High-dimensional Graphical Models for Noisy Data

Shomoita Alam

Department of Mathematics and Statistics, McGill University, Montreal

April, 2023

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of

Doctor of Philosophy

Abstract

The problem of estimating the inverse covariance or precision matrix for graphical models under a high-dimensional setting is a well-known challenge in modern statistics. Numerous theoretical and applied works have been proposed to date, particularly when the data are fully observed and follow a multivariate normal distribution. However, in the presence of measurement errors, such as additive or multiplicative errors, different surrogate estimates have been suggested in the literature to obtain unbiased estimates of the true covariance matrix.

Unfortunately, these surrogate estimators may not necessarily be positive semi-definite, leading to a non-convex objective function. To address this issue, the surrogate estimators can be projected onto the nearest positive semi-definite matrix, transforming the objective function into a convex problem. While consistency bounds for tail deviations of the estimated and true covariance matrix have been well-studied for fully observed data with sub-Gaussian distributions or bounded moments, such bounds have not been established for the presence of measurement errors.

Therefore, the first part of this thesis focuses on developing consistency bounds for random variables that are sub-Gaussian or have bounded moments in the presence of additive or multiplicative measurement errors. We also perform simulation studies and real data analysis to compare the performance of the covariance projection method with existing methods for precision matrix estimation for corrupted data.

Next, we address the problem of joint estimation of regression coefficients and precision matrix in the presence of missing data, a common issue in genetics. We restrict our attention to the scenario where both the data and measurement error are sub-Gaussian. We employ similar techniques to project the surrogate estimate of the sample covariance matrix to ensure convexity of the objective function and derive consistency bounds. Additionally, we conduct simulation studies to compare our method with existing approaches.

Abrégé

Le problème de l'estimation de la matrice inverse de covariance ou de précision pour les modèles graphiques dans un cadre à haute dimension est un défi bien connu des statistiques modernes. De nombreux travaux théoriques et appliqués ont été proposés à ce jour, en particulier lorsque les données sont entièrement observées et suivent une distribution normale multivariée. Toutefois, en présence d'erreurs de mesure, telles que des erreurs additives ou multiplicatives, différentes estimations de substitution ont été proposées dans la littérature pour obtenir des estimations non biaisées de la véritable matrice de covariance.

Malheureusement, ces estimateurs de substitution ne sont pas nécessairement semi-définis positifs, ce qui conduit à une fonction objective non convexe. Pour résoudre ce problème, les estimateurs de substitution peuvent être projetés sur la matrice semi-définie positive la plus proche, ce qui transforme la fonction objective en un problème convexe. Alors que les limites de cohérence pour les écarts de queue de la matrice de covariance estimée et réelle ont été bien étudiées pour les données entièrement observées avec des distributions sub-gaussiennes ou des moments limités, de telles limites n'ont pas été établies en présence d'erreurs de mesure.

Par conséquent, la première partie de cette thèse se concentre sur le développement de bornes de cohérence pour les variables aléatoires qui sont sous-gaussiennes ou qui ont des moments limités en présence d'erreurs de mesure additives ou multiplicatives. Nous réalisons également des études de simulation et des analyses de données réelles afin de

comparer les performances de la méthode de projection de la covariance avec les méthodes existantes d'estimation de la matrice de précision pour les données corrompues.

Ensuite, nous abordons le problème de l'estimation conjointe des coefficients de régression et de la matrice de précision en présence de données manquantes, un problème courant en génétique. Nous limitons notre attention au scénario où les données et l'erreur de mesure sont sub-gaussiennes. Nous utilisons des techniques similaires pour projeter l'estimation de substitution de la matrice de covariance de l'échantillon afin de garantir la convexité de la fonction objective et de dériver des limites de cohérence. En outre, nous menons des études de simulation pour comparer notre méthode aux approches existantes.

Acknowledgements

I wish to express my utmost gratitude to my supervisors, Professor David A. Stephens and Professor Archer Yang, for their invaluable support and guidance throughout my doctoral studies. Their insights and knowledge have significantly contributed to the quality of my thesis. I am deeply grateful for their continued feedback and encouragement. I would also like to extend my appreciation to Professor Abbas Khalili, a member of my thesis committee, for his insightful comments.

I would also like to thank the Department of Mathematics and Statistics faculty at McGill University for providing me with ample opportunities to strengthen my foundation in mathematics and statistics during my graduate studies.

Furthermore, I am indebted to the Natural Sciences and Engineering Research Council of Canada (NSERC) for their generous financial support. I am also grateful for the financial support provided by my department and supervisors. Additionally, I would like to acknowledge Professor Erica E. M. Moodie from the Department of Epidemiology, Biostatistics, and Occupational Health for her unwavering support and guidance throughout my graduate studies.

I would also like to express my gratitude to my parents for their unwavering support throughout my studies. Additionally, I am thankful to my partner, Zayd Omar, for his constant motivation and assistance in proofreading some of my calculations. I would also like to thank my friends Guanbo Wang, Shouao Wang, and Peng Tang for their constant encouragement and support.

Contribution to Original Knowledge

Chapter 3:

In this chapter, we introduