatorname{tr}(\mathcal{X}^T\mathcal{X}\circ\widehat{M}) &= \frac{1}{n}E\operatorname{tr}(\mathcal{X}^T\mathcal{X}) + \frac{1}{n}\sum_{i\neq j}^{n}\operatorname{tr}(E(M^i\circ M^j)\circ E(y^i\otimes y^i)) \\
&= \frac{1}{n}E\operatorname{tr}(\mathcal{X}^T\mathcal{X}) + \operatorname{tr}(\operatorname{diag}(\zeta_1^2,\ldots,\zeta_m^2)\circ\frac{1}{n}\sum_{i\neq j}^{n}E(y^i\otimes y^i)) \\
&\overset{(a)}{=} \frac{1}{n}E\operatorname{tr}(\mathcal{X}^T\mathcal{X}) + \operatorname{tr}(\operatorname{diag}(\zeta_1^2,\ldots,\zeta_m^2)\circ\frac{1}{n}(n-1)A_0\operatorname{tr}(B_0)) \\
&= \frac{1}{n}E\operatorname{tr}(\mathcal{X}^T\mathcal{X}) + \frac{n-1}{n}\operatorname{tr}(B_0)\sum_{i=1}^{m}a_{ii}\zeta_i^2 \qquad (5.19)
\end{aligned}$$

where (a) is because

$$\sum_{i\neq j}^{n}E(y^i\otimes y^i) = (n-1)\sum_{i=1}^{n}E(y^i\otimes y^i) = (n-1)A_0\operatorname{tr}(B_0)$$

Rearranging (5.19) proves the result. □

Proposition 5.3.11. *With probability at least* $1-\frac{n^2}{m^4}$,

$$\left|\frac{S_c}{ES_c}-1\right| \leq \alpha$$

Also, with probability at least $1-1/m^2$,

$$\left|\frac{\operatorname{tr}(\mathcal{X}^T\mathcal{X})}{E\operatorname{tr}(\mathcal{X}^T\mathcal{X})}-1\right| \leq \alpha_{\mathrm{diag}}\frac{\|B_0\|_F}{\operatorname{tr}(B_0)}$$

Let Λ_M *be the even that both of these hold, which occurs with probability at least* $1-1/m^2-n^2/m^4$.

Proof. **Concentration of $\mathbf{S_c}$.** We first decompose S_c into a sum

$$\begin{aligned}
S_c &= \frac{n}{n-1}\operatorname{tr}(\mathcal{X}^T\mathcal{X}\circ \widehat{M}) - \frac{1}{n-1}\operatorname{tr}(\mathcal{X}^T\mathcal{X}) \\
&= \frac{1}{n-1}\sum_{j=1}^{n}\sum_{i=1}^{n}\operatorname{tr}(M^i\circ M^j\circ(y^i\otimes y^i)) - \frac{1}{n-1}\sum_{i=1}^{n}\operatorname{tr}(M^i\circ(y^i\otimes y^i)) \\
&= \frac{1}{n-1}\sum_{j=1}^{n}\sum_{i\neq j}^{n}\operatorname{tr}(M^i\circ M^j\circ(y^i\otimes y^i)) = \frac{1}{n-1}\sum_{j=1}^{n}\sum_{i\neq j}^{n}\operatorname{tr}((v^i\otimes v^j)\circ(y^i\otimes y^i)) \\
&= \frac{1}{n-1}\sum_{j=1}^{n}\sum_{i\neq j}^{n}(y^i)^T\operatorname{diag}(v^i\otimes v^j)y^i,
\end{aligned}$$

so

$$|S_c - ES_c| \le \frac{1}{n-1}\sum_{j=1}^{n}\sum_{i\neq j}^{n}|(y^i)^T\operatorname{diag}(v^i\otimes v^j)y^i - E(y^i)^T\operatorname{diag}(v^i\otimes v^j)y^i|.$$

Note that $y^i = (b_{ii}A_0)^{1/2}Z_{i\cdot}$, so we can apply Theorem 5.3.4 to each of these terms with $D_0 = (b_{ii}A_0)^{1/2}$ and $D_\xi = \operatorname{diag}(v^i\otimes v^j)$ to get that

$$\begin{aligned}
&P\left(\forall\, i\neq j, \left|(y^i)^T\operatorname{diag}(v^i\otimes v^j)y^i - E(y^i)^T\operatorname{diag}(v^i\otimes v^j)y^i\right| \le b_{ii}\tau\right) \\
&\qquad \ge 1 - 2n^2\exp\left(-c_2\min\left(\frac{\tau^2}{4K^4(\sum_{k=1}^{m}\zeta_k^2 a_{kk}^2 + \sum_{k\neq\ell}a_{k\ell}^2\zeta_k^2\zeta_\ell^2)}, \frac{\tau}{2K^2\|A_0\|_2}\right)\right)
\end{aligned}$$

where we applied a union bound over all $i\neq j$.

Choosing $\tau = CK^2\log m\|A_0\|_2 + CK^2\log^{1/2}(m\vee n)\sqrt{\sum_{k=1}^{m}\zeta_k^2 a_{kk}^2 + \sum_{k\neq\ell}a_{k\ell}^2\zeta_k^2\zeta_\ell^2}$, we get that with probability at least $1 - n^2/m^4$, for all $i,j = 1,\ldots,n$ and $i\neq j$,

$$\begin{aligned}
&\frac{1}{n-1}\sum_{j\neq i}^{n}\left|(y^i)^T\operatorname{diag}(v^i\otimes v^j)y^i - E(y^i)^T\operatorname{diag}(v^i\otimes v^j)y^i\right| \le b_{ii}\tau \\
\Longrightarrow\ &\frac{1}{n-1}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\left|(y^i)^T\operatorname{diag}(v^i\otimes v^j)y^i - E(y^i)^T\operatorname{diag}(v^i\otimes v^j)y^i\right| \le \operatorname{tr}(B_0)\tau
\end{aligned}$$

Recall that $ES_c = \text{tr}(B_0)\sum_{k=1}^n a_{kk}\zeta_k^2$. So

$$\begin{aligned}
|S_c - ES_c| &\le \text{tr}(B_0)\tau \\
\left|\frac{S_c}{ES_c} - 1\right| &\le \frac{\tau}{\sum_{k=1}^n a_{kk}\zeta_k^2} \\
&= \frac{CK^2 \log m\|A_0\|_2 + CK^2\log^{1/2}(m \vee n)\sqrt{\sum_{k=1}^m \zeta_k^2 a_{kk}^2 + \sum_{k\neq\ell} a_{k\ell}^2\zeta_k^2\zeta_\ell^2}}{\sum_{k=1}^n a_{kk}\zeta_k^2} \\
&\le \alpha
\end{aligned}$$

for α as defined in (3.18).

Concentration of $\text{tr}(\mathcal{X}^T\mathcal{X})$. We first note that $\text{tr}(\mathcal{X}^T\mathcal{X}) = (\text{vec}(\mathbb{U}) \circ \text{vec}(\mathbb{X}))^T(\text{vec}(\mathbb{U}) \circ \text{vec}(\mathbb{X}))$, where $\text{vec}(\mathbb{X})$ has covariance $A \otimes B$. So we can write $\text{vec}(\mathbb{X}) = (A\otimes B)^{1/2}g$, where $g \in \mathbb{R}^{nm}$ has subgaussian entries with subgaussian constant K, mean zero, and unit variance.

Let $D = \text{diag}(\text{vec}(\mathbb{U}))$, where $ED = \text{diag}((\zeta_1, \ldots, \zeta_1, \zeta_2, \ldots, \zeta_2, \ldots, \zeta_m, \ldots, \zeta_m))$. Then we get that

$$\text{tr}(\mathcal{X}^T\mathcal{X}) = g(A\otimes B)^{1/2}D(A\otimes B)^{1/2}g^T$$

Applying Theorem 5.3.4 to this quadratic form and plugging in

$$t = CK^2\log(m)\|A_0\|_2\|B_0\|_2 + CK^2\log^{1/2}(m)\sqrt{\|\text{diag}(B_0)\|_F^2\sum_{k=1}^m a_{kk}^2\zeta_k + \|B_0\|_F^2\sum_{k\neq\ell} a_{k\ell}^2\zeta_k\zeta_\ell}$$

we get that, with probability at least $1 - 1/m^2$,

$$\begin{aligned}
&|\text{tr}(\mathcal{X}^T\mathcal{X}) - E\,\text{tr}(\mathcal{X}^T\mathcal{X})| \le t \\
&\qquad \le CK^2\log(m)\|A_0\|_2\|B_0\|_2 + CK^2\log^{1/2}(m)\|B_0\|_F\sqrt{\sum_{k=1}^m a_{kk}^2\zeta_k + \sum_{k\neq\ell} a_{k\ell}^2\zeta_k\zeta_\ell}
\end{aligned}$$

So, since $E\operatorname{tr}(\mathcal{X}^T\mathcal{X}) = \sum_{i=1}^m \zeta_i a_{ii} \operatorname{tr}(B_0)$ as shown in Proposition 5.3.10, we have that

$$\begin{aligned}
&\left|\frac{\operatorname{tr}(\mathcal{X}^T\mathcal{X})}{E\operatorname{tr}(\mathcal{X}^T\mathcal{X})} - 1\right| \\
&\quad \leq \frac{CK^2\log(m)\|A_0\|_2}{\sum_{i=1}^m \zeta_i a_{ii}} \frac{\|B_0\|_2}{\operatorname{tr}(B_0)} + \frac{CK^2\log^{1/2}(m)\sqrt{\sum_{k=1}^m a_{kk}^2\zeta_k + \sum_{k\neq\ell} a_{k\ell}^2\zeta_k\zeta_\ell}}{\sum_{i=1}^m \zeta_i a_{ii}} \frac{\|B_0\|_F}{\operatorname{tr}(B_0)} \\
&\quad \leq \alpha_{\text{diag}}\frac{\|B_0\|_F}{\operatorname{tr}(B_0)}
\end{aligned}$$

□

5.4 Estimation of B with unknown group means

Assume our rows are sorted by group. Denote

$$\mathbb{U} = \begin{bmatrix} \mathbb{U}^1 \\ \mathbb{U}^2 \end{bmatrix}, \quad \mathbb{X} = \begin{bmatrix} \mathbb{X}^1 \\ \mathbb{X}^2 \end{bmatrix}, \qquad \text{where } \mathbb{U}^1, \mathbb{X}^1 \in \mathbb{R}^{n_1 \times m} \text{ and } \mathbb{U}^2, \mathbb{X}^2 \in \mathbb{R}^{n_2 \times m}$$

Let n_{1j}, n_{2j} be the number of observed values for variables j in group 1 and 2, respectively. Then we note that

$$\tilde{\mathcal{X}}_{\cdot j} = u^j \circ (I - P^j)\mathcal{X}_{\cdot j} \qquad \text{where } P^j = \begin{bmatrix} \frac{1}{n_{1j}}\vec{1}_{n_1}\mathbb{U}_{\cdot j}^{1T} & 0 \\ 0 & \frac{1}{n_{2j}}\vec{1}_{n_2}\mathbb{U}_{\cdot j}^{2T} \end{bmatrix} \tag{5.20}$$

and that $(I - P^j)\mathcal{X}_{\cdot j} = (I - P^j)x^j$.

Recall that for some absolute constants C, C_1 we define

$$\alpha = CK^2\frac{\log m\|A_0\|_2}{\sum_{k=1}^m a_{kk}\zeta_k^2} + CK^2\log^{1/2}(m \vee n)\frac{\sqrt{a_\infty\|A_0\|_2}}{a_{\min}\|\zeta\|_2} \tag{3.18}$$

Table 5.1: Notation

Notation	Meaning
Sample Sizes	
$n = n_1 + n_2$	Total number of rows
n_1, n_2	Number of rows per group
$n_{1k} = \sum_{i=1}^{n_1} \mathbb{U}_{ik}, n_{2k} = \sum_{i=n_1+1}^{n} \mathbb{U}_{ik}$	Number of observed values per group for column k
$n_{1k\ell} = \sum_{i=1}^{n_1} \mathbb{U}_{ik}\mathbb{U}_{i\ell}, n_{2k\ell} = \sum_{i=n_1+1}^{n} \mathbb{U}_{ik}\mathbb{U}_{i\ell}$	Number of times cols k, ℓ are both observed per group
$n_{1\min} = \min_k n_{1k}, n_{2\min} = \min_k n_{2k}$	Minimum number of observed values per group
$\hat{\zeta}_k = (n_{1k} + n_{2k})/n$	Estimated sampling probability in column k
Matrices	
A	Column-wise covariance
$B = \begin{pmatrix} B^1 & B^{12} \\ B^{21} & B^2 \end{pmatrix}$	Row-wise covariance
$B_{i\cdot}, B_{\cdot j}$	ith row and jth column of B, respectively
P^j, defined in (5.40)	Proj matrix, calculates observed group means for col j
Probabilities	
$\tilde{\zeta}_{2k} = P(\mathbb{U}_{ik}\mathbb{U}_{jk} = 1 \mid n_{1k})$	Probs conditional on the number of observed values
$\tilde{\zeta}_{3k} = P(\mathbb{U}_{\ell k}\mathbb{U}_{ik}\mathbb{U}_{jk} = 1 \mid n_{1k})$	

and

$$\begin{aligned} \alpha_{\text{mean}} = \frac{C_1}{\zeta_{\min}^2 n_{\min}} \left(\|B\|_1 + \frac{1}{\zeta_{\min}} \frac{\operatorname{tr}(B)}{n_{\min}} \right) + C_1 K^2 \frac{\log m}{\zeta_{\min}^3 \operatorname{tr}(A)} \|A\|_2 \\ + C_1 K^2 \log^{1/2}(m \vee n) \frac{\|A\|_F}{\zeta_{\min}^3 \operatorname{tr}(A)} + \alpha \end{aligned} \tag{3.20}$$

Proofs of Theorems 3.2.1 and 3.2.2. To prove the B-side of these results, we simply apply Theorem 5.2.1 using the correlation convergence results in Theorem 5.4.1. We perform a similar application in Section 5.5. □

Theorem 5.4.1. *Consider data generating random matrices as in* (3.1) *and* (3.2) *and suppose Assumptions 1 and 2 hold.*

Let $m \vee n \geq 3$ and $\alpha_{\text{mean}} < 1/3$. Then, with probability at least $1 - 25/(m \vee n)^2$ and $\widetilde{\Gamma}(B_0)$ as defined in (3.10)*, we get that*

$$\forall i \neq j, \quad |\widetilde{\Gamma}_{ij}(B_0) - \rho_{ij}(B_0)| \leq 3\alpha_{\text{mean}}$$

Similarly, with probability at least $1 - 27/(m \vee n)^2$ and $\widehat{\Gamma}(B_0)$ as defined in (3.15)*, we get that*

$$\forall i \neq j, \quad |\widehat{\Gamma}_{ij}(B_0) - \rho_{ij}(B_0)| \leq 9\frac{\alpha_{\text{mean}}}{\zeta_{\min}}$$

Proof. Using Corollary 5.4.5 and the proof of the first statement proceeds identically to the proof of Theorem 5.3.2.

The proof of the second statement is similar to the proof of Theorem 5.3.3 using Proposition 5.4.8.

$$\begin{aligned}
|\widehat{\Gamma}_{ij}(B_0) - \widetilde{\Gamma}_{ij}(B_0)| &\leq |\widetilde{\Gamma}_{ij}(B_0)| \left| \frac{\operatorname{tr}(\mathcal{X}\mathcal{X}^T)}{\operatorname{tr}(B_0)\sum_{i=1}^m \zeta_i a_{ii}} \frac{\operatorname{tr}(B_0)\sum_{i=1}^n \zeta_i^2 a_{ii}}{S_c} - 1 \right| \\
&\leq |\widetilde{\Gamma}_{ij}(B_0)| \max\left(\left| \frac{1+\alpha_{\text{mean}}}{1-\alpha_{\text{mean}}/\zeta_{\min}} - 1 \right|, \left| \frac{1-\alpha_{\text{mean}}}{1+\alpha_{\text{mean}}/\zeta_{\min}} - 1 \right| \right) \\
&\leq |\widetilde{\Gamma}_{ij}(B_0)| \max\left(1 + 3\frac{\alpha_{\text{mean}}}{\zeta_{\min}} - 1, 1 - \left(1 - \alpha_{\text{mean}} - \frac{\alpha_{\text{mean}}}{\zeta_{\min}} \right) \right) \\
&\leq (1 + 3\alpha_{\text{mean}})3\frac{\alpha_{\text{mean}}}{\zeta_{\min}} = 3\frac{\alpha_{\text{mean}}}{\zeta_{\min}} + 9\frac{\alpha_{\text{mean}}^2}{\zeta_{\min}}
\end{aligned}$$

and therefore, when $\alpha_{\text{mean}} < 1/3$,

$$|\widehat{\Gamma}_{ij}(B_0) - \rho_{ij}(B_0)| \leq |\widehat{\Gamma}_{ij}(B_0) - \widetilde{\Gamma}_{ij}(B_0)| + |\widetilde{\Gamma}_{ij}(B_0) - \rho_{ij}(B_0)| \leq 9\frac{\alpha_{\text{mean}}}{\zeta_{\min}}$$

□

5.4.1 Concentration results

Define the Gram matrix $\hat{S}(B) = \tilde{\mathcal{X}}\tilde{\mathcal{X}}^T$. We get that

$$\begin{aligned}\hat{S}(B) &= \sum_{k=1}^{m} \tilde{\mathcal{X}}_{\cdot k}\tilde{\mathcal{X}}_{\cdot k}^T \\ &= \sum_{k=1}^{m} (u^k \circ (I - P^k)x^k)(u^{kT} \circ x^{kT}(I - P^k)^T) \\ &= \sum_{k=1}^{m} (u^k u^{kT}) \circ ((I - P^k)x^k x^{kT}(I - P^k)^T) \qquad (5.21)\end{aligned}$$

We now consider the (i, j)th entry of this matrix. There are three cases to consider, when $i = j$, $i \neq j$ but i and j are in the same group, and when i and j are in different groups. The following propositions, which we prove in Section 5.4.2, present concentraion results for each of these cases.

Proposition 5.4.2. *For $i \neq j$, where i and j are both in group $g \in \{1, 2\}$, we get that, with probability at least $1 - 1m/(n \vee m)^c - 24n_g/(n \vee m)^c$ for some constant c,*

$$\begin{aligned}\left|\frac{\hat{S}(B)_{ij}}{\sqrt{b_{ii}b_{jj}}\mathcal{M}_{ij}} - \rho_{ij}(B)\right| \leq C_2 \frac{1}{\zeta_{\min}^2 n_g}\left(\max_{1\leq k\leq n_g} |B^g_{\cdot k}|_1 + \frac{\operatorname{tr}(B^g)}{n_g}\right) + CK^2 \frac{\log m}{\zeta_{\min}^3 \operatorname{tr}(A)}\|A\|_2 \\ + CK^2 \log^{1/2}(m \vee n)\frac{\|A\|_F}{\zeta_{\min}^3 \operatorname{tr}(A)} + \alpha_{\text{offd}}\end{aligned} \qquad (5.22)$$

Proposition 5.4.3. *For $i = j$, where i is in group $g \in \{1, 2\}$, we get that, with probability at least $1 - 1m/(n \vee m)^c - 18n_g/(n \vee m)^c$ for some constant c,*

$$\begin{aligned}\left|\frac{\hat{S}(B)_{ii}}{b_{ii}\mathcal{M}_{ii}} - 1\right| \leq \frac{C_2}{\zeta_{\min}^2 n_g}\left(\max_{1\leq k\leq n_g} |B^g_{\cdot k}|_1 + \frac{1}{\zeta_{\min}}\frac{\operatorname{tr}(B^g)}{n_g}\right) + CK^2 \frac{\log m}{\zeta_{\min}^3 \operatorname{tr}(A)}\|A\|_2 \\ + CK^2 \log^{1/2}(m \vee n)\frac{\|A\|_F}{\zeta_{\min}^3 \operatorname{tr}(A)} + \alpha_{\text{offd}}\end{aligned} \qquad (5.23)$$

Proposition 5.4.4. *Let $n_{\min} = \min_g n_g$. For $i \neq j$, where i and j are in different groups,* *we get that, with probability at least* $1 - 1m/(n \vee m)^c - 18n_g/(n \vee m)^c$ *for some constant c,*

$$\left| \frac{\hat{S}(B)_{ij}}{\sqrt{b_{ii}b_{jj}}\mathcal{M}_{ij}} - \rho_{ij}(B) \right| \leq C_2 \frac{1}{\zeta_{\min}^2 n_{\min}} \max_g \max_{1 \leq k \leq n_g} |B^g_{\cdot k}|_1 + CK^2 \frac{\log m}{\zeta_{\min}^3 \operatorname{tr}(A)} \|A\|_2 \\ + CK^2 \log^{1/2}(m \vee n) \frac{\|A\|_F}{\zeta_{\min}^3 \operatorname{tr}(A)} + \alpha_{\text{diag}} \tag{5.24}$$

We can combine these results with a union bound to obtain the following overall concentration bound.

Corollary 5.4.5. *Let* $n_{\min} = \min_g n_g$*. Then, with probability at least* $1 - 25/(n \vee m)^2$*, the following event holds simultaneously for all* $i, j = 1, \ldots, n$*.*

$$\left| \frac{\hat{S}(B)_{ij}}{\sqrt{b_{ii}b_{jj}}\mathcal{M}_{ij}} - \rho_{ij}(B) \right| \leq \frac{C_2}{\zeta_{\min}^2 n_{\min}} \max_g \left(\max_{1 \leq k \leq n_g} |B^g_{\cdot k}|_1 + \frac{1}{\zeta_{\min}} \frac{\operatorname{tr}(B^g)}{n_{\min}} \right) + CK^2 \frac{\log m}{\zeta_{\min}^3 \operatorname{tr}(A)} \|A\|_2 \\ + CK^2 \log^{1/2}(m \vee n) \frac{\|A\|_F}{\zeta_{\min}^3 \operatorname{tr}(A)} + \alpha =: \alpha_{\text{mean}}$$

5.4.2 Concentration proofs

Proof of Proposition 5.4.2. Without loss of generality, assume i and j are in group 1. Define the standard basis vectors $e'_i = (\mathbb{1}_{k=i})_{k=1}^n \in \mathbb{R}^n$ and $e_i = (\mathbb{1}_{k=i})_{k=1}^{n_1} \in \mathbb{R}^{n_1}$. Then we get that

$$\begin{aligned} \hat{S}(B)_{ij} &= \sum_{k=1}^m \mathbb{U}_{ik}\mathbb{U}_{jk}(e'_i - P^k_{i\cdot})^T x^k x^{kT}(e'_j - P^k_{j\cdot}) \\ &= \sum_{k=1}^m \mathbb{U}_{ik}\mathbb{U}_{jk}\mathbb{X}_{ik}\mathbb{X}_{jk} - \sum_{k=1}^m \mathbb{U}_{ik}\mathbb{U}_{jk}\frac{1}{n_{1k}}(e_i^T X^1_{\cdot k}(X^1_{\cdot k})^T U^1_{\cdot k}) \\ &\quad - \sum_{k=1}^m \mathbb{U}_{ik}\mathbb{U}_{jk}\frac{1}{n_{1k}}(e_j^T X^1_{\cdot k}(X^1_{\cdot k})^T U^1_{\cdot k}) + \sum_{k=1}^m \mathbb{U}_{ik}\mathbb{U}_{jk}\frac{1}{n_{1k}^2}((U^1_{\cdot k})^T X^1_{\cdot k}(X^1_{\cdot k})^T U^1_{\cdot k}) \\ &= \sum_{k=1}^m \mathbb{U}_{ik}\mathbb{U}_{jk}\mathbb{X}_{ik}\mathbb{X}_{jk} - S_{B1} - S_{B2} + S_{B3} \end{aligned} \tag{5.25}$$

Rearranging, we can decompose our excess error over the mean-zero case as

$$\begin{aligned}&\left|\hat{S}(B)_{ij} - \sum_{k=1}^{m} \mathbb{U}_{ik}\mathbb{U}_{jk}\mathbb{X}_{ik}\mathbb{X}_{jk}\right| \\ &\qquad \le |ES_{B1}| + |S_{B1} - ES_{B1}| + |ES_{B2}| + |S_{B2} - ES_{B2}| + |ES_{B3}| + |S_{B3} - ES_{B3}|\end{aligned} \tag{5.26}$$

From here, condition on $\{n_{1k}\}_{k=1}^{m}$. Note that even after conditioning on the column counts, for $k \neq h$ we still get that

$$\mathbb{U}_{ik} \perp \mathbb{U}_{ih} \mid \{n_{1k}\} \qquad \mathbb{U}_{ik}\mathbb{U}_{jk} \perp \mathbb{U}_{ih}\mathbb{U}_{jh} \mid \{n_{1k}\}$$
$$\mathbb{U}_{\ell k}\mathbb{U}_{ik}\mathbb{U}_{jk} \perp \mathbb{U}_{\ell h}\mathbb{U}_{ih}\mathbb{U}_{jh} \mid \{n_{1k}\}$$

Since the conditioning only induces dependences bewteen entries of $\mathbb{U}$ in the same column, but here we always consider entries across different columns.

We then bound the conditional probabilities of these products given fixed column counts $\{n_{1k}\}$. So let $\tilde{\zeta}_{3k} = P(\mathbb{U}_{\ell k}\mathbb{U}_{ik}\mathbb{U}_{jk} = 1 \mid n_{1k})$ for $\ell \neq i, j$ and $\tilde{\zeta}_{2k} = P(\mathbb{U}_{ik}\mathbb{U}_{jk} = 1 \mid n_{1k})$. Note that $\tilde{\zeta}_{3k} \le \tilde{\zeta}_{2k}$. Then we can calculate

$$\tilde{\zeta}_{2k} = P(\mathbb{U}_{ik}\mathbb{U}_{jk} = 1 \mid n_{1k}) = \frac{n_{1k}(n_{1k}-1)}{n_1(n_1-1)}$$

So

$$\begin{aligned}\frac{\tilde{\zeta}_{2k}}{n_{1k}} &= \frac{n_{1k}-1}{n_1(n_1-1)} \le \frac{1}{n_1} \\ \frac{\tilde{\zeta}_{2k}}{n_{1k}^2} &\le \frac{1}{n_1(n_1-1)}\frac{n_{1k}(n_{1k}-1)}{n_{1k}^2} \le \frac{2}{n_1^2}\end{aligned} \tag{5.27}$$

Lemma 5.4.6 bounds each of the terms in (5.26) conditional on fixing $\{n_{1k}\}_{k=1}^{m}$. It is proved later in this section.

Lemma 5.4.6. *We get that*

$$|E(S_{B1} \mid \{n_{1k}\})| \le \frac{2}{n_1}\operatorname{tr}(A)|B^1_{\cdot i}|_1 \qquad |E(S_{B2} \mid \{n_{1k}\})| \le \frac{2}{n_1}\operatorname{tr}(A)|B^1_{\cdot j}|_1 \tag{5.28}$$

Define

$$\tau_2 = CK^2 \log m \frac{1}{n_{1\min}} \|A\|_2 + CK^2 \frac{\log^{1/2}(m \vee n)}{n_{1\min}} \|A\|_F \tag{5.29}$$

Then, conditional on $\{n_{1k}\}_{k=1}^m$*, each of the following events holds with probability at least* $1 - 6n_1/(n \vee m)^c$ *for some constant* c*,*

$$|S_{B1} - ES_{B1}| \le n_1\tau_2 \qquad |S_{B2} - ES_{B2}| \le n_1\tau_2 \tag{5.30}$$

We can also show that

$$|E(S_{B3} \mid \{n_{1k}\})| \le \frac{2}{n_1^2} \operatorname{tr}(A) \operatorname{tr}(B^1) \tag{5.31}$$

and that with probability at least $1 - 6n_1/(n \vee m)^c$*,*

$$|S_{B3} - ES_{B3}| \le \frac{n_1}{n_{1\min}} \tau_2 \tag{5.32}$$

Using Lemma 5.4.6, we get that, conditional on $\{n_{1k}\}$, with probability at least $1 - 18n_1/(n \vee m)^c$,

$$\left| \hat{S}(B)_{ij} - \sum_{k=1}^m \mathbb{U}_{ik}\mathbb{U}_{jk}\mathbb{X}_{ik}\mathbb{X}_{jk} \right| \le \frac{4}{n_1} \operatorname{tr}(A) \max_{1 \le k \le n_1} |B^1_{\cdot k}|_1 + \frac{2}{n_1^2} \operatorname{tr}(A) \operatorname{tr}(B^1) + n_1\tau_2 \tag{5.33}$$

The following lemma helps us decondition on $\{n_{1k}\}$, which still remains in this bound through the terms involving $n_{1\min}$ inside τ_2.

Lemma 5.4.7. *Assume* $\zeta_{\min} \gtrsim \sqrt{\log(m \vee n)/n_1}$ *and* $\frac{1}{\zeta_{\min}}\sqrt{3/2}\sqrt{\log(n \vee m)/n_1} < 1/3$*. Then for any value* τ*, we can show that, with probability at least* $1 - 1/(n \vee m)^2$*,*

$$\frac{n_1}{n_{1\min}} \le C_3 \frac{1}{\zeta_{\min}}$$

Proof. Using Hoeffding's bound, we can get that, for some constant c,

$$\begin{aligned} &P\left(\frac{n_{1k}}{n_1} - \zeta_k \geq -\sqrt{3/2}\sqrt{\log(n \vee m)/n_1}\right) \geq 1 - 1/(n \vee m)^3 \\ \implies &P\left(\frac{n_{1\min}}{n_1} - \zeta_{\min} \geq -\sqrt{3/2}\sqrt{\log(n \vee m)/n_1}\right) \geq 1 - m/(n \vee m)^3 \end{aligned}$$

Rearranging, we get that this event also implies

$$\frac{\zeta_{\min}}{n_{1\min}/n_1} \leq \frac{1}{1 - \frac{1}{\zeta_{\min}}\sqrt{3/2}\sqrt{\log(n \vee m)/n_1}} \leq 1 + (3/2)^{3/2}\frac{1}{\zeta_{\min}}\sqrt{\frac{\log(m \vee n)}{n_1}}$$

where the last inequality is true for $\frac{1}{\zeta_{\min}}\sqrt{3/2}\sqrt{\log(n \vee m)/n_1} < 1/3$.

We can therefore rewrite

$$\begin{aligned} \frac{n_1}{n_{1\min}} &= \frac{n_1}{n_{1\min}} - \frac{1}{\zeta_{\min}} + \frac{1}{\zeta_{\min}} = \left(\frac{n_1}{n_{1\min}} - \frac{1}{\zeta_{\min}}\right) + \frac{1}{\zeta_{\min}} \\ &= \frac{1}{\zeta_{\min}}\left(\frac{\zeta_{\min}}{n_{1\min}/n_1} - 1\right) + \frac{1}{\zeta_{\min}} \\ &\leq C_3\frac{1}{\zeta_{\min}} \end{aligned}$$

Where the last inequality uses the assumption that $\zeta_{\min} \gtrsim \sqrt{\log(m \vee n)/n_1}$. □

Combining this lemma with (5.33), we get that with probability at least $1 - 1/(n \vee m)^2 - 18/(n \vee m)^2$,

$$\begin{aligned} \left|\hat{S}(B)_{ij} - \sum_{k=1}^{m} \mathbb{U}_{ik}\mathbb{U}_{jk}\mathbb{X}_{ik}\mathbb{X}_{jk}\right| \leq C_2\frac{\operatorname{tr}(A)}{n_1}\left(\max_{1 \leq k \leq n_1} |B^1_{\cdot k}|_1 + \frac{\operatorname{tr}(B^1)}{n_1}\right) \\ + CK^2\frac{\log m}{\zeta_{\min}}\|A\|_2 + CK^2\frac{\log^{1/2}(m \vee n)}{\zeta_{\min}}\|A\|_F \end{aligned}$$

Recall that for $i \neq j$, $\mathcal{M}_{ij} = \sum_{k=1}^m a_{kk}\zeta_k^2 \geq \zeta_{\min}^2 \operatorname{tr}(A)$. Now finally can now show that

$$\begin{aligned}
&\left| \frac{\hat{S}(B)_{ij}}{\sqrt{b_{ii}b_{jj}}\mathcal{M}_{ij}} - \rho_{ij}(B) \right| \\
&\quad \leq \frac{1}{\sqrt{b_{ii}b_{jj}}\mathcal{M}_{ij}} \left| \hat{S}(B)_{ij} - \sum_{k=1}^m \mathbb{U}_{ik}\mathbb{U}_{jk}\mathbb{X}_{ik}\mathbb{X}_{jk} \right| + \left| \frac{\sum_{k=1}^m \mathbb{U}_{ik}\mathbb{U}_{jk}\mathbb{X}_{ik}\mathbb{X}_{jk}}{\sqrt{b_{ii}b_{jj}}\mathcal{M}_{ij}} - \rho_{ij}(B) \right| \\
&\quad \leq C_2 \frac{1}{\zeta_{\min}^2 n_1} \left(\max_{1 \leq k \leq n_1} |B^1_{\cdot k}|_1 + \frac{\operatorname{tr}(B^1)}{n_1} \right) + CK^2 \frac{\log m}{\zeta_{\min}^3 \operatorname{tr}(A)} \|A\|_2 \\
&\qquad + CK^2 \log^{1/2}(m \vee n) \frac{\|A\|_F}{\zeta_{\min}^3 \operatorname{tr}(A)} + \alpha_{\text{offd}}
\end{aligned}$$

□

Proof of Lemma 5.4.6. We can rewrite

$$S_{B1} = \sum_{\ell=1}^{n_1} \sum_{k=1}^m \frac{1}{n_{1k}} \mathbb{U}_{\ell k}\mathbb{X}_{\ell k}\mathbb{U}_{jk}\mathbb{U}_{ik}\mathbb{X}_{ik} = \sum_{\ell=1}^{n_1} S_{B1,\ell}$$

Here we will bound the bias and deviation of each of the $S_{B1,\ell}$ terms separately using the Sparse Hanson-Wright results, and then combine them using union bounds. Note that there may be some looseness in this union bound, and that directly bounding S_{B1} using the techniques in *Zhou* (2019) may help us achieve tighter raters, but we leave this to future work.

Then we first can calculate the expectation.

$$E(S_{B1,\ell} \mid \{n_{1k}\}) = \begin{cases} b_{\ell i} \sum_{k=1}^m \frac{1}{n_{1k}} \tilde{\zeta}_{3k} a_{kk} & \text{if } \ell \neq i, j \\ b_{\ell i} \sum_{k=1}^m \frac{1}{n_{1k}} \tilde{\zeta}_{2k} a_{kk} & \text{otherwise} \end{cases}$$

So this is bounded by

$$|E(S_{B1,\ell} \mid \{n_{1k}\})| \leq |b_{\ell i}| \max_k \frac{\tilde{\zeta}_{2k}}{n_{1k}} \operatorname{tr}(A) \leq |b_{\ell i}| \frac{2}{n_1} \operatorname{tr}(A)$$

Which implies that

$$|E(S_{B1} \mid \{n_{1k}\})| \leq \frac{2}{n_1} \operatorname{tr}(A)|B^1_{\cdot i}|_1 \tag{5.34}$$

Now, for $\ell \neq i, j$, we can write

$$\begin{aligned} S_{B1,\ell} &= \sum_{k=1}^{m} \frac{1}{n_{1k}} \mathbb{U}_{\ell k}\mathbb{X}_{\ell k}\mathbb{U}_{jk}\mathbb{U}_{ik}\mathbb{X}_{ik} = (N^{1/2}\mathbb{X}_{\ell\cdot})^T \operatorname{diag}(\mathbb{U}_{\ell\cdot} \circ \mathbb{U}_{i\cdot} \circ \mathbb{U}_{j\cdot})(N^{1/2}\mathbb{X}_{i\cdot}) \\ &= g(N^{1/2}A^{1/2})^T \operatorname{diag}(\mathbb{U}_{\ell\cdot} \circ \mathbb{U}_{i\cdot} \circ \mathbb{U}_{j\cdot})(N^{1/2}A^{1/2})g^T \end{aligned}$$

Where $N = \operatorname{diag}(\{1/n_{1k}\}_k)$.

We then follow the same proof as the off-diagonal case in Theorem 5.3.6, replacing $D = \operatorname{diag}(\mathbb{U}_{\ell\cdot} \circ \mathbb{U}_{i\cdot} \circ \mathbb{U}_{j\cdot})$, $A_0^{1/2} D A_0^{1/2}$ with $(N^{1/2}A^{1/2})^T D (N^{1/2}A^{1/2})$, and A with $\tilde{A} = (\tilde{a}_{ij}) = A^{1/2}NA^{1/2}$. Note that when Theorems 5.3.4 and 5.3.5 are used, here we use the variants that allow for D_0 to be non-symmetric, as we have $D_0 = N^{1/2}A^{1/2}$.

We therefore get that

$$\begin{aligned} &P(|S_{B1,\ell} - ES_{B1,\ell}| > t \mid \{n_{1k}\}) \\ &\qquad \leq 6 \exp\left(-c_2 \min\left\{\frac{t^2}{K^4}\left(\sum_{k=1}^{m} \tilde{\zeta}_{3k}\tilde{a}_{kk}^2 + \sum_{h\neq k} \tilde{\zeta}_{3h}\tilde{\zeta}_{3k}\tilde{a}_{hk}^2\right)^{-1}, \frac{t}{K^2\|\tilde{A}\|_2}\right\}\right) \end{aligned}$$

For constant C, let

$$\tau' = CK^2 \log m \|\tilde{A}\|_2 + CK^2 \log^{1/2}(m \vee n) \left(\sum_{k=1}^{m} \tilde{\zeta}_{3k}\tilde{a}_{kk}^2 + \sum_{h\neq k} \tilde{\zeta}_{3h}\tilde{\zeta}_{3k}\tilde{a}_{hk}^2\right)^{1/2}$$

Which implies that, for some constant c,

$$P(|S_{B1,\ell} - ES_{B1,\ell}| > \tau' \mid \{n_{1k}\}) \leq \frac{6}{(n \vee m)^c} \tag{5.35}$$

When $\ell = i$ or j, we follow the same procedure expect with $D = \operatorname{diag}(\mathbb{U}_{i\cdot} \circ \mathbb{U}_{j\cdot})$, and therefore get the same result except for

$$\tau'' = CK^2 \log m \|\tilde{A}\|_2 + CK^2 \log^{1/2}(m \vee n) \left(\sum_{k=1}^{m} \tilde{\zeta}_{2k} \tilde{a}_{kk}^2 + \sum_{h \neq k} \tilde{\zeta}_{2h} \tilde{\zeta}_{2k} \tilde{a}_{hk}^2 \right)^{1/2} >= \tau'$$

We let $n_{1\min} = \min_k n_{1k}$ and define τ_2 as follows, where $\tau_2 \geq \tau', \tau''$.

$$\begin{aligned}
\tau'' &= CK^2 \log m \|\tilde{A}\|_2 + CK^2 \log^{1/2}(m \vee n) \left(\sum_{k=1}^{m} \tilde{\zeta}_{2k} \tilde{a}_{kk}^2 + \sum_{h \neq k} \tilde{\zeta}_{2h} \tilde{\zeta}_{2k} \tilde{a}_{hk}^2 \right)^{1/2} \\
&\leq CK^2 \log m \frac{1}{n_{1\min}} \|A\|_2 + CK^2 \log^{1/2}(m \vee n) \left(\sum_{k=1}^{m} \frac{1}{n_{1\min}^2} a_{kk}^2 + \sum_{h \neq k} \frac{1}{n_{1\min}^2} a_{hk}^2 \right)^{1/2} \\
&\leq CK^2 \log m \frac{1}{n_{1\min}} \|A\|_2 + CK^2 \frac{\log^{1/2}(m \vee n)}{n_{1\min}} \|A\|_F =: \tau_2 \qquad (5.36)
\end{aligned}$$

Combining these with a union bound over $\ell = 1, \ldots, n_1$,

$$P(|S_{B1} - ES_{B1}| < n_1 \tau_2 \mid \{n_{1k}\}) \geq 1 - \sum_{\ell=1}^{n_1} P(|S_{B1,\ell} - ES_{B1,\ell}| > \tau_2 \mid \{n_{1k}\}) \geq 1 - \frac{6n_1}{(n \vee m)^c}.$$

So, conditional on $\{n_{1k}\}$, with probability at least $1 - 6n_1/(n \vee m)^c$, we get that $|S_{B1} - ES_{B1}| \leq n_1 \tau_2$.

S_{B2} is symmetric with S_{B1}, so we can use the same arguments as above for it.

For the last term, we can write

$$S_{B3} = \sum_{\ell=1}^{n_1} \sum_{k=1}^{m} \frac{1}{n_{1k}^2} \mathbb{U}_{\ell k} \mathbb{U}_{jk} \mathbb{U}_{ik} \mathbb{X}_{\ell k}^2 = \sum_{\ell=1}^{n_1} S_{B3,\ell}$$

Then

$$E(S_{B3,\ell} \mid \{n_{1k}\}) = \begin{cases} b_{\ell\ell} \sum_{k=1}^{m} \frac{1}{n_{1k}^2} \tilde{\zeta}_{3k} a_{kk} & \text{if } \ell \neq i, j \\ b_{\ell\ell} \sum_{k=1}^{m} \frac{1}{n_{1k}^2} \tilde{\zeta}_{2k} a_{kk} & \text{otherwise} \end{cases}$$

So

$$|E(S_{B3,\ell} \mid \{n_{1k}\})| \leq b_{\ell\ell} \frac{2}{n_1^2} \operatorname{tr}(A)$$

Which implies that

$$|E(S_{B3} \mid \{n_{1k}\})| \leq \frac{2}{n_1^2} \operatorname{tr}(A) \operatorname{tr}(B^1)$$

Using the same concentration arguments as above, we can show that conditional on $\{n_{1k}\}$, with probability at least $1 - 6n_1/(n \vee m)^c$,

$$|S_{B3} - ES_{B3}| \leq \frac{n_1}{n_{1\min}} \tau_2$$

□

Proof of Proposition 5.4.3. Without loss of generality, assume i is in group 1. Again, define the standard basis vectors $e_i' = (\mathbb{1}_{k=i})_{k=1}^n \in \mathbb{R}^n$ and $e_i = (\mathbb{1}_{k=i})_{k=1}^{n_1} \in \mathbb{R}^{n_1}$. Then we get that

$$\begin{aligned}
\hat{S}(B)_{ii} &= \sum_{k=1}^m \mathbb{U}_{ik}(e_i' - P_{i\cdot}^k)^T \mathbb{X}_{\cdot k} \mathbb{X}_{\cdot k}^T (e_i' - P_{i\cdot}^k) \\
&= \sum_{k=1}^m \mathbb{U}_{ik} \mathbb{X}_{ik}^2 - 2 \sum_{k=1}^m \mathbb{U}_{ik} \frac{1}{n_{1k}} (e_i^T X_{\cdot k}^1 (X_{\cdot k}^1)^T U_{\cdot k}^1) \\
&\quad + \sum_{k=1}^m \mathbb{U}_{ik} \frac{1}{n_{1k}^2} ((U_{\cdot k}^1)^T X_{\cdot k}^1 (X_{\cdot k}^1)^T U_{\cdot k}^1) \\
&= \sum_{k=1}^m \mathbb{U}_{ik} \mathbb{X}_{ik}^2 - 2S_{B4} + S_{B5} \qquad (5.37)
\end{aligned}$$

Rearranging,

$$\left| \hat{S}(B)_{ii} - \sum_{k=1}^m \mathbb{U}_{ik} \mathbb{X}_{ik}^2 \right| \leq 2|ES_{B4}| + 2|S_{B4} - ES_{B4}| + |ES_{B5}| + |S_{B5} - ES_{B5}| \qquad (5.38)$$

From here, we follow very similar steps to those taken in the proof of Lemma 5.4.6, some of which we omit for bevity. We first note that $P(U_{ik} = 1 \mid n_{1k}) = n_{1k}/n_1$. Then we can rewrite

$$S_{B4} = \sum_{\ell=1}^{n_1} \sum_{k=1}^m \frac{1}{n_{1k}} \mathbb{U}_{\ell k} \mathbb{X}_{\ell k} \mathbb{U}_{ik} \mathbb{X}_{ik} = \sum_{\ell=1}^{n_1} S_{B4,\ell}$$

Note that this is identical to the decomposition of S_{B1}, except that since $i = j$ we have that $\mathbb{U}_{jk}\mathbb{U}_{ik} = \mathbb{U}_{ik}$. So following the same steps as the bound of $E(S_{B1} \mid \{n_{1k}\})$, we get the same upper bound on the expectation

$$|E(S_{B4} \mid \{n_{1k})\}| \leq \frac{2}{n_1}\operatorname{tr}(A)|B^1_{\cdot i}|_1$$

We also follow the same steps as those used to bound $|S_{B1} - ES_{B1}|$, replacing $D = \operatorname{diag}(\mathbb{U}_{\ell\cdot} \circ \mathbb{U}_{i\cdot})$ when $\ell \neq i$ and $D = \operatorname{diag}(\mathbb{U}_{i\cdot})$ when $\ell = i$. This results in the same upper bound of $|S_{B4} - ES_{B4}| \leq n_1\tau_2$ with probability at least $1 - 6n_1/(n \vee m)^c$.

S_{B5} shares a similar relationship with S_{B3}, so we again follow the same steps. This results in bounds of

$$|E(S_{B5} \mid \{n_{1k}\})| \leq \frac{2}{n_1 n_{1\,\mathrm{min}}}\operatorname{tr}(A)\operatorname{tr}(B^1)$$

(noting the difference of a factor of $n_1/n_{1\,\mathrm{min}}$) and

$$|S_{B3} - ES_{B3}| \leq \frac{n_1}{n_{1\,\mathrm{min}}}\tau_2$$

So, following the same steps as in the proof of Proposition 5.4.2, we can therefore show that

$$\begin{aligned}\left|\frac{\hat{S}(B)_{ii}}{b_{ii}\mathcal{M}_{ii}} - 1\right| &\leq \frac{1}{b_{ii}\mathcal{M}_{ii}}\left|\hat{S}(B)_{ii} - \sum_{k=1}^{m}\mathbb{U}_{ik}\mathbb{X}_{ik}^2\right| + \left|\frac{\sum_{k=1}^{m}\mathbb{U}_{ik}\mathbb{X}_{ik}^2}{b_{ii}\mathcal{M}_{ii}} - 1\right| \\ &\leq \frac{C_2}{\zeta_{\min}^2 n_1}\left(\max_{1\leq k\leq n_1}|B^1_{\cdot k}|_1 + \frac{1}{\zeta_{\min}}\frac{\operatorname{tr}(B^1)}{n_1}\right) \\ &\quad + CK^2\frac{\log m}{\zeta_{\min}^3\operatorname{tr}(A)}\|A\|_2 + CK^2\log^{1/2}(m \vee n)\frac{\|A\|_F}{\zeta_{\min}^3\operatorname{tr}(A)} + \alpha_{\mathrm{offd}}\end{aligned}$$

□

Proof of Proposition 5.4.4. Without loss of generality assume i is in group 1 and j is in group 2. Define the standard basis vectors $e'_i = (\mathbb{1}_{k=i})_{k=1}^{n} \in \mathbb{R}^n$, $e^1_i = (\mathbb{1}_{k=i})_{k=1}^{n_1} \in \mathbb{R}^{n_1}$, and

$e_i^2 = (\mathbb{1}_{k=i})_{k=1}^{n_2} \in \mathbb{R}^{n_2}$ Then we can rewrite

$$
\begin{aligned}
\hat{S}(B)_{ij} &= \sum_{k=1}^{m} \mathbb{U}_{ik}\mathbb{U}_{jk}(e'_i - P^k_{i\cdot})^T \mathbb{X}_{\cdot k}\mathbb{X}^T_{\cdot k}(e'_j - P^k_{j\cdot}) \\
&= \sum_{k=1}^{m} \mathbb{U}_{ik}\mathbb{U}_{jk}\mathbb{X}_{ik}\mathbb{X}_{jk} - \sum_{k=1}^{m} \mathbb{U}_{ik}\mathbb{U}_{jk}\frac{1}{n_{2k}}(e_i^{2T}X^2_{\cdot k}(X^2_{\cdot k})^T U^2_{\cdot k}) \\
&\quad - \sum_{k=1}^{m} \mathbb{U}_{ik}\mathbb{U}_{jk}\frac{1}{n_{1k}}(e_j^{1T}X^1_{\cdot k}(X^1_{\cdot k})^T U^1_{\cdot k}) \\
&= \sum_{k=1}^{m} \mathbb{U}_{ik}\mathbb{U}_{jk}\mathbb{X}_{ik}\mathbb{X}_{jk} - S_{B6} - S_{B7} \qquad (5.39)
\end{aligned}
$$

Note that S_{B6} and S_{B7} are identical to S_{B1} when $g = 2, 1$, respectively. So we can use the steps and results in Proposition 5.4.2 and Lemma 5.4.6 to immediately obtain our result. □

5.4.3 Mask estimation with demeaning

Define

$$\tilde{S}_c = \frac{n}{n-1}\,\mathrm{tr}(\tilde{\mathcal{X}}^T\tilde{\mathcal{X}} \circ \widehat{M}) - \frac{1}{n-1}\,\mathrm{tr}(\tilde{\mathcal{X}}^T\tilde{\mathcal{X}}) \tag{5.1}$$

Where, since we are using the demeaned estimators here, we place $\mathcal{X}$ with the demeaned $\tilde{\mathcal{X}}$ in the standard definition of S_c.

Proposition 5.4.8. *Under the events in Proposition 5.3.11 and Proposition 5.4.3, we get that*

$$\left|\frac{\mathrm{tr}(\tilde{\mathcal{X}}^T\tilde{\mathcal{X}})}{\mathrm{tr}(B_0)\sum_{i=1}^{m}\zeta_i a_{ii}} - 1\right| \leq \alpha_{\mathrm{mean}}$$

and

$$\left|\frac{\tilde{S}_c}{\mathrm{tr}(B_0)\sum_{i=1}^{n}\zeta_i^2 a_{ii}} - 1\right| \leq \frac{1}{\zeta_{\min}}\alpha_{\mathrm{mean}}$$

Proof. First note that $\mathrm{tr}((\mathbb{U}\circ\mathbb{X})^T(\mathbb{U}\circ\mathbb{X}))$ is equivalent to estimator studied in Proposition 5.3.11, where there is no mean. Then we can use the triangle inequality to decompose the excess error of $\mathrm{tr}(\tilde{\mathcal{X}}^T\tilde{\mathcal{X}})$ to the sum of errors of the diagonal terms.

$$
\begin{aligned}
|\mathrm{tr}(\tilde{\mathcal{X}}^T\tilde{\mathcal{X}}) - \mathrm{tr}((\mathbb{U}\circ\mathbb{X})^T(\mathbb{U}\circ\mathbb{X}))| &= \left|\sum_{i=1}^{n} \hat{S}(B)_{ii} - \sum_{k=1}^{m} \mathbb{U}_{ik}\mathbb{X}_{ik}^2\right| \\
&\leq \sum_{i=1}^{n}\left|\hat{S}(B)_{ii} - \sum_{k=1}^{m} \mathbb{U}_{ik}\mathbb{X}_{ik}^2\right| \leq \mathrm{tr}(B_0)\sum_{i=1}^{m}\zeta_i a_{ii}\alpha_{\mathrm{mean}}
\end{aligned}
$$

Where the last inequality uses Proposition 5.4.3 since the second-to-last term is the same as the LHS of (5.38). We can therefore use Proposition 5.3.11 to show that

$$
\begin{aligned}
&|\mathrm{tr}(\tilde{\mathcal{X}}^T\tilde{\mathcal{X}}) - \mathrm{tr}(B_0)\sum_{i=1}^{m}\zeta_i a_{ii}| \\
&\qquad \leq |\mathrm{tr}(\tilde{\mathcal{X}}^T\tilde{\mathcal{X}}) - \mathrm{tr}((\mathbb{U}\circ\mathbb{X})^T(\mathbb{U}\circ\mathbb{X}))| + |\mathrm{tr}((\mathbb{U}\circ\mathbb{X})^T(\mathbb{U}\circ\mathbb{X})) - \mathrm{tr}(B_0)\sum_{i=1}^{m}\zeta_i a_{ii}| \\
&\qquad \leq \mathrm{tr}(B_0)\sum_{i=1}^{m}\zeta_i a_{ii}\alpha_{\mathrm{mean}} + \sum_{i=1}^{m}\zeta_i a_{ii}\|B_0\|_F\alpha
\end{aligned}
$$

Implying that

$$
\left|\frac{\mathrm{tr}(\tilde{\mathcal{X}}^T\tilde{\mathcal{X}})}{\mathrm{tr}(B_0)\sum_{i=1}^{m}\zeta_i a_{ii}} - 1\right| \leq \alpha_{\mathrm{mean}} + \alpha\frac{\|B_0\|_F}{\mathrm{tr}(B_0)} \leq c\alpha_{\mathrm{mean}}
$$

For $\tilde{S}_c$, we will bound the error of $\mathrm{tr}(\tilde{\mathcal{X}}^T\tilde{\mathcal{X}})\circ\widehat{M}$ by the error of $\mathrm{tr}(\tilde{\mathcal{X}}^T\tilde{\mathcal{X}})$. To do this, note that

$$
\begin{aligned}
\operatorname{tr}(\tilde{\mathcal{X}}^T\tilde{\mathcal{X}}) &= \sum_{k=1}^m \tilde{\mathcal{X}}_{\cdot k}^T\tilde{\mathcal{X}}_{\cdot k} \\
&= \sum_{k=1}^m (\mathbb{U}_{\cdot k} \circ (I - P^k)\mathcal{X}_{\cdot k})^T(\mathbb{U}_{\cdot k} \circ (I - P^k)\mathcal{X}_{\cdot k}) \\
&= \sum_{k=1}^m (\mathbb{U}_{\cdot k} \circ (I - P^k)\mathbb{X}_{\cdot k})^T(\mathbb{U}_{\cdot k} \circ (I - P^k)\mathbb{X}_{\cdot k}) \\
&= \sum_{k=1}^m (\mathbb{U}_{\cdot k} \circ \mathbb{X}_{\cdot k} - \mathbb{U}_{\cdot k} \circ P^k\mathbb{X}_{\cdot k})^T(\mathbb{U}_{\cdot k} \circ \mathbb{X}_{\cdot k} - \mathbb{U}_{\cdot k} \circ P^k\mathbb{X}_{\cdot k}) \\
&= \operatorname{tr}((\mathbb{U} \circ \mathbb{X})^T(\mathbb{U} \circ \mathbb{X})) \\
&\qquad - 2\sum_{k=1}^m (\mathbb{U}_{\cdot k} \circ P^k\mathbb{X}_{\cdot k})^T(\mathbb{U}_{\cdot k} \circ \mathbb{X}_{\cdot k}) + \sum_{k=1}^m (\mathbb{U}_{\cdot k} \circ P^k\mathbb{X}_{\cdot k})^T(\mathbb{U}_{\cdot k} \circ P^k\mathbb{X}_{\cdot k}) \\
&= \operatorname{tr}((\mathbb{U} \circ \mathbb{X})^T(\mathbb{U} \circ \mathbb{X})) \\
&\qquad - \sum_{k=1}^m \frac{1}{n_{1k}} \left(\sum_{\ell=1}^{n_1} \mathbb{U}_{\ell k}^1\mathbb{X}_{\ell k}^1\right)^2 - \sum_{k=1}^m \frac{1}{n_{2k}} \left(\sum_{\ell=1}^{n_2} \mathbb{U}_{\ell k}^2\mathbb{X}_{\ell k}^2\right)^2
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\operatorname{tr}(\tilde{\mathcal{X}}^T\tilde{\mathcal{X}} \circ \widehat{M}) &= \sum_{k=1}^m \tilde{\mathcal{X}}_{\cdot k}^T\tilde{\mathcal{X}}_{\cdot k}\hat{\zeta}_k \\
&= \operatorname{tr}((\mathbb{U} \circ \mathbb{X})^T(\mathbb{U} \circ \mathbb{X}) \circ \widehat{M}) \\
&\qquad - \sum_{k=1}^m \frac{\hat{\zeta}_k}{n_{1k}} \left(\sum_{\ell=1}^{n_1} \mathbb{U}_{\ell k}^1\mathbb{X}_{\ell k}^1\right)^2 - \sum_{k=1}^m \frac{\hat{\zeta}_k}{n_{2k}} \left(\sum_{\ell=1}^{n_2} \mathbb{U}_{\ell k}^2\mathbb{X}_{\ell k}^2\right)^2
\end{aligned}
$$

and therefore

$$\begin{aligned}
&|\mathrm{tr}(\tilde{\mathcal{X}}^T\tilde{\mathcal{X}} \circ \widehat{M}) - \mathrm{tr}((\mathbb{U} \circ \mathbb{X})^T(\mathbb{U} \circ \mathbb{X}) \circ \widehat{M})| \\
&\qquad = \sum_{k=1}^{m} \frac{\hat{\zeta}_k}{n_{1k}} \left(\sum_{\ell=1}^{n_1} \mathbb{U}^1_{\ell k}\mathbb{X}^1_{\ell k} \right)^2 + \sum_{k=1}^{m} \frac{\hat{\zeta}_k}{n_{2k}} \left(\sum_{\ell=1}^{n_2} \mathbb{U}^2_{\ell k}\mathbb{X}^2_{\ell k} \right)^2 \\
&\qquad \le \sum_{k=1}^{m} \frac{1}{n_{1k}} \left(\sum_{\ell=1}^{n_1} \mathbb{U}^1_{\ell k}\mathbb{X}^1_{\ell k} \right)^2 + \sum_{k=1}^{m} \frac{1}{n_{2k}} \left(\sum_{\ell=1}^{n_2} \mathbb{U}^2_{\ell k}\mathbb{X}^2_{\ell k} \right)^2 \\
&\qquad = |\mathrm{tr}(\tilde{\mathcal{X}}\tilde{\mathcal{X}}^T) - \mathrm{tr}((\mathbb{U} \circ \mathbb{X})^T(\mathbb{U} \circ \mathbb{X}))|
\end{aligned}$$

So if we let S_c be the oracle estimator with perfect demaning (as studied in Section 5.3.3), we can show that

$$\begin{aligned}
|\tilde{S}_c - S_c| &\le \frac{n}{n-1} |\mathrm{tr}(\tilde{\mathcal{X}}^T\tilde{\mathcal{X}} \circ \widehat{M}) - \mathrm{tr}((\mathbb{U} \circ \mathbb{X})^T(\mathbb{U} \circ \mathbb{X}) \circ \widehat{M})| \\
&\quad - \frac{1}{n-1} |\mathrm{tr}(\tilde{\mathcal{X}}^T\tilde{\mathcal{X}}) - \mathrm{tr}((\mathbb{U} \circ \mathbb{X})^T(\mathbb{U} \circ \mathbb{X}))| \\
&\le |\mathrm{tr}(\tilde{\mathcal{X}}^T\tilde{\mathcal{X}}) - \mathrm{tr}((\mathbb{U} \circ \mathbb{X})^T(\mathbb{U} \circ \mathbb{X}))| \le \mathrm{tr}(B_0) \sum_{i=1}^{m} \zeta_i a_{ii} \alpha_{\mathrm{mean}}
\end{aligned}$$

Using Proposition 5.3.11, we can therefore show that

$$\left| \frac{\tilde{S}_c}{\mathrm{tr}(B_0) \sum_{i=1}^n \zeta_i^2 a_{ii}} - 1 \right| \le \frac{\sum_{i=1}^n \zeta_i a_{ii}}{\sum_{i=1}^n \zeta_i^2 a_{ii}} \alpha_{\mathrm{mean}} + \alpha \le \frac{1}{\zeta_{\min}} \alpha_{\mathrm{mean}}$$

□

5.5 Estimation of A with unknown group means

For convenience we restate our data model and some necessary notation for this section. Assume our rows are sorted by group. Denote

$$\mathbb{U} = \begin{bmatrix} \mathbb{U}^1 \\ \mathbb{U}^2 \end{bmatrix}, \quad \mathbb{X} = \begin{bmatrix} \mathbb{X}^1 \\ \mathbb{X}^2 \end{bmatrix}, \quad \text{where } \mathbb{U}^1, \mathbb{X}^1 \in \mathbb{R}^{n_1 \times m} \text{ and } \mathbb{U}^2, \mathbb{X}^2 \in \mathbb{R}^{n_2 \times m}$$

Table 5.2: Notation

Notation	Meaning
Sample Sizes	
$n = n_1 + n_2$	Total number of rows
n_1, n_2	Number of rows per group
$n_{1k} = \sum_{i=1}^{n_1} \mathbb{U}_{ik}, n_{2k} = \sum_{i=n_1+1}^{n} \mathbb{U}_{ik}$	Number of observed values per group for column k
$n_{1k\ell} = \sum_{i=1}^{n_1} \mathbb{U}_{ik}\mathbb{U}_{i\ell}, n_{2k\ell} = \sum_{i=n_1+1}^{n} \mathbb{U}_{ik}\mathbb{U}_{i\ell}$	Number of times cols k, ℓ are both observed per group
$n_{1\min} = \min_k n_{1k}, n_{2\min} = \min_k n_{2k}$	Minimum number of observed values per group
$\hat{\zeta}_k = (n_{1k} + n_{2k})/n$	Estimated sampling probability in column k
Matrices	
A	Column-wise covariance
$B = \begin{pmatrix} B^1 & B^{12} \\ B^{21} & B^2 \end{pmatrix}$	Row-wise covariance
$B_{i\cdot}, B_{\cdot j}$	ith row and jth column of B, respectively
P^j, defined in (5.40)	Proj matrix, calculates observed group means for col j

Let n_{1j}, n_{2j} be the number of observed values for variables j in group 1 and 2, respectively. Then we note that

$$\tilde{\mathcal{X}}_{\cdot j} = u^j \circ (I - P^j)\mathcal{X}_{\cdot j} \qquad \text{where } P^j = \begin{bmatrix} \frac{1}{n_{1j}}\vec{1}_{n_1}\mathbb{U}_{\cdot j}^{1T} & 0 \\ 0 & \frac{1}{n_{2j}}\vec{1}_{n_2}\mathbb{U}_{\cdot j}^{2T} \end{bmatrix} \tag{5.40}$$

and that $(I - P^j)\mathcal{X}_{\cdot j} = (I - P^j)x^j$.

Recall that for some absolute constants C, C_2 we define

$$\beta = CK^2 \frac{\log n \|B_0\|_2}{\zeta_{\min} \operatorname{tr}(B_0)} + CK^2 \frac{\log^{1/2}(m \vee n)}{\sqrt{n}} \frac{1}{\zeta_{\min}} \frac{\sqrt{n}\|B_0\|_F}{\operatorname{tr}(B_0)} \tag{3.19}$$

and

$$\beta_{\text{mean}} = C_2 \frac{1}{\zeta_{\min} \operatorname{tr}(B_0)} \left(\frac{\operatorname{tr}(B^1)}{n_1} + \frac{\operatorname{tr}(B^2)}{n_2} \right) + C_2 \frac{1}{\zeta_{\min} \operatorname{tr}(B_0)} \left(\frac{1}{n_1} \left| \sum_{k \neq \ell} B^1_{k\ell} \right| + \frac{1}{n_2} \left| \sum_{k \neq \ell} B^2_{k\ell} \right| \right)$$
$$+ C_2 K^2 \frac{\log n}{\zeta_{\min}^2} \frac{\|B^1\|_2 + \|B^2\|_2}{\operatorname{tr}(B_0)} + C_2 K^2 \frac{\log^{1/2}(m \vee n)}{\zeta_{\min}^{5/2}} \frac{\|B^1\|_F + \|B^2\|_F}{\operatorname{tr}(B_0)} + \beta \tag{3.21}$$

Proofs of Theorems 3.2.1 and 3.2.2. Just as in Section 5.5, to prove the A-side of these results, we simply apply Theorem 5.2.1 using the correlation convergence results in Theorem 5.5.1. □

5.5.1 Correlation results

Theorem 5.5.1. *Consider data generating random matrices as in* (3.1) *and* (3.2)*. Define* $\zeta_{\min} = \min_{j=1,\ldots,m} \zeta_j$*. Assume* $n_1, n_2 \equiv n$ *and* $\zeta_{\min} \gtrsim \sqrt{\frac{\log(m \vee n)}{n}}$*. Let* C *be some absolute constant, and define*

$$\beta_{\text{mean}} = C \frac{1}{\zeta_{\min} \operatorname{tr}(B_0)} \left(\frac{\operatorname{tr}(B^1)}{n_1} + \frac{\operatorname{tr}(B^2)}{n_2} \right) + C \frac{1}{\zeta_{\min} \operatorname{tr}(B_0)} \left(\frac{1}{n_1} \left| \sum_{k \neq \ell} B^1_{k\ell} \right| + \frac{1}{n_2} \left| \sum_{k \neq \ell} B^2_{k\ell} \right| \right)$$
$$+ C K^2 \frac{\log n}{\zeta_{\min}^2} \frac{\|B^1\|_2 + \|B^2\|_2}{\operatorname{tr}(B_0)} + C K^2 \frac{\log^{1/2}(m \vee n)}{\zeta_{\min}^{5/2}} \frac{\|B^1\|_F + \|B^2\|_F}{\operatorname{tr}(B_0)} + \beta \tag{3.21}$$

Let $m \vee n \geq 3$*. Then, with probability at least* $1 - 26/(m \vee n)^2$*, for* $\beta_{\text{mean}} < 1/3$*, and* $\widetilde{\Gamma}(A_0)$ *as defined in* (3.7)*, we get that*

$$\forall i \neq j, \quad |\widetilde{\Gamma}_{ij}(A_0) - \rho_{ij}(A_0)| \leq 3\beta_{\text{mean}}$$

Similarly, with probability at least $1 - 26/(m \vee n)^2$ *and* $\widehat{\Gamma}(A_0)$ *as defined in* (3.14)*, we get that*

$$\forall i \neq j, \quad |\widehat{\Gamma}_{ij}(A_0) - \rho_{ij}(A_0)| \leq 12\beta + 3\beta_{\text{mean}}.$$

Proof. Using Proposition 5.5.2 the proof of the first statement proceeds identically to the proof of Theorem 5.3.2.

From the proof of Theorem 5.3.3 we know that, with probability at least $1 - 2/(m \vee n)^2$,

$$\left| \frac{\sqrt{\zeta_i \zeta_j}}{\sqrt{\hat{\zeta}_i \hat{\zeta}_j}} - 1 \right| \leq 6\beta \tag{5.41}$$

And therefore, using Proposition 5.5.2, we get that with high probability,

$$\begin{aligned} |\hat{\Gamma}_{ij}(A_0) - \rho_{ij}(A_0)| &\leq |\hat{\Gamma}_{ij}(A_0) - \tilde{\Gamma}_{ij}(A_0)| + |\tilde{\Gamma}_{ij}(A_0) - \rho_{ij}(A_0)| \\ &\leq \tilde{\Gamma}_{ij}(A_0) \left| \frac{\sqrt{\zeta_i \zeta_j}}{\sqrt{\hat{\zeta}_i \hat{\zeta}_j}} - 1 \right| + 3\beta_{\text{mean}} \\ &\leq (1 + 3\beta_{\text{mean}}) 6\beta + 3\beta_{\text{mean}} \\ &= 6\beta + 3\beta_{\text{mean}} + 18\beta\beta_{\text{mean}} \leq 12\beta + 3\beta_{\text{mean}} \end{aligned}$$

□

5.5.2 Covariance concentration

Consider the the (i, j)-th entry of $\hat{S}(A)$.

$$\begin{aligned} \hat{S}(A)_{ij} &= \mathcal{X}_{\text{cen},i}^T \mathcal{X}_{\text{cen},j} \\ &= (\mathcal{X}_i^T (I - P^i)^T \circ u^{iT})(u^j \circ (I - P^j)\mathcal{X}_j) \\ &= x^{iT} (I - P^i)^T \text{diag}(u^i \circ u^j)(I - P^j) x^j \end{aligned} \tag{5.42}$$

Proposition 5.5.2. *Assume* $\zeta_{\min} \gtrsim \sqrt{\log(m \vee n)/n_{\min}}$ *and* $\frac{1}{\zeta_{\min}} \sqrt{3/2} \sqrt{\log(n \vee m)/n_{\min}} < 1/3$. *Recall that* $\mathcal{N}_{jj} = \text{tr}(B_0)\zeta_j$ *for all* j *and* $\mathcal{N}_{ij} = \zeta_i \zeta_j \, \text{tr}(B_0)$ *for all* $i \neq j$. *Then with* *probability at least* $1-26/(n\vee m)^2$*, the following event holds simultaneously for all* i,j*.*

$$\begin{aligned}
&\left|\frac{\hat{S}(A)_{ij}}{\sqrt{a_{ii}a_{jj}}\mathcal{N}_{ij}}-\rho_{ij}(A)\right| \\
&\quad\leq C\frac{1}{\zeta_{\min}\operatorname{tr}(B_0)}\left(\frac{\operatorname{tr}(B^1)}{n_1}+\frac{\operatorname{tr}(B^2)}{n_2}\right)+C\frac{1}{\zeta_{\min}\operatorname{tr}(B_0)}\left(\frac{1}{n_1}\left|\sum_{k\neq\ell}B^1_{k\ell}\right|+\frac{1}{n_2}\left|\sum_{k\neq\ell}B^2_{k\ell}\right|\right) \\
&\qquad+CK^2\frac{\log n}{\zeta_{\min}^2}\frac{\|B^1\|_2+\|B^2\|_2}{\operatorname{tr}(B_0)}+CK^2\frac{\log^{1/2}(m\vee n)}{\zeta_{\min}^{5/2}}\frac{\|B^1\|_F+\|B^2\|_F}{\operatorname{tr}(B_0)}+\beta:=\beta_{\text{mean}}
\end{aligned}$$

Proof. For now we consider the matrix from the center of the quadratic form (5.42). We can write

$$(I-P^i)^T\operatorname{diag}(u^i\circ u^j)(I-P^j)=\operatorname{diag}(u^i\circ u^j)-\begin{bmatrix}D^{1ij} & 0\\ 0 & D^{2ij}\end{bmatrix}$$

where

$$\begin{aligned}
D^{1ij}&=-\operatorname{diag}(u^{1i}\circ u^{1j})\left(\frac{1}{n_{1j}}\vec{1}_{n_1}u^{1jT}\right)-\left(\frac{1}{n_{1i}}\vec{1}_{n_1}u^{1iT}\right)^T\operatorname{diag}(u^{1i}\circ u^{1j}) \\
&\quad+\left(\frac{1}{n_{1i}}\vec{1}_{n_1}u^{1iT}\right)^T\operatorname{diag}(u^{1i}\circ u^{1j})\left(\frac{1}{n_{1j}}\vec{1}_{n_1}u^{1jT}\right) \\
D^{2ij}&=-\operatorname{diag}(u^{2i}\circ u^{2j})\left(\frac{1}{n_{2j}}\vec{1}_{n_2}u^{2jT}\right)-\left(\frac{1}{n_{2i}}\vec{1}_{n_2}u^{1iT}\right)^T\operatorname{diag}(u^{2i}\circ u^{2j}) \\
&\quad+\left(\frac{1}{n_{2i}}\vec{1}_{n_2}u^{1iT}\right)^T\operatorname{diag}(u^{2i}\circ u^{2j})\left(\frac{1}{n_{2j}}\vec{1}_{n_2}u^{2jT}\right)
\end{aligned}$$

We then get that, for $k\neq\ell$,

$$\begin{aligned}
D^{1ij}_{kk}&=\frac{n_{1ij}-n_{1i}-n_{1j}}{n_{1i}n_{1j}}u^{1i}_k u^{1j}_k \\
D^{1ij}_{k\ell}&=\frac{n_{1ij}}{n_{1i}n_{1j}}u^{1i}_k u^{1j}_\ell-\frac{1}{n_{1i}}u^{1i}_\ell u^{1i}_k u^{1j}_\ell-\frac{1}{n_{1j}}u^{1j}_k u^{1i}_k u^{1j}_\ell
\end{aligned}$$

So for $i \neq j$, we can decompose our excess error as

$$\begin{aligned}\hat{S}(A)_{ij} - x^{iT}\text{diag}(u^i \circ u^j)x^j &= \sum_{k=1}^{n_1} D_{kk}^{1ij} x_k^{1i} x_k^{1j} + \sum_{k \neq \ell} D_{k\ell}^{1ij} x_k^{1i} x_\ell^{1j} \\ &\quad + \sum_{k=1}^{n_2} D_{kk}^{2ij} x_k^{2i} x_k^{2j} + \sum_{k \neq \ell} D_{k\ell}^{2ij} x_k^{2i} x_\ell^{2j} \\ &= S_{A1}^1 + S_{A2}^1 + S_{A1}^2 + S_{A2}^2\end{aligned}$$

So

$$\begin{aligned}|\hat{S}(A)_{ij} - x^{iT}\text{diag}(u^i \circ u^j)x^j| &\leq |S_{A1}^1| + |S_{A2}^1 - ES_{A2}^1| + |ES_{A2}^1| \\ &\quad + |S_{A1}^2| + |S_{A2}^2 - ES_{A2}^2| + |ES_{A2}^2|\end{aligned} \tag{5.43}$$

We first consider S_{A1}^1 and S_{A1}^2.

$$S_{A1}^1 = \sum_{k=1}^{n_1} D_{kk}^{1ij} x_k^{1i} x_k^{1j} = \frac{n_{1ij} - n_{1i} - n_{1j}}{n_{1i} n_{1j}} \sum_{k=1}^{n_1} u_k^{1i} u_k^{1j} x_k^{1i} x_k^{1j} \tag{5.44}$$

Lemma 5.5.3. *For $g = 1, 2$, we can bound, for all $i \neq j$ simultaneously,*

$$\begin{aligned}|S_{A1}^g| &\leq C(\zeta_i + \zeta_j)\frac{tr(B^g)}{n_g}|a_{ij}| + CK^2 \frac{1}{n_g(\zeta_i \wedge \zeta_j)} \log n \|B^g\|_2 \sqrt{a_{ii}a_{jj}} \\ &\quad + CK^2 \frac{1}{n_g\sqrt{\zeta_i \wedge \zeta_j}} \log^{1/2}(m \vee n)\|B^g\|_F \sqrt{a_{ii}a_{jj}}\end{aligned}$$

with probability at least $1 - 7/(n \vee m)^2$.

Proof. Without loss of generality assume $g = 1$. From Lemma 5.3.8, we get that

$$\begin{aligned}&\left|\sum_{k=1}^{n_1} u_k^{1i} u_k^{1j} x_k^{1i} x_k^{1j} - E\sum_{k=1}^{n_1} u_k^{1i} u_k^{1j} x_k^{1i} x_k^{1j}\right| \\ &\qquad \leq CK^2 \log n \|B^1\|_2 \sqrt{a_{ii}a_{jj}} + CK^2 \log^{1/2}(m \vee n)\|B^1\|_F \sqrt{\zeta_i\zeta_j}\sqrt{a_{ii}a_{jj}}\end{aligned}$$

Note that, modifying the proof of Lemma 5.4.7 slightly, we can also show that, for all i, j

simultaneously,

$$\begin{aligned}\left|\frac{n_{1ij} - n_{1i} - n_{1j}}{n_{1i}n_{1j}}\right| &\leq \frac{n_{1i} + n_{1j}}{n_{1i}n_{1j}} = \frac{1}{n_{1i}} + \frac{1}{n_{1j}} \\ &\leq \frac{C_3}{n_1\zeta_i} + \frac{C_3}{n_1\zeta_j} = \frac{C_3}{n_1}\frac{\zeta_i + \zeta_j}{\zeta_i\zeta_j} \leq \frac{C_3}{n_1}\frac{1}{\zeta_i \wedge \zeta_j}\end{aligned}$$

holds with probability at least $1 - 1/(n \vee m)^2$.

Therefore, we can combine these with a union bound to get that, for all $i \neq j$, with probability at least $1 - 7/(n \vee m)^2$, the following holds.

$$\begin{aligned}|S^1_{A1}| &\leq \left|\frac{n_{1ij} - n_{1i} - n_{1j}}{n_{1i}n_{1j}}\right|\left|\sum_{k=1}^{n_1} u_k^{1i}u_k^{1j}x_k^{1i}x_k^{1j}\right| \\ &\leq \frac{C_3}{n_1}\frac{\zeta_i + \zeta_j}{\zeta_i\zeta_j}\left(\left|E\sum_{k=1}^{n_1} u_k^{1i}u_k^{1j}x_k^{1i}x_k^{1j}\right| + CK^2 \log n \|B^1\|_2\sqrt{a_{ii}a_{jj}}\right. \\ &\qquad\qquad \left. + CK^2 \log^{1/2}(m \vee n)\|B^1\|_F\sqrt{\zeta_i\zeta_j}\sqrt{a_{ii}a_{jj}}\right) \\ &\leq \frac{C_3}{n_1}(\zeta_i + \zeta_j)tr(B^1)|a_{ij}| + CK^2\frac{\log n}{n_1}\frac{1}{\zeta_i \wedge \zeta_j}\|B^1\|_2\sqrt{a_{ii}a_{jj}} \\ &\quad + CK^2\frac{\log^{1/2}(m \vee n)}{n_1}\|B^1\|_F\frac{\sqrt{\zeta_i\zeta_j}}{\zeta_i \wedge \zeta_j}\sqrt{a_{ii}a_{jj}} \\ &\leq \frac{C_3}{n_1}(\zeta_i + \zeta_j)tr(B^1)|a_{ij}| + CK^2\frac{\log n}{n_1}\frac{1}{\zeta_i \wedge \zeta_j}\|B^1\|_2\sqrt{a_{ii}a_{jj}} \\ &\quad + CK^2\frac{\log^{1/2}(m \vee n)}{n_1}\|B^1\|_F\frac{1}{\sqrt{\zeta_i \wedge \zeta_j}}\sqrt{a_{ii}a_{jj}}\end{aligned}$$

□

Lemma 5.5.4. *For $g = 1, 2$, we can bound, for all $i \neq j$ simultaneously,*

$$\begin{aligned}|S^g_{A2}| \leq C\frac{1}{n_g\zeta_{\min}}|a_{ij}|\left|\sum_{k\neq\ell} B^g_{k\ell}\right| + CK^2\frac{1}{\zeta_{\min}}\log n\|B^g\|_2\sqrt{a_{ii}a_{jj}} \\ + CK^2\frac{1}{\zeta_{\min}}\log^{1/2}(m \vee n)\|B^g\|_F\sqrt{a_{ii}a_{jj}}\end{aligned}$$

with probability at least $1 - 20/(n \vee m)^2$.

Proof. Without loss of generality assume $g = 1$. We separate S^1_{A2}, into three terms

$$\begin{aligned} S^1_{A2} = \sum_{k\neq\ell} D^{1ij}_{k\ell} x^{1i}_k x^{1j}_\ell &= \frac{n_{1ij}}{n_{1i}n_{1j}} \sum_{k\neq\ell} u^{1i}_k u^{1j}_\ell x^{1i}_k x^{1j}_\ell - \frac{1}{n_{1i}} \sum_{k\neq\ell} u^{1i}_\ell u^{1i}_k u^{1j}_\ell x^{1i}_k x^{1j}_\ell \\ &\quad - \frac{1}{n_{1j}} \sum_{k\neq\ell} u^{1j}_k u^{1i}_k u^{1j}_\ell x^{1i}_k x^{1j}_\ell \\ &= \frac{n_{1ij}}{n_{1i}n_{1j}} S^1_{A2,1} - \frac{1}{n_{1i}} S^1_{A2,2} - \frac{1}{n_{1j}} S^1_{A2,3} \end{aligned} \tag{5.45}$$

Concentration comes from the following proposition

Proposition 5.5.5. *With probability at least $1 - 19/(n \vee m)^2$, we get that the following hold simultaneously for all $i \neq j$.*

$$\begin{aligned} \left|S^1_{A2,1} - ES^1_{A2,1}\right| &\leq CK^2 \log n \|B^1\|_2 n_1 (\zeta_i \wedge \zeta_j) \sqrt{a_{ii}a_{jj}} \\ &\quad + CK^2 \log^{1/2}(m \vee n) \|B^1\|_F n_1 \sqrt{\zeta_i\zeta_j} \sqrt{a_{ii}a_{jj}} \\ \left|S^1_{A2,2} - ES^1_{A2,2}\right| &\leq CK^2 \log n \|B^1\|_2 n_1 (\zeta_i \wedge \zeta_j) \sqrt{a_{ii}a_{jj}} \\ &\quad + CK^2 \log^{1/2}(m \vee n) \|B^1\|_F n_1 \sqrt{\zeta_i\zeta_j} \sqrt{a_{ii}a_{jj}} \\ \left|S^1_{A2,3} - ES^1_{A2,3}\right| &\leq CK^2 \log n \|B^1\|_2 n_1 (\zeta_i \wedge \zeta_j) \sqrt{a_{ii}a_{jj}} \\ &\quad + CK^2 \log^{1/2}(m \vee n) \|B^1\|_F n_1 \sqrt{\zeta_i\zeta_j} \sqrt{a_{ii}a_{jj}} \end{aligned}$$

Again, a slight modification to Lemma 5.4.7 shows that, for all i, j simultaneously

$$\left|\frac{n_{1ij}}{n_{1i}n_{1j}}\right| \leq \left|\frac{1}{n_{1i}}\right|, \left|\frac{1}{n_{1j}}\right| \leq \frac{1}{n_{1i} \wedge n_{1j}} \leq C_3 \frac{1}{n_1(\zeta_i \wedge \zeta_j)} \tag{5.46}$$

holds with probability at least $1 - 1/(n \vee m)^2$.

So we get that, with probability at least $1-20/(n\vee m)^2$, the following holds for all $i \neq j$.

$$\begin{aligned}
|S^1_{A2}| &\le \left|\frac{n_{1ij}}{n_{1i}n_{1j}}S^1_{A2,1}\right| + \left|\frac{1}{n_{1i}}S^1_{A2,2}\right| + \left|\frac{1}{n_{1j}}S^1_{A2,3}\right| \\
&\le \frac{n_{1ij}}{n_{1i}n_{1j}}(|ES^1_{A2,1}| + |S^1_{A2,1} - ES^1_{A2,1}|) + \frac{1}{n_{1i}}(|ES^1_{A2,2}| + |S^1_{A2,2} - ES^1_{A2,2}|) \\
&\quad + \frac{1}{n_{1j}}(|ES^1_{A2,3}| + |S^1_{A2,3} - ES^1_{A2,3}|) \\
&\le C_3\frac{1}{n_1(\zeta_i \wedge \zeta_j)}(|ES^1_{A2,1}| + |ES^1_{A2,2}| + |ES^1_{A2,3}| \\
&\quad + |S^1_{A2,1} - ES^1_{A2,1}| + |S^1_{A2,2} - ES^1_{A2,2}| + |S^1_{A2,3} - ES^1_{A2,3}|) \\
&\le C_3\frac{\zeta_i\zeta_j}{n_1(\zeta_i \wedge \zeta_j)}\left(|a_{ij}|\left|\sum_{k\neq\ell}B^1_{k\ell}\right| + \zeta_i|a_{ij}|\left|\sum_{k\neq\ell}B^1_{k\ell}\right| + \zeta_j|a_{ij}|\left|\sum_{k\neq\ell}B^1_{k\ell}\right|\right) \\
&\quad + CK^2\frac{1}{n_1(\zeta_i \wedge \zeta_j)}\log n\|B^1\|_2 n_1(\zeta_i\wedge\zeta_j)\sqrt{a_{ii}a_{jj}} \\
&\quad + CK^2\frac{1}{n_1(\zeta_i \wedge \zeta_j)}\log^{1/2}(m\vee n)\|B^1\|_F n_1\sqrt{\zeta_i\zeta_j}\sqrt{a_{ii}a_{jj}} \\
&\le C\frac{\zeta_i\zeta_j}{n_1(\zeta_i \wedge \zeta_j)}|a_{ij}|\left|\sum_{k\neq\ell}B^1_{k\ell}\right| + CK^2\log n\|B^1\|_2\sqrt{a_{ii}a_{jj}} \\
&\quad + CK^2\frac{1}{\sqrt{\zeta_i\wedge\zeta_j}}\log^{1/2}(m\vee n)\|B^1\|_F\sqrt{a_{ii}a_{jj}}
\end{aligned}$$

□

Putting these all together, we get that

$$\begin{aligned}
&|\hat{S}(A)_{ij} - x^{iT}\mathrm{diag}(u^i \circ u^j)x^j| \\
&\quad \le C(\zeta_i+\zeta_j)|a_{ij}|\left(\frac{\mathrm{tr}(B^1)}{n_1} + \frac{\mathrm{tr}(B^2)}{n_2}\right) + \frac{\zeta_i\zeta_j|a_{ij}|}{(\zeta_i\wedge\zeta_j)}\left(\frac{1}{n_1}\left|\sum_{k\neq\ell}B^1_{k\ell}\right| + \frac{1}{n_2}\left|\sum_{k\neq\ell}B^2_{k\ell}\right|\right) \\
&\qquad + CK^2\frac{\log^{1/2}(m\vee n)}{\sqrt{\zeta_i\wedge\zeta_j}}\left(\frac{\|B^1\|_F}{n_1} + \frac{\|B^2\|_F}{n_2}\right)\sqrt{a_{ii}a_{jj}} \\
&\qquad + CK^2\frac{\log^{1/2}(m\vee n)}{\sqrt{\zeta_i\wedge\zeta_j}}(\|B^1\|_F + \|B^2\|_F)\sqrt{a_{ii}a_{jj}} \\
&\qquad + CK^2\frac{\log n}{\zeta_i\wedge\zeta_j}\left(\frac{\|B^1\|_2}{n_1} + \frac{\|B^2\|_2}{n_2}\right)\sqrt{a_{ii}a_{jj}} + CK^2\log n(\|B^1\|_2 + \|B^2\|_2)\sqrt{a_{ii}a_{jj}}
\end{aligned}$$

The proof is completed by noting that

$$
\begin{aligned}
&|\hat{S}(A)_{ij} - \rho_{ij}(A)\sqrt{a_{ii}a_{jj}}\mathcal{N}_{ij}| \\
&\qquad \leq |\hat{S}(A)_{ij} - x^{iT}\mathrm{diag}(u^i \circ u^j)x^j| + |x^{iT}\mathrm{diag}(u^i \circ u^j)x^j - \rho_{ij}(A)\sqrt{a_{ii}a_{jj}}\mathcal{N}_{ij}|
\end{aligned}
$$

and using Lemma 5.3.8 with the last term and then rearranging.

For $i = j$, we get that

$$
\begin{aligned}
\hat{S}(A)_{ii} - x^{iT}\mathrm{diag}(u^i)x^i = &-\sum_{k=1}^{n_1} \frac{1}{n_{1i}} u_k^{1i}(x_k^{1i})^2 - \sum_{k \neq \ell} \frac{1}{n_{1i}} u_k^{1i} u_\ell^{1i} x_k^{1i} x_\ell^{1i} \\
&-\sum_{k=1}^{n_2} \frac{1}{n_{2i}} u_k^{2i}(x_k^{2i})^2 - \sum_{k \neq \ell} \frac{1}{n_{2i}} u_k^{2i} u_\ell^{2i} x_k^{2i} x_\ell^{2i}
\end{aligned}
$$

The result is then shown by following the same steps as the $i \neq j$ case, though often simplified due to the reduction of terms.

□

Proof of Proposition 5.5.5. We start by rewriting $S^1_{A2,1}$.

$$
\begin{aligned}
S^1_{A2,1} &= (x^{1i})^T(u^{1i}u^{1jT} - \mathrm{diag}(u^{1i}u^{1jT}))x^{1j} \\
&= (x^{1i})^T D x^{1j} \qquad (5.47)
\end{aligned}
$$

Without loss of generality, let $i = 1, j = 2$. From here, we condition on $\mathbb{U}$ (and therefore D), holding the sparsity pattern of our data constant.

We concatenate the vectors x^{11}, x^{12} to form $(x^{11}, x^{12}) \in \mathbb{R}^{2n_1}$ with covariance matrix $A_{0,(1,2)} \otimes B_0$. We $g_1, \ldots, g_{2n_1} \overset{\text{i.i.d}}{\sim} Z$ for Z subgaussian with subgaussian constant K, mean zero, and unit variance, and let

$$
\begin{pmatrix} c_{ii} & c_{ij} \\ c_{ij} & c_{jj} \end{pmatrix}
$$

be the symmetrix square root of $A_{0,(1,2)}$ and define

$$D'(i,j) = \begin{pmatrix} c_{ii}c_{ij} & c_{ii}c_{jj} \\ c_{ij}c_{ij} & c_{ij}c_{jj} \end{pmatrix} \otimes B_0^{1/2} D B_0^{1/2}.$$

Then

$$S^1_{A2,1} = (x^{1i})^T D x^{1j} = \sum_{k=1}^{2n_1} \sum_{\ell=1}^{2n_1} D'_{k\ell}(1,2) g_k g_\ell$$

where

$$E \sum_{k=1}^{2n_1} \sum_{\ell=1}^{2n_1} D'_{k\ell}(1,2) g_k g_\ell = \operatorname{tr}(D'(1,2)) = a_{12} \operatorname{tr}(B_0 D).$$

For the concentration, we will need the following proposition, which is a reflection of Proposition 5.3.9 for this direction.

Proposition 5.5.6. *Let $A_{0,(i,j)} = (a_{ij})_{i,j=1}^2 \in \mathbb{R}^{2\times 2}$ be the positive definite submatrix of A_0 with rows and columns i, j. Denote it's unique symmetric square root as*

$$\begin{pmatrix} c_{ii} & c_{ij} \\ c_{ij} & c_{jj} \end{pmatrix}$$

Define

$$D'(i,j) = B_0^{1/2} D B_0^{1/2} \otimes \begin{pmatrix} c_{ii}c_{ij} & c_{ii}c_{jj} \\ c_{ij}c_{ij} & c_{ij}c_{jj} \end{pmatrix}$$

Then

$$\|D'(i,j)\|_2 \le \sqrt{a_{ii}a_{jj}} \|B_0^{1/2} D B_0^{1/2}\|_2 \tag{5.48}$$

$$\|D'(i,j)\|_F \le \sqrt{a_{ii}a_{jj}} \|B_0^{1/2} D B_0^{1/2}\|_F \tag{5.49}$$

$$\tag{5.50}$$

And, recalling that $\rho_{ij}(A_0) = a_{ij}/\sqrt{a_{ii}a_{jj}}$,

$$\left|\rho_{ij}(A_0)\frac{c_{ii}c_{jj} + c_{ij}^2}{a_{ij}}\right| < 1 \tag{5.51}$$

Proof. This follows the same proof as Proposition 5.3.9. □

If we partition $g = (g_1, g_2)$, consider the following quadratic forms

$$\begin{aligned} Z &= g_1(B_0^{1/2}DB_0^{1/2})g_1^T - Eg_1(B_0^{1/2}DB_0^{1/2})g_1^T \\ Z' &= g_2(B_0^{1/2}DB_0^{1/2})g_2^T - Eg_2(B_0^{1/2}DB_0^{1/2})g_2^T \\ U &= g_1(B_0^{1/2}DB_0^{1/2})g_2^T - Eg_1(B_0^{1/2}DB_0^{1/2})g_2^T \end{aligned}$$

For Z, Z' independent.

Then

$$\begin{aligned} &\frac{1}{\sqrt{a_{ii}a_{jj}}}gD'(i,j)g^T - E\frac{1}{\sqrt{a_{ii}a_{jj}}}gD'(i,j)g^T \\ &\quad \leq \left|\frac{a_{ij}}{\sqrt{a_{ii}a_{jj}}}\left(\frac{1}{a_{ij}}(gD'(i,j)g^T - EgD'(i,j)g^T)\right)\right| \\ &\quad \leq |\rho_{ij}(A_0)|\left(\left|\frac{c_{ii}c_{ij}}{a_{ij}}Z + \frac{c_{ij}c_{jj}}{a_{ij}}Z'\right| + \left|\frac{c_{ii}c_{jj} + c_{ij}^2}{a_{ij}}U\right|\right) \\ &\quad = |\rho_{ij}(A_0)|\left(|s_1Z + s_2Z'| + \left|\frac{c_{ii}c_{jj} + c_{ij}^2}{a_{ij}}U\right|\right) \\ &\quad \leq |\rho_{ij}(A_0)||s_1Z + s_2Z'| + |U| \end{aligned}$$

where $s_1 = (c_{ii}c_{ij})/a_{ij}$, $s_2 = (c_{ij}c_{jj})/a_{ij}$, and $s_1, s_2 \in [0,1]$ with $s_1 + s_2 = 1$ since $A_{0,(i,j)}^{1/2}$ is positive definite and $c_{ii}c_{ij} + c_{ij}c_{jj} = \mathrm{tr}(c_ic_j^T) = a_{ij}$.

Therefore, we can use this to get that

$$
\begin{aligned}
&P\left(\left|\frac{S^1_{A2,1}}{\sqrt{a_{ii}a_{jj}}} - E\frac{S^1_{A2,1}}{\sqrt{a_{ii}a_{jj}}}\right| > t\right) \\
&\quad = P\left(\left|\frac{gD'(i,j)g^T}{\sqrt{a_{ii}a_{jj}}} - E\frac{gD'(i,j)g^T}{\sqrt{a_{ii}a_{jj}}}\right| > t\right) \\
&\quad \leq P(|\rho_{ij}(A_0)||s_1 Z + s_2 Z'| > t/2) + P(|U| > t/2) \\
&\quad \overset{(a)}{\leq} P(|\rho_{ij}(A_0)||Z| > t/2) + P(|\rho_{ij}(A_0)||Z'| > t/2) + P(|U| > t/2) \\
&\quad =: p_1 + p_2 + p_3
\end{aligned}
$$

where (a) is because $s_1, s_2 \in [0,1]$ and $s_1 + s_2 = 1$ means that $|s_1 Z + s_2 Z'| > k \implies |Z| > k$ and/or $|Z'| > k$.

We note that $p_1 = P(|\rho_{ij}(B_0)||Z| > t/2) \leq P(|Z| > t/2)$. Then we can apply Theorem 1.1 from *Rudelson and Vershynin* (2013) to p_1, p_2, and p_3[1] with Proposition 5.5.6 and Lemma 5.5.7 to get that

$$
p_1, p_2, p_3 \leq 2\exp\left(-c\min\left(\frac{t^2}{K^4\|B\|_F^2 n_{1i}n_{1j}}, \frac{t}{K^2\|B\|_2\min(n_{1i}n_{1j})}\right)\right).
$$

So after adding these probabilities, we can show that, conditional on $\mathbb{U}$, with probability at least $1 - 6/(n \vee m)^c$,

$$
\begin{aligned}
&\left|S^1_{A2,1} - ES^1_{A2,1}\right| \\
&\quad \leq \sqrt{a_{ii}a_{jj}}\left(CK^2\log n\|B^1\|_2\min(n_{1i}, n_{1j}) + CK^2\log^{1/2}(m\vee n)\|B^1\|_F\sqrt{n_{1i}n_{1j}}\right) \\
&\quad \leq CK^2\log n\|B^1\|_2\min(n_{1i}, n_{1j})\sqrt{a_{ii}a_{jj}} + CK^2\log^{1/2}(m\vee n)\|B^1\|_F\sqrt{n_{1i}n_{1j}}\sqrt{a_{ii}a_{jj}}
\end{aligned}
$$

[1] Note that for U, since g_1 and g_2 are independnet, we actually only need the easier version of Theorem 1.1 that does not require the decoupling argument.

Rewriting $S^1_{A2,2}$, we can similarly show that

$$\begin{aligned} S^1_{A2,2} &= \sum_{k \neq \ell} u^{1i}_\ell u^{1i}_k u^{1j}_\ell x^{1i}_k x^{1j}_\ell \\ &= (x^{1i})^T \left(u^{1i} (u^{1i} \circ u^{1j})^T - \operatorname{diag}(u^{1i} (u^{1i} \circ u^{1j})^T) \right) x^{1j} \end{aligned}$$

Following the same steps as the above proof for $S^1_{A2,2}$, we get that

$$\begin{aligned} &\left| S^1_{A2,2} - ES^1_{A2,2} \right| \\ &\quad \leq \sqrt{a_{ii} a_{jj}} \left(CK^2 \log n \|B^1\|_2 \min(n_{1i}, n_{1j}) + CK^2 \log^{1/2}(m \vee n) \|B^1\|_F \sqrt{n_{1i} n_{1j}} \right) \\ &\quad \leq CK^2 \log n \|B^1\|_2 \min(n_{1i}, n_{1j}) \sqrt{a_{ii} a_{jj}} + CK^2 \log^{1/2}(m \vee n) \|B^1\|_F \sqrt{n_{1i} n_{1j}} \sqrt{a_{ii} a_{jj}} \end{aligned}$$

The argument for $S^1_{A2,3}$ is symmetric to that of $S^1_{A2,2}$. Then by using a union bound over these events, we get that, with probability at least $1 - 18/(n \vee m)^2$, conditional on the mask $\mathbb{U}$ we get that the following hold simultaneously for all $i \neq j$.

$$\begin{aligned} \left| S^1_{A2,1} - ES^1_{A2,1} \right| &\leq CK^2 \log n \|B^1\|_2 (n_{1i} \wedge n_{1j}) \sqrt{a_{ii} a_{jj}} + CK^2 \log^{1/2}(m \vee n) \|B^1\|_F \sqrt{n_{1i} n_{1j}} \sqrt{a_{ii} a_{jj}} \\ \left| S^1_{A2,2} - ES^1_{A2,2} \right| &\leq CK^2 \log n \|B^1\|_2 (n_{1i} \wedge n_{1j}) \sqrt{a_{ii} a_{jj}} + CK^2 \log^{1/2}(m \vee n) \|B^1\|_F \sqrt{n_{1i} n_{1j}} \sqrt{a_{ii} a_{jj}} \\ \left| S^1_{A2,3} - ES^1_{A2,3} \right| &\leq CK^2 \log n \|B^1\|_2 (n_{1i} \wedge n_{1j}) \sqrt{a_{ii} a_{jj}} + CK^2 \log^{1/2}(m \vee n) \|B^1\|_F \sqrt{n_{1i} n_{1j}} \sqrt{a_{ii} a_{jj}} \end{aligned}$$

From here, using Hoeffding's bound, we can get that,

$$\begin{aligned} & P\left(\frac{n_{1i}}{n_1} - \zeta_i \leq \sqrt{3/2} \sqrt{\log(n \vee m)/n_1} \right) \geq 1 - 1/(n \vee m)^3 \\ \implies & P\left(\frac{n_{1i}}{n_1} - \zeta_i \leq \sqrt{3/2} \sqrt{\log(n \vee m)/n_1} \; \forall \; i \right) \geq 1 - 1m/(n \vee m)^3 \end{aligned}$$

Rearranging, we get that this event also implies

$$\frac{n_{1i}}{n_1 \zeta_i} \leq \sqrt{3/2} \frac{1}{\zeta_i} \sqrt{\frac{\log(n \vee m)}{n_1}} \leq C_4$$

Where the last inequality uses the assumption that $\zeta_{\min} \gtrsim \sqrt{\log(m \vee n)/n_1}$.

Combining this with the above event with a union bound provides our final result. □

5.5.3 Additional results

Lemma 5.5.7. *We can show that*

$$\|(B^1)^{1/2}(u^{1i}u^{1jT} - \text{diag}(u^{1i}u^{1jT}))(B^1)^{1/2}\|_F \le \|B^1\|_2\sqrt{n_{1i}n_{1j}}$$
$$\|(B^1)^{1/2}(u^{1i}u^{1jT} - \text{diag}(u^{1i}u^{1jT}))(B^1)^{1/2}\|_2 \le \|B^1\|_2 \min(n_{1i}, n_{1j})$$

and, similarly,

$$\|(B^1)^{1/2}\left(u^{1i}(u^{1i} \circ u^{1j})^T - \text{diag}(u^{1i}(u^{1i} \circ u^{1j})^T)\right)(B^1)^{1/2}\|_F \le \|B^1\|_2\sqrt{n_{1i}n_{1j}}$$
$$\|(B^1)^{1/2}\left(u^{1i}(u^{1i} \circ u^{1j})^T - \text{diag}(u^{1i}(u^{1i} \circ u^{1j})^T)\right)(B^1)^{1/2}\|_2 \le \|B^1\|_2 \min(n_{1i}, n_{1j}).$$

Proof of Lemma 5.5.7. We begin by showing

$$\|u^{1i}u^{1jT} - \text{diag}(u^{1i}u^{1jT})\|_F = \sqrt{|u^{1i}||u^{1j}| - |u^{1i} \circ u^{1j}|} \le \sqrt{n_{1i}n_{1j}}$$
$$\|u^{1i}u^{1jT} - \text{diag}(u^{1i}u^{1jT})\|_2 \le \min(|u^{1i}|, |u^{1j}|) \le \min(n_{1i}, n_{1j}) \tag{5.52}$$

and

$$\|u^{1i}(u^{1i} \circ u^{1j})^T - \text{diag}(u^{1i}(u^{1i} \circ u^{1j})^T)\|_F = \sqrt{|u^{1i}||u^{1i} \circ u^{1j}| - |u^{1i} \circ u^{1j}|} \le \sqrt{n_{1i}n_{1j}}$$
$$\|u^{1i}(u^{1i} \circ u^{1j})^T - \text{diag}(u^{1i}(u^{1i} \circ u^{1j})^T)\|_2 \le |u^i \circ u^j| \le \min(n_{1i}, n_{1j}). \tag{5.53}$$

The Frobenius norm bounds come from a simple counting exercise; every element of $u^{1i}u^{1jT} - \text{diag}(u^{1i}u^{1jT})$ and $u^{1i}(u^{1i} \circ u^{1j})^T - \text{diag}(u^{1i}(u^{1i} \circ u^{1j})^T)$ is either a 1 or a 0, so the squared frobenius norm of each matrix is the number of 1's in that matrix.

For the first spectral norm bound (5.52), we note that since all of the elements are 1 or

0, adding more 1's to the matrix will never decrease the spectral norm. So

$$\begin{aligned}
\|u^{1i}u^{1jT} - \text{diag}(u^{1i}u^{1jT})\|_2 &\leq \|u^{1i}\vec{1}_n^T\|_2 = |u^{1i}| \\
\|u^{1i}u^{1jT} - \text{diag}(u^{1i}u^{1jT})\|_2 &\leq \|\vec{1}_n u^{1jT}\|_2 = |u^{1j}|.
\end{aligned}$$

For the second spectral norm bound (5.53), we can similarly show that

$$\begin{aligned}
\|u^{1i}(u^{1i} \circ u^{1j})^T - \text{diag}(u^{1i}(u^{1i} \circ u^{1j})^T)\|_2 &\leq \|\vec{1}_n(u^{1i} \circ u^{1j})^T\|_2 = |u^{1i} \circ u^{1j}| \\
&\leq \min(|u^{1i}|, |u^{1j}|)
\end{aligned}$$

To get the final bounds, we bound each of the norms as

$$\begin{aligned}
&\|(B^1)^{1/2}(u^{1i}u^{1jT} - \text{diag}(u^{1i}u^{1jT}))(B^1)^{1/2}\|_F \\
&\qquad \leq \|(B^1)^{1/2}\|_2\|u^{1i}u^{1jT} - \text{diag}(u^{1i}u^{1jT})\|_F\|(B^1)^{1/2}\|_2 \leq \|B^1\|_2\sqrt{n_{1i}n_{1j}} \\
&\|(B^1)^{1/2}(u^{1i}u^{1jT} - \text{diag}(u^{1i}u^{1jT}))(B^1)^{1/2}\|_2 \\
&\qquad \leq \|(B^1)^{1/2}\|_2\|u^{1i}u^{1jT} - \text{diag}(u^{1i}u^{1jT})\|_2\|(B^1)^{1/2}\|_2 \leq \|B^1\|_2 \min(n_{1i}, n_{1j})
\end{aligned}$$

and

$$\begin{aligned}
&\|(B^1)^{1/2}\left(u^{1i}(u^{1i} \circ u^{1j})^T - \text{diag}(u^{1i}(u^{1i} \circ u^{1j})^T)\right)(B^1)^{1/2}\|_F \\
&\qquad \leq \|(B^1)^{1/2}\|_2\|u^{1i}(u^{1i} \circ u^{1j})^T - \text{diag}(u^{1i}(u^{1i} \circ u^{1j})^T)\|_F\|(B^1)^{1/2}\|_2 \\
&\qquad \leq \|B\|_2\sqrt{n_{1i}n_{1j}} \\
&\|(B^1)^{1/2}\left(u^{1i}(u^{1i} \circ u^{1j})^T - \text{diag}(u^{1i}(u^{1i} \circ u^{1j})^T)\right)(B^1)^{1/2}\|_2 \\
&\qquad \leq \|(B^1)^{1/2}\|_2\|u^{1i}(u^{1i} \circ u^{1j})^T - \text{diag}(u^{1i}(u^{1i} \circ u^{1j})^T)\|_2\|(B^1)^{1/2}\|_2 \\
&\qquad \leq \|B\|_2 \min(n_{1i}, n_{1j}).
\end{aligned}$$

□

CHAPTER VI

Future Work

6.1 Matrix-variate binary data

We present some initial methodology for using our matrix-variate graph estimators for a Kronecker-product Ising model in Chapter IV. Although we present some simulation evidence that we can indeed recover the structure of binary graphical models, there are no theoretical results for this setting that we know of. We are interested in developing more theoretical justifications for using this method and the approximations involved, akin to extending the work of *Banerjee et al.* (2008) to the matrix-variate case, as well as adapting other estimators designed for use with binary and discrete data. In particular, we are interested in models that utilize underlying Gaussian latent variables that are then discretized (e.g. *Suggala et al.*, 2017; *Fan et al.*, 2017; *Feng and Ning*, 2019) and in how we might replace the i.i.d. latent factors with matrix-variate versions.

6.2 Theoretical results for flexible mean estimation with dependence

The two-group mean structure with known group labels studied in Chapters III and V is quite restrictive even though it is commonly used. Although we provide some guidance on methods to use when more flexible mean estimation is desired, there is a considerable gap here in terms of theoretical results. Developing theory for the low-rank estimator may be out of reach at the current moment, but it may be possible to provide guarantees for two- or K-group means with unknown labels by first applying a spectral clustering step to estimate group labels. The setting in *Blum et al.* (2007) is a different setting than ours, but provides a framework for the theoretical tools and results that would be needed.

6.3 Extensions to more general missing structures

Recently, *Pavez and Ortega* (2019) and *Park et al.* (2020) have taken steps towards showing results for covariance estimation with more general missing structures, though both use settings with independent observations. We make two fairly strong assumptions on the missingness structure, that whether each entry is observed is independent of whether other entries are observed, and that the missingness is independent of the data (missing completely at random, or MCAR). These assumptions are standard, but in application are rarely expected to be fully satisfied. Extensions of our theory in the matrix-variate case to allow for relaxations of these assumptions would allow our methods to be used with significantly more confidence in a wide range of applications.

APPENDICES

APPENDIX A

Appendices to Precision Matrix Estimation with Noisy and Missing Data

A.1 Proofs of Propositions

Proof of Proposition 1. The optimization problem (2.6) is equivalently

$$\min_{\Theta,V}\ \phi(\Theta,V)=\min_{\Theta,V}\{f_1(\Theta)+f_2(V)\}\ \text{ s.t. }\ A\text{vec}(V)+B\text{vec}(\Theta)=0 \tag{A.1}$$

where $f_1(\Theta)=\text{tr}(\hat{\Gamma}_n\Theta)-\log\det(\Theta)+\mathbb{1}_{\mathcal{X}_R}(V)$, $f_2(V)=g_\lambda(V)$, $A=-I_{m^2}$, and $B=I_{m^2}$.

Boyd et al. (2010) show that if f_1 and f_2 are proper convex functions and if (A.1) is solveable then ADMM converges in terms of the objective value $\phi(\Theta^t,V^t)\to\phi^*$ and dual variable $\Lambda^t\to\Lambda^*$. *Bertsekas and Tsitsiklis* (1989, Proposition 4.2) and *Mota et al.* (2011) show that if in addition A and B have full column rank then we get convergence of the primal iterates $\Theta^t\to\Theta^*$ and $V^t\to V^*$, where (Θ^*,V^*) is the solution to (A.1). □

Before we prove Proposition 2, we first define directional derivatives and stationary points.

Definition. The *directional derivative* of a lower semi-continuous function h at Θ in the direction Δ is

$$h'(\Theta; \Delta) = \lim_{t \searrow 0} \frac{h(\Theta + t\Delta) - h(\Theta)}{t}.$$

Note that we allow $h'(\Theta; \Delta) = +\infty$. We say that Θ is a *stationary point* of h if it satisfies the first-order necessary conditions to be a local extrema, i.e.

$$h'(\Theta; \Delta) \geq 0 \text{ for all directions } \Delta \in \mathbb{R}^{m \times m}$$

Note that this coincides with the definition of stationary point used in *Loh and Wainwright* (2017), though they use slightly different notation. Also note that $h'(\Theta; \Delta) = \langle \nabla h(\Theta), \Delta \rangle$ when h is continuously differentiable.

Proof of Proposition 2. From the first-order necessary conditions of the subproblems (2.8)-(2.9), we get that, for all $\Delta \in \mathbb{R}^{m \times m}$,

$$\begin{aligned} &0 \leq g'_\lambda(V^{k+1}; \Delta) - \langle \rho(\Theta^k - V^{k+1}) + \Lambda^k, \Delta \rangle \\ &0 \leq \langle \hat{\Gamma}_n - (\Theta^{k+1})^{-1} + \rho(\Theta^{k+1} - V^{k+1}) + \Lambda^k, \Delta \rangle + \mathbb{1}'_{\mathcal{X}_R}(\Theta^{k+1}; \Delta) \end{aligned} \tag{A.2}$$

And recall that

$$\Lambda^{k+1} = \Lambda^k + \rho(\Theta^{k+1} - V^{k+1}). \tag{A.3}$$

We can rewrite (A.2)-(A.3) as

$$g'_\lambda(V^{k+1}; \Delta) \geq \langle \rho(\Theta^k - \Theta^{k+1}) + \Lambda^{k+1}, \Delta \rangle \tag{A.4}$$

$$0 \leq \langle \hat{\Gamma}_n - (\Theta^{k+1})^{-1} + \Lambda^{k+1}, \Delta \rangle + \mathbb{1}'_{\mathcal{X}_R}(\Theta^{k+1}; \Delta) \tag{A.5}$$

$$\frac{1}{\rho}(\Lambda^{k+1} - \Lambda^k) = \Theta^{k+1} - V^{k+1}. \tag{A.6}$$

Now consider a fixed point $(\Theta^*, V^*, \Lambda^*)$ and consider (A.4)-(A.6) evaluated at this limit point. From (A.6) we get that $\Theta^* = V^*$. This combined with (A.4) gives us that, for all

$\Delta \in \mathbb{R}^{m \times m}$,

$$g'_\lambda(\Theta^*; \Delta) \geq \langle \Lambda^*, \Delta \rangle$$

Finally, (A.5) gives us that

$$0 \leq \langle \hat{\Gamma}_n - (\Theta^*)^{-1} + \Lambda^*, \Delta \rangle + \mathbb{1}'_{\mathcal{X}_R}(\Theta^*; \Delta)$$

Using the above and recalling the objective f as defined in (2.6), we get that, for all $\Delta \in \mathbb{R}^{m \times m}$,

$$\begin{aligned} 0 &\leq \langle \hat{\Gamma}_n - (\Theta^*)^{-1}, \Delta \rangle + \langle \Lambda^*, \Delta \rangle + \mathbb{1}'_{\mathcal{X}_R}(\Theta^*; \Delta) \\ &\leq \langle \hat{\Gamma}_n - (\Theta^*)^{-1}, \Delta \rangle + g'_\lambda(\Theta^*; \Delta) + \mathbb{1}'_{\mathcal{X}_R}(\Theta^*; \Delta) = f'(\Theta^*; \Delta) \end{aligned}$$

So Θ^* is a stationary point of f by definition. □

Comparison to Guo and Zhang (2017)

Guo and Zhang (2017) study the problem of condition number-constrained precision matrix estimation, where they consider the estimator

$$\hat{\Theta} = \underset{\Theta \succ 0, \text{cond}(\Theta) \leq \kappa}{\arg\min} -\log\det\Theta + \text{tr}(\hat{\Gamma}_n\Theta) + \lambda\|\Theta\|_{1,\text{off}} \tag{A.7}$$

Note that this is quite similar to the estimators we consider in (2.2), as they simply replace the maximum eigenvalue constraint with a constraint on the ratio of the maximum to minimum eigenvalues.

However, they do not study the application of their estimator to cases with indefinite input or its performance in noisy and missing data situations. In particular, constraining the condition number does not necessarily guarantee that the graphical Lasso objective (2.1) will be lower bounded, especially when using nonconvex penalties.

As a simple example, consider the case with an input matrix and iterates

$$\hat{\Gamma}_n = \begin{pmatrix} 1 & 0 \\ 0 & -0.2 \end{pmatrix} \qquad \Theta^t = t \begin{pmatrix} 0.1 & 0 \\ 0 & 1 \end{pmatrix}$$

In this case the objective is

$$f(\Theta^t) = \operatorname{tr}(\hat{\Gamma}_n \Theta^t) - \log\det\Theta^t = -0.1 \times t - \log(0.1 \times t)$$

which is unbounded below as t grows even though the condition numbers of the iterates are constant.

More generally, whenever $\hat{\Gamma}_n \in \mathbb{R}^{m \times m}$ has eigenvalues $\sigma_1, \ldots, \sigma_m$, where $\sigma_1 \geq \cdots \geq \sigma_{m_1} \geq 0$ and $0 > \sigma_{m_1+1} \geq \cdots \geq \sigma_m$. Denote $S_1 = \sum_{i=1}^{m_1} \sigma_i$ and $S_2 = \sum_{i=m_1+1}^{m} -\sigma_i$. Let $VDV^T = \hat{\Gamma}_n$ be the eigendecomposition of the covariance estimate. Then for some condition number bound κ, we can consider iterates of the form $\Theta^t = tVMV^T$, where M is a diagonal matrix with entries

$$M_{ii} = \begin{cases} 1 & \text{if } i \leq m_1 \\ \kappa & \text{if } i > m_1 \end{cases}$$

Which we note has a condition number of κ. Then we can see that the objective becomes

$$\begin{aligned} f(\Theta^t) &= t\operatorname{tr}(VDV^TVMV^T) - (m - m_1)\log(\kappa) + g_\lambda(tVMV^T) \\ &= t(S_1 - \kappa S_2) - (m - m_1)\log(\kappa) + g_\lambda(tVMV^T) \end{aligned}$$

So if $\kappa > S_1/S_2$ then this objective is still unbounded below.

Using a spectral norm bound $\|\Theta\|_2 \leq R$ as the side constraint with a indefinite input guarantees a lower bound on the graphical Lasso objective regardless of the choice of R and is therefore a more natural side constraint to use.

A.2 ADMM for general side constraints

In this section we develop an ADMM algorithm for general side constraints, i.e. the following variant of (2.2).[1]

$$\hat{\Theta} \in \underset{\Theta \succeq 0, h(\Theta) \leq R}{\arg\min} \; \mathrm{tr}(\hat{\Gamma}_n \Theta) - \log\det(\Theta) + g_\lambda(\Theta).$$

This algorithm has the same convergence guarantees as Algorithm 1, but in practice we find that Algorithm 1 converges faster and more consistently when the spectral norm side constraint is used.

Derivation

We first rewrite the objective as

$$f(\Theta) = \mathrm{tr}(\hat{\Gamma}_n \Theta) - \log\det(\Theta) + g_\lambda(\Theta) + \mathbb{1}_{\mathcal{X}_{h,R}}(\Theta) \tag{A.8}$$

where $\mathcal{X}_{h,R} = \{\Theta : \Theta \succeq 0, h(\Theta) \leq R\}$ and

$$\mathbb{1}_{\mathcal{X}}(\Theta) = \begin{cases} 0 & \text{if } \Theta \in \mathcal{X} \\ \infty & \text{otherwise.} \end{cases}$$

We can then introduce auxiliary optimization variables $V_1, V_2 \in \mathbb{R}^{m \times m}$ and reformulate the optimization problem as

$$\hat{\Theta} = \underset{\Theta, V_1, V_2}{\arg\max} \left\{ \mathrm{tr}(\hat{\Gamma}_n \Theta) - \log\det(\Theta) + g_\lambda(V_1) + \mathbb{1}_{\mathcal{X}_{h,R}}(V_2) \right\} \text{ s.t. } \Theta = V_1 = V_2$$

For a penalty parameter $\rho > 0$ and Lagrange multiplier matrices $\Lambda_1, \Lambda_2 \in \mathbb{R}^{m \times m}$, we

[1] Note that we switch the notation of the side constraint function from ρ to h to avoid confusion with the ADMM penalty parameter ρ.

consider the augmented Lagrangian of this problem

$$\begin{aligned}\mathcal{L}_\rho(\Theta, V_1, V_2, \Lambda_1, \Lambda_2) = \operatorname{tr}(\hat{\Gamma}_n\Theta) - \log\det(\Theta) + g_\lambda(V_1) + \mathbb{1}_{\mathcal{X}_{h,R}}(V_2) \\ + \frac{\rho}{2}\|\Theta - V_1\|_F^2 + \frac{\rho}{2}\|\Theta - V_2\|_F^2 + \langle\Lambda_1, \Theta - V_1\rangle \quad + \langle\Lambda_2, \Theta - V_2\rangle\end{aligned} \tag{A.9}$$

The ADMM algorithm is then, given current iterates Θ^k, V_1^k, V_2^k, Λ_1^k, and Λ_2^k,

$$V_1^{k+1} = \operatorname*{arg\,min}_{V_1 \in \mathbb{R}^{m\times m}} \left\{ g_\lambda(V_1) + \frac{\rho}{2}\|\Theta^k - V_1\|_F^2 + \langle\Lambda_1^k, \Theta^k - V_1\rangle \right\} \tag{A.10}$$

$$V_2^{k+1} = \operatorname*{arg\,min}_{V_2 \in \mathbb{R}^{m\times m}} \left\{ \mathbb{1}_{\mathcal{X}_{h,R}}(V_2) + \frac{\rho}{2}\|\Theta^k - V_2\|_F^2 + \langle\Lambda_2^k, \Theta^k - V_2\rangle \right\} \tag{A.11}$$

$$\begin{aligned}\Theta^{k+1} = \operatorname*{arg\,min}_{\Theta \in \mathbb{R}^{m\times m}} \Big\{ -\log\det\Theta + \operatorname{tr}(\hat{\Gamma}_n\Theta) + \frac{\rho}{2}\|\Theta - V_1^{k+1}\|_F^2 + \frac{\rho}{2}\|\Theta - V_2^{k+1}\|_F^2 \\ + \langle\Lambda_1^k, \Theta - V_1^{k+1}\rangle + \langle\Lambda_2^k, \Theta - V_2^{k+1}\rangle \Big\}\end{aligned} \tag{A.12}$$

$$\Lambda_1^{k+1} = \Lambda_1^k + \rho(\Theta^{k+1} - V_1^{k+1}) \tag{A.13}$$

$$\Lambda_2^{k+1} = \Lambda_2^k + \rho(\Theta^{k+1} - V_2^{k+1}) \tag{A.14}$$

Considering the V_1-subproblem, we can show that the minimization problem in (A.10) is equivalent to

$$V_1^{k+1} = \operatorname*{arg\,min}_{V_1 \in \mathbb{R}^{m\times m}} \left\{ \frac{1}{\rho} g_\lambda(V_1) + \frac{1}{2}\left\| V_1 - \frac{\rho\Theta^k + \Lambda_1^k}{\rho} \right\|_F^2 \right\}.$$

Which is a prox operator of g_λ/ρ. These have the same form as described in Section 2.2.1.

For the V_2-subproblem, we similarly see that (A.11) is equivalent to

$$V_2^{k+1} = \operatorname*{arg\,min}_{V_2 \in \mathbb{R}^{m\times m}} \left\{ \mathbb{1}_{\mathcal{X}_{h,R}}(V_2) + \frac{1}{2}\left\| V_2 - \frac{\rho\Theta^k + \Lambda_2^k}{\rho} \right\|_F^2 \right\}.$$

which is equivalent to the projection operator

$$\operatorname{Proj}_{\mathcal{X}_{h,R}}\left(\frac{\rho\Theta^k + \Lambda_2^k}{\rho}\right) = \min_{V_2 \in \mathcal{X}_{h,R}} \left\| V_2 - \frac{\rho\Theta^k + \Lambda_2^k}{\rho} \right\|_F^2 \tag{A.15}$$

Note that if directly projecting onto $\mathcal{X}_{h,R}$ does not have an closed-form solution, we can perform this step using Dykstra's alternating projection algorithm.

Finally, for the Θ-subproblem, we can again show that (A.12) is equivalent to

$$\Theta = \underset{\Theta \in \mathbb{R}^{m \times m}}{\arg\min} \Bigg\{ -\log\det\Theta + \rho \left\| \Theta - \frac{\rho V_1^{k+1} + \rho V_2^{k+1} - \hat{\Gamma}_n - \Lambda_1^k - \Lambda_2^k}{2\rho} \right\|_F^2 \Bigg\} \tag{A.16}$$

Let us define the operator

$$\tilde{T}_\rho(A) = \underset{\Theta}{\arg\min} \left\{ -\log\det\Theta + \rho\|\Theta - A\|_F^2 \right\} = \frac{1}{2}(A + (A^2 + (2/\rho)I)^{1/2}) \tag{A.17}$$

whose solution is derived in Section 2.2.1 if we set $R = \infty$. Then the solution to (A.12) is $\tilde{T}_\rho((\rho V_1^{k+1} + \rho V_2^{k+1} - \hat{\Gamma}_n - \Lambda_1^k - \Lambda_2^k)/(2\rho))$.

Using these results, the algorithm in (A.10)-(A.14) becomes

$$\begin{aligned}
V_1^{k+1} &= \mathrm{Prox}_{g_\lambda/\rho}\left(\frac{\rho\Theta^k + \Lambda_1^k}{\rho}\right) \\
V_2^{k+1} &= \mathrm{Proj}_{\mathcal{X}_{h,R}}\left(\frac{\rho\Theta^k + \Lambda_2^k}{\rho}\right) \\
\Theta^{k+1} &= \tilde{T}_\rho\left(\frac{\rho V_1^{k+1} + \rho V_2^{k+1} - \hat{\Gamma}_n - \Lambda_1^k - \Lambda_2^k}{2\rho}\right) \\
\Lambda_1^{k+1} &= \Lambda_1^k + \rho(\Theta^{k+1} - V_1^{k+1}) \\
\Lambda_2^{k+1} &= \Lambda_2^k + \rho(\Theta^{k+1} - V_2^{k+1})
\end{aligned} \tag{A.18}$$

Convergence

Analogues to Propositions 1 and 2 can also be shown for this algorithm using similar methods. To do this, we first note that we can rewrite the optimization problem (A.8) as

$$\min_{\Theta, V}\ \phi(\Theta, V) = \min_{\Theta, V} \{f_1(\Theta) + f_2(V)\}\ \text{ s.t. }\ A\mathrm{vec}(V) + B\mathrm{vec}(\Theta) = 0 \tag{A.19}$$

where

$$f_1(\Theta) = \mathrm{tr}(\hat{\Gamma}_n\Theta) - \log\det(\Theta) \qquad f_2(V) = g_\lambda(A_1V) + \mathbb{1}_{\mathcal{X}_{h,R}}(A_2V)$$

and

$$A = -I_{2m^2} \qquad B = \begin{pmatrix} I_{m^2} \\ I_{m^2} \end{pmatrix}$$

$$V = \begin{pmatrix} V_1 \\ V_2 \end{pmatrix} \qquad A_1 = \begin{pmatrix} I_m & 0 \end{pmatrix} \qquad A_2 = \begin{pmatrix} 0 & I_m \end{pmatrix}$$

This results in the following augmented Lagrangian that is equivalent to (A.9).

$$\mathcal{L}_\rho(\Theta, V, \Lambda) = f_1(\Theta) + f_2(V) + \frac{\rho}{2}\|B\Theta + AV\|_F^2 + \langle \Lambda, B\Theta + AV \rangle$$

Even though we present our algorithm as a three-block ADMM in Section A.2, this formulation makes it clear that we are using a two-block splitting scheme where (A.10) and (A.11) are the separable subproblems of the V-step.

Showing similar convergence results to Propositions 1 and 2 can then be done using the same techniques as in Section A.1

A.3 Additional simulation results

Penalty nonconvexity and R

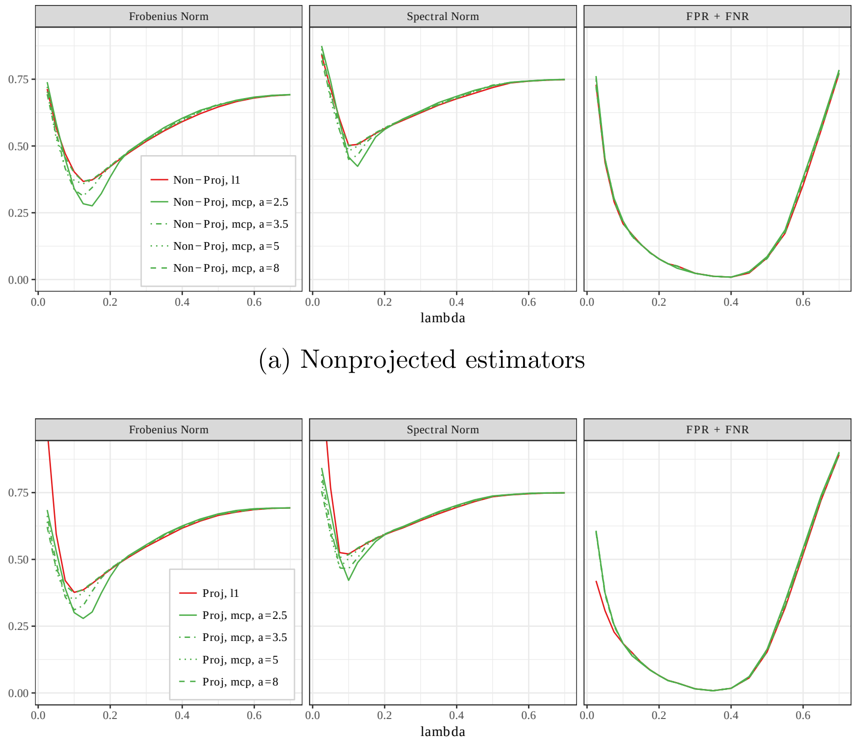
Suppose g_λ is μ-weakly convex and $R \leq \sqrt{\frac{2}{\mu}}$. Then, as shown in Lemma 6 of *Loh and Wainwright* (2017), the overall objective function is strictly convex over the feasible set, and Proposition 2 therefore shows that any limiting point of ADMM algorithm corresponds to the unique global optimum of the objective. However, this choice of R radius on the $\|\cdot\|_2$ side constraint is quite restrictive. In particular, since we also require $R \geq \|\Theta^*\|_2$ we therefore need to choose large values of a in the SCAD or MCP penalties to make μ small enough, which means in practice we simply recover the performance of the ℓ_1 penalized methods. Though *Loh and Wainwright* (2017) show statistical properties for when the parameters are chosen satisfying this condition, in practice we can often do better by allowing the objective to be nonconvex even though no global optimum will exist.

Once we relax this condition ($R > \sqrt{2/\mu}$), the objective becomes nonconvex, and Proposition 2 simply shows that any limiting point of our ADMM algorithm will be a stationary point of the objective. In our simulations, we generally set μ and R such that this condition is violated, and yet we show that our algorithm still results in good estimators. In fact, Figure A.1 demonstrates how, in practice, choosing μ such that this condition is met tends to eliminate the advantages that nonconvex penalties provide. Here the choice of $a = 8$ is the only one that satisfies the condition, and this choice has identical performance as the ℓ_1 penalty. Using a smaller value of a violates this condition but allows the estimator to take advantage of the unbiasedness of the penalty, resulting in better performance in this setting.

Note that for both of these cases, our ADMM algorithm provides a new feasible method of implementing estimation of this type of side-constrained graphical Lasso objective. This consideration is related to tuning, where satisfying the (R, μ) condition allows the support recovery without incoherence statistical results of *Loh and Wainwright* (2017) but in practice results in suboptimal performance, as the nonconvex penalties have to be chosen such that they lose their unbiased advantage over the ℓ_1 penalty.

Method comparisons

Tables A.1-A.3 present more detailed comparison based on the models from the Kronecker sum (KS) and the missing data (MD) models. We compared performance in terms of relative Frobenius and nuclear norm to the true precision matrix, as well as false positive rate plus false negative rate (FPRFNR). The Kronecker sum results are reported for two sample sizes and two values of the noise parameter τ_B, while the missing data results are reported for



(a) Nonprojected estimators

(b) Projected estimators

Figure A.1: Comparing the performance of the graphical Lasso estimators as a (and therefore the weak convexity constant μ) is changed. Here we present the results using the MCP penalty, so $\mu = 1/a$. We set R to be the oracle Note that $a = 8$ is the only value of a that satisfies the $R \leq \sqrt{2/\mu}$ condition from *Loh and Wainwright* (2017). Data is from a missing data model with $A = \mathrm{AR1}(0.6)$, $m = 400$, $n = 80$, and $\zeta = 0.9$.

two covariance models and three settings of the sample size and sampling rate ζ.[2]

Comparing the projected and nonprojected methods, we see that these two methods are fairly competitive. In terms of model selection, the nonprojected methods tend to perform similarly or better than the projected methods. This improvement is particularly evident in the $n = 80$ settings in Table A.1. If we focus on the methods using the ℓ_1 penalty, the nonprojected method performs at least similarly and sometimes significantly better than the projected method in terms of norm error. The lower sampling rate regime in Tables A.2 and A.3 shows this trend as well. Overall these results suggest a small but sometimes

[2] Note that in the initial covariance estimator for the missing data model the effective sample size for estimating an off-diagonal element of the covariance is $n\zeta^2$; four settings are designed to keep this effective sample size roughly constant while changing the sampling rate ζ. The effective sample sizes for the $n = 80$, $n = 130$, and $n = 250$ settings are 64.8, 63.7, and 62.5, respectively.

Table A.1: The relative norm error and FPR + FNR performance of the Kronecker sum estimator using different methods. Here we set A to be from an AR(0.5) model and choose B from an Erdos-Renyi random graph. We set $m = 400$ and let $\tau_B = 0.5$. Metrics are reported as the minimum value over a range of penalty parameters λ. The MCP penalty is chosen with $a = 2.5$, and we set $R = 1.5\|A\|_2$.

n	τ_B	method	penalty	Frobenius	Spectral	Nuclear	FPRFNR
80	0.3	Nonproj	ℓ_1	0.422	0.598	0.406	0.107
			MCP	0.450	0.613	0.422	0.106
		Proj	ℓ_1	0.424	0.610	0.411	0.113
			MCP	0.444	0.616	0.429	0.111
		Nodewise	ℓ_1	0.391	0.517	0.383	0.130
160	0.3	Nonproj	ℓ_1	0.342	0.509	0.327	0.013
			MCP	0.363	0.518	0.345	0.013
		Proj	ℓ_1	0.356	0.525	0.343	0.016
			MCP	0.341	0.493	0.321	0.015
		Nodewise	ℓ_1	0.288	0.429	0.280	0.017
80	0.5	Nonproj	ℓ_1	0.469	0.642	0.452	0.174
			MCP	0.481	0.659	0.458	0.177
		Proj	ℓ_1	0.464	0.651	0.450	0.194
			MCP	0.483	0.658	0.467	0.197
		Nodewise	ℓ_1	0.466	0.600	0.455	0.250
160	0.5	Nonproj	ℓ_1	0.389	0.573	0.369	0.052
			MCP	0.422	0.596	0.393	0.054
		Proj	ℓ_1	0.407	0.593	0.384	0.056
			MCP	0.399	0.587	0.377	0.055
		Nodewise	ℓ_1	0.358	0.538	0.349	0.083

significant advantage for the nonprojected methods, supporting the idea that the projected methods pay a cost in terms of efficiency due to the loss of information in the projection.

There is no significant difference in model selection between MCP and the ℓ_1 penalty. In fact, the different penalties perform almost identically across scenarios regardless of the ℓ_∞-projection step. Intuitively, the primary benefit of nonconvex penalties is their ability to more accurately estimate large entries, which are easy for the estimators to select.

In terms of norm error, however, there are significant differences depending on the indefiniteness of the optimization problem. Table A.4 reports some statistics on the eigenspectrum

Table A.2: The relative norm error and FPR + FNR performance of the missing data estimator using different methods. Here we set A to be from an AR(0.6) model and set $m = 400$. Recall that ζ is the sampling rate. Metrics are reported as the minimum value over a range of penalty parameters λ. The MCP penalty is chosen with $a = 2.5$, and we set R to be 1.5 times the oracle value for each method.

A Model	n	ζ	method	penalty	Frobenius	Spectral	Nuclear	FPRFNR
AR(0.6)	80	0.9	Nonproj	ℓ_1	0.367	0.506	0.363	0.0089
				MCP	0.308	0.533	0.296	0.0088
			Proj	ℓ_1	0.377	0.520	0.375	0.0085
				MCP	0.308	0.527	0.284	0.0083
			Nodewise	ℓ_1	0.292	0.487	0.280	0.0097
	130	0.7	Nonproj	ℓ_1	0.397	0.597	0.388	0.017
				MCP	0.384	0.632	0.363	0.016
			Proj	ℓ_1	0.417	0.599	0.407	0.019
				MCP	0.348	0.626	0.326	0.018
			Nodewise	ℓ_1	0.356	0.592	0.347	0.029
	250	0.5	Nonproj	ℓ_1	0.420	0.619	0.403	0.028
				MCP	0.457	0.680	0.436	0.026
			Proj	ℓ_1	0.437	0.626	0.429	0.031
				MCP	0.391	0.600	0.369	0.032
			Nodewise	ℓ_1	0.412	0.632	0.400	0.078
	700	0.3	Nonproj	ℓ_1	0.431	0.633	0.411	0.043
				MCP	0.505	0.718	0.470	0.040
			Proj	ℓ_1	0.450	0.644	0.431	0.034
				MCP	0.422	0.664	0.391	0.031
			Nodewise	ℓ_1	0.555	0.704	0.517	0.131

of the input matrix. Nonprojected methods with MCP tends to perform relatively better than its ℓ_1 counterpart if the input matrix is close to the positive semidefinite space. Simulation results from the missing data model Tables A.2 and A.3 further support this relationship between the most negative eigenvalue and the relative performance. Here we see how the MCP nonprojected estimator goes from being significantly better than its ℓ_1 counterpart in terms of Frobenius error in the $\zeta = 0.9$ case to significantly worse when $\zeta = 0.5$. In the projected case, which projects away this indefinite issue, the MCP estimator consistently outperforms its ℓ_1 counterpart in terms of Frobenius error.

The nonconvexity of the penalty interacts poorly with indefiniteness of the input matrix.

Table A.3: The relative norm error and FPR + FNR performance of the missing data estimator using different methods. Here we set A to be from an Erdos-Renyi random graph and set $m = 400$. Recall that ζ is the sampling rate. Metrics are reported as the minimum value over a range of penalty parameters λ. The MCP penalty is chosen with $a = 2.5$, and we set R to be 1.5 times the oracle value for each method.

A Model	n	ζ	method	penalty	Frobenius	Spectral	Nuclear	FPRFNR
ER	80	0.9	Nonproj	ℓ_1	0.398	0.426	0.369	0.133
				MCP	0.379	0.444	0.355	0.132
			Proj	ℓ_1	0.405	0.420	0.375	0.129
				MCP	0.367	0.383	0.346	0.126
			Nodewise	ℓ_1	0.349	0.357	0.334	0.160
	130	0.7	Nonproj	ℓ_1	0.409	0.495	0.372	0.137
				MCP	0.410	0.562	0.372	0.137
			Proj	ℓ_1	0.423	0.497	0.385	0.135
				MCP	0.388	0.465	0.354	0.131
			Nodewise	ℓ_1	0.372	0.463	0.346	0.194
	250	0.5	Nonproj	ℓ_1	0.421	0.556	0.379	0.163
				MCP	0.463	0.680	0.401	0.170
			Proj	ℓ_1	0.437	0.556	0.394	0.163
				MCP	0.406	0.535	0.364	0.171
			Nodewise	ℓ_1	0.431	0.654	0.376	0.241
	700	0.3	Nonproj	ℓ_1	0.427	0.604	0.383	0.193
				MCP	0.485	0.701	0.415	0.189
			Proj	ℓ_1	0.445	0.575	0.401	0.184
				MCP	0.423	0.638	0.380	0.191
			Nodewise	ℓ_1	0.500	0.719	0.413	0.276

When the ℓ_1 penalty is used, it is better able to "control" the indefiniteness of the input due to its linear scaling, resulting in better norm error performance. The nonconvex penalty's inability to resolve the indefiniteness issue results in a degradation of its relative performance as the input matrix becomes more indefinite.

Turning to the nodewise estimator, we see similar patterns. Again referring to Table A.4, it seems that the relative performance of the nodewise estimator varies significantly with the indefiniteness of the input matrix. When the input matrix is closer to positive semidefinite, such as the $n = 160$ situations in Table A.1 or the $\zeta = 0.9$ cases in Tables A.2 and A.3, it performs comparably in terms of model selection and significantly better in terms of norm

Table A.4: Measures of the indefiniteness of the input matrix $\hat{\Gamma}_n$. σ_i denote the eigenvalues of $\hat{\Gamma}_n$, while σ_i^+ denote the eigenvalues of $\hat{\Gamma}_n^+$ as defined in Section 2.3.1. We set $m = 400$. For data generated from each model, we report the most negative eigenvalue, the maimum eigenvalues of both the nonprojected and projected sample covariances, the sum of the negative eigenvalues, and the number of negative eigenvalues.

Model	A	n	$\min \sigma_i$	$\max \sigma_i$	$\max \sigma_i^+$	$\sum_{\sigma_i<0} \sigma_i$	$\#\{\sigma_i < 0\}$
KS	AR(0.5)	$n = 80, \tau_B = 0.3$	-0.51	17.0	15.3	-100.5	320
		$n = 160, \tau_B = 0.3$	-0.42	10.3	9.6	-74.1	240
		$n = 80, \tau_B = 0.5$	-0.93	21.3	18.1	-170.1	320
		$n = 160, \tau_B = 0.5$	-0.78	12.0	10.7	-124.6	243
MD	AR(0.6)	$n = 80, \zeta = 0.9$	-0.26	14.2	13.6	-36.2	320
		$n = 130, \zeta = 0.7$	-0.63	12.3	11.0	-116.6	270
		$n = 250, \zeta = 0.5$	-1.19	11.4	9.7	-183.6	218
		$n = 700, \zeta = 0.3$	-2.17	9.2	7.5	-228.9	188
	ER	$n = 80, \zeta = 0.9$	-0.26	13.4	12.7	-36.6	320
		$n = 130, \zeta = 0.7$	-0.62	11.7	10.4	-116.7	270
		$n = 250, \zeta = 0.5$	-1.20	10.3	8.7	-180.7	214
		$n = 700, \zeta = 0.3$	-2.17	8.5	6.9	-223.0	184

error. But when the input matrix is very indefinite, such as the $\zeta = 0.5$ cases in Tables A.2 and A.3 its relative performance quickly degrades.

BIBLIOGRAPHY

Abramowitz, A. I., and K. L. Saunders (2008), Is polarization a myth?, *The Journal of Politics*, *70*(2), 542–555.

Agarwal, A., S. Negahban, and M. J. Wainwright (2012), Fast global convergence of gradient methods for high-dimensional statistical recovery, *The Annals of Statistics*, *40*(5), 2452–2482.

Allen, G. I., and R. Tibshirani (2010), Transposable regularized covariance models with an application to missing data imputation, *The Annals of Applied Statistics*, *4*(2), 764–790.

Ashford, J., and R. Sowden (1970), Multi-variate probit analysis, *Biometrics*, pp. 535–546.

Banerjee, O., L. E. Ghaoui, and A. d'Aspremont (2008), Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data, *Journal of Machine Learning Research*, *9*(Mar), 485–516.

Belloni, A., M. Rosenbaum, and A. B. Tsybakov (2017), Linear and conic programming estimators in high dimensional errors-in-variables models, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *79*(3), 939–956.

Bertsekas, D. P., and J. N. Tsitsiklis (1989), *Parallel and distributed computation: Numerical methods*, Prentice Hall Englewood Cliffs, NJ, republished by Athena Scientific in 1997.

Bickel, P. J., and E. Levina (2008), Covariance regularization by thresholding, *The Annals of Statistics*, *36*(6), 2577–2604.

Blum, A., A. Coja-Oghlan, A. Frieze, and S. Zhou (2007), Separating populations with wide data: A spectral analysis, in *International Symposium on Algorithms and Computation*, pp. 439–451, Springer.

Boyd, S., N. Parikh, E. Chu, B. Peleato, and J. Eckstein (2010), Distributed optimization and statistical learning via the alternating direction method of multipliers, *Foundations and Trends® in Machine Learning*, *3*(1), 1–122.

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

Buchanan, A. M., and A. W. Fitzgibbon (2005), Damped Newton algorithms for matrix factorization with missing data, in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, pp. 316–322, IEEE.

Cai, T., and W. Liu (2011), Adaptive thresholding for sparse covariance matrix estimation, *Journal of the American Statistical Association*, *106*(494), 672–684.

Chen, X., and W. Liu (2019), Graph estimation for matrix-variate Gaussian data, *Statistica Sinica*, *29*, 479–504.

Chen, X., D. Yang, Y. Xu, Y. Xia, D. Wang, and H. Shen (2020), Testing and support recovery of correlation structures for matrix-valued observations with an application to stock market data, *arXiv preprint arXiv:2006.16501*.

Chib, S., and E. Greenberg (1998), Analysis of multivariate probit models, *Biometrika*, *85*(2), 347–361.

Datta, A., and H. Zou (2017), Cocolasso for high-dimensional error-in-variables regression, *The Annals of Statistics*, *45*(6), 2400–2426.

Dutilleul, P. (1999), The MLE algorithm for the matrix normal distribution, *Journal of statistical computation and simulation*, *64*(2), 105–123.

Efron, B. (2009), Are a set of microarrays independent of each other?, *Ann. App. Statist.*, *3*(3), 922–942.

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

Fan, J., L. Xue, and H. Zou (2014), Strong oracle optimality of folded concave penalized estimation, *Annals of Statistics*, *42*(3), 819.

Fan, J., H. Liu, Y. Ning, and H. Zou (2017), High dimensional semiparametric latent graphical model for mixed data, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *2*(79), 405–421.

Fan, R., B. Jang, Y. Sun, and S. Zhou (2019), Precision matrix estimation with noisy and missing data, in *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2810–2819.

Farrell, R. H. (1985), Multivariate calculation: Use of the continuous groups.

Feng, H., and Y. Ning (2019), High-dimensional mixed graphical model with ordinal data: Parameter estimation and statistical inference, in *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 654–663.

Friedman, J., T. Hastie, and R. Tibshirani (2008), Sparse inverse covariance estimation with the graphical lasso, *Biostatistics*, *9*(3), 432–441.

Galecki, A. T. (1994), General class of covariance structures for two or more repeated factors in longitudinal data analysis, *Communications in Statistics-Theory and Methods*, *23*(11), 3105–3119.

Genton, M. G. (2007), Separable approximations of space-time covariance matrices, *Environmetrics: The official journal of the International Environmetrics Society*, *18*(7), 681–695.

Glanz, H., and L. Carvalho (2018), An expectation–maximization algorithm for the matrix normal distribution with an application in remote sensing, *Journal of Multivariate Analysis*, *167*, 31–48.

Greenewald, K., S. Park, S. Zhou, and A. Giessing (2017), Time-dependent spatially varying graphical models, with application to brain fMRI data analysis, in *Advances in Neural Information Processing Systems*, pp. 5834–5842.

Guo, J., J. Cheng, E. Levina, G. Michailidis, and J. Zhu (2015a), Estimating heterogeneous graphical models for discrete data with an application to roll call voting, *The Annals of Applied Statistics*, *9*(2), 821.

Guo, J., E. Levina, G. Michailidis, and J. Zhu (2015b), Graphical models for ordinal data, *Journal of Computational and Graphical Statistics*, *24*(1), 183–204.

Guo, X., and C. Zhang (2017), The effect of L_1 penalization on condition number constrained estimation of precision matrix, *Statistica Sinica*, *27*, 1299–1317.

Han, J.-H. (2007), Analysing roll calls of the European Parliament: A Bayesian application, *European Union Politics*, *8*(4), 479–507.

Hare, C., and K. T. Poole (2014), The polarization of contemporary American politics, *Polity*, *46*(3), 411–429.

Hatfield, L. A., and A. M. Zaslavsky (2018), Separable covariance models for health care quality measures across years and topics, *Statistics in Medicine*, *37*(12), 2053–2066.

Hong, M., Z.-Q. Luo, and M. Razaviyayn (2016), Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems, *SIAM Journal on Optimization*, *26*(1), 337–364.

Hornstein, M., R. Fan, K. Shedden, and S. Zhou (2019), Joint mean and covariance estimation with unreplicated matrix-variate data, *Journal of the American Statistical Association*, *114*(526), 682–696.

Hsieh, C.-J., M. A. Sustik, I. S. Dhillon, and P. Ravikumar (2014), QUIC: Quadratic approximation for sparse inverse covariance estimation., *Journal of Machine Learning Research*, *15*(1), 2911–2947.

Hwang, J. T. (1986), Multiplicative errors-in-variables models with applications to recent data released by the US Department of Energy, *Journal of the American Statistical Association*, *81*(395), 680–688.

Iyengar, S., Y. Lelkes, M. Levendusky, N. Malhotra, and S. J. Westwood (2019), The origins and consequences of affective polarization in the United States, *Annual Review of Political Science*, *22*, 129–146.

Jamshidian, M., and P. M. Bentler (1999), ML estimation of mean and covariance structures with missing data using complete data routines, *Journal of Educational and behavioral Statistics*, *24*(1), 21–24.

Kolar, M., and E. P. Xing (2008), Improved estimation of high-dimensional ising models, *arXiv preprint arXiv:0811.1239*.

Kolar, M., L. Song, A. Ahmed, and E. P. Xing (2010), Estimating time-varying networks, *The Annals of Applied Statistics*, *4*(1), 94–123.

Leng, C., and C. Y. Tang (2012), Sparse matrix graphical models, *Journal of the American Statistical Association*, *107*(499), 1187–1200.

Little, R. J., and D. B. Rubin (2014), *Statistical analysis with missing data*, John Wiley & Sons.

Liu, H., F. Han, M. Yuan, J. Lafferty, and L. Wasserman (2012), High-dimensional semiparametric Gaussian copula graphical models, *The Annals of Statistics*, *40*(4), 2293–2326.

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

Loh, P.-L., and M. J. Wainwright (2013), Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses, *The Annals of Statistics*, *41*(6), 3022.

Loh, P.-L., and M. J. Wainwright (2015), Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima, *Journal of Machine Learning Research*, *16*, 559–616.

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

Lounici, K. (2014), High-dimensional covariance matrix estimation with missing observations, *Bernoulli*, *20*(3), 1029–1058.

Mazumder, R., and T. Hastie (2012), Exact covariance thresholding into connected components for large-scale graphical lasso, *The Journal of Machine Learning Research*, *13*(1), 781–794.

Meinshausen, N., and P. Bühlmann (2006), High-dimensional graphs and variable selection with the lasso, *The Annals of Statistics*, pp. 1436–1462.

Mota, J. F., J. M. Xavier, P. M. Aguiar, and M. Püschel (2011), A proof of convergence for the alternating direction method of multipliers applied to polyhedral-constrained functions, *arXiv preprint arXiv:1112.2295*.

Naik, D. N., and S. S. Rao (2001), Analysis of multivariate repeated measures data with a Kronecker product structured covariance matrix, *Journal of Applied Statistics*, *28*(1), 91–105.

Ng, A. Y., M. I. Jordan, and Y. Weiss (2002), On spectral clustering: Analysis and an algorithm, in *Advances in neural information processing systems*, pp. 849–856.

Park, S. (2016), Selected problems for high-dimensional data–quantile and errors-in-variables regressions, Ph.D. thesis, University of Michigan.

Park, S., K. Shedden, and S. Zhou (2017), Non-separable covariance models for spatio-temporal data, with applications to neural encoding analysis, *arXiv preprint arXiv:1705.05265*.

Park, S., X. Wang, and J. Lim (2020), Estimating high-dimensional covariance and precision matrices under general missing dependence, *arXiv preprint arXiv:2006.04632*.

Pavez, E., and A. Ortega (2019), Covariance matrix estimation with non uniform and data dependent missing observations, *arXiv preprint arXiv:1910.00667*.

Qiu, H., F. Han, H. Liu, and B. Caffo (2016), Joint estimation of multiple graphical models from high dimensional time series, *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *78*(2), 487.

Ravikumar, P., M. J. Wainwright, and J. D. Lafferty (2010), High-dimensional Ising model selection using l1-regularized logistic regression, *The Annals of Statistics*, *38*(3), 1287–1319.

Ravikumar, P., M. J. Wainwright, G. Raskutti, and B. Yu (2011), High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence, *Electronic Journal of Statistics*, *5*, 935–980.

Rosenbaum, M., and A. B. Tsybakov (2010), Sparse recovery under matrix uncertainty, *The Annals of Statistics*, pp. 2620–2651.

Rosenbaum, M., and A. B. Tsybakov (2013), Improved matrix uncertainty selector, in *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*, pp. 276–290, Institute of Mathematical Statistics.

Rothman, A. J., P. J. Bickel, E. Levina, and J. Zhu (2008), Sparse permutation invariant covariance estimation, *Electronic Journal of Statistics*, *2*, 494–515.

Rudelson, M., and R. Vershynin (2013), Hanson-Wright inequality and sub-gaussian concentration, *Electronic Communications in Probability*, *18*.

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

Shvartsman, M., N. Sundaram, M. Aoi, A. Charles, T. Willke, and J. Cohen (2018), Matrix-normal models for fMRI analysis, in *International Conference on Artificial Intelligence and Statistics*, pp. 1914–1923.

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

Städler, N., D. J. Stekhoven, and P. Bühlmann (2014), Pattern alternating maximization algorithm for missing data in high-dimensional problems, *The Journal of Machine Learning Research*, *15*(1), 1903–1928.

Stekhoven, D. J., and P. Bühlmann (2011), Missforest: Non-parametric missing value imputation for mixed-type data, *Bioinformatics*, *28*(1), 112–118.

Suggala, A. S., E. Yang, and P. Ravikumar (2017), Ordinal graphical models: A tale of two approaches, in *International Conference on Machine Learning*, pp. 3260–3269.

Troyanskaya, O., M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman (2001), Missing value estimation methods for dna microarrays, *Bioinformatics*, *17*(6), 520–525.

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

Van Buuren, S., and K. Oudshoorn (1999), *Flexible multivariate imputation by MICE*, Leiden: TNO.

Viallon, V., O. Banerjee, E. Jougla, G. Rey, and J. Coste (2014), Empirical comparison study of approximate methods for structure selection in binary graphical models, *Biometrical Journal*, *56*(2), 307–331.

Wang, D., X. Liu, and R. Chen (2019), Factor models for matrix-valued high-dimensional time series, *Journal of Econometrics*, *208*(1), 231–248.

Wang, Y., W. Yin, and J. Zeng (2015), Global convergence of ADMM in nonconvex nonsmooth optimization, *arXiv preprint arXiv:1511.06324*.

Wang, Y., J. Kang, P. B. Kemmer, and Y. Guo (2016), An efficient and reliable statistical method for estimating functional connectivity in large scale brain networks using partial correlation, *Frontiers in Neuroscience*, *10*, 123.

Werner, K., M. Jansson, and P. Stoica (2008), On estimation of covariance matrices with Kronecker product structure, *IEEE Transactions on Signal Processing*, *56*(2), 478–491.

Wiberg, T. (1976), Computation of principal components when data are missing, in *Proc. Second Symp. Computational Statistics*, pp. 229–236.

Yin, J., and H. Li (2012), Model selection and estimation in the matrix normal graphical model, *Journal of multivariate analysis*, *107*, 119–140.

Yuan, M. (2010), High dimensional inverse covariance matrix estimation via linear programming, *Journal of Machine Learning Research*, *11*(Aug), 2261–2286.

Yuan, M., and Y. Lin (2007), Model selection and estimation in the Gaussian graphical model, *Biometrika*, *94*(1), 19–35.

Zhang, C.-H. (2010), Nearly unbiased variable selection under minimax concave penalty, *The Annals of Statistics*, *38*(2), 894–942.

Zhang, C.-H., and T. Zhang (2012), A general theory of concave regularization for high-dimensional sparse estimation problems, *Statistical Science*, pp. 576–593.

Zhou, S. (2014), GEMINI: Graph estimation with matrix variate normal instances, *The Annals of Statistics*, *42*(2), 532–562.

Zhou, S. (2019), Sparse Hanson–Wright inequalities for subgaussian quadratic forms, *Bernoulli*, *25*(3), 1603–1639.

Zhou, S. (2020), The tensor quadratic forms, *arXiv preprint arXiv:*.

Zhou, S., J. Lafferty, and L. Wasserman (2010), Time varying undirected graphs, *Machine Learning*, *80*, 295–319.

Zhou, S., P. Rütimann, M. Xu, and P. Bühlmann (2011), High-dimensional covariance estimation based on Gaussian graphical models, *Journal of Machine Learning Research*, *12*(Oct), 2975–3026.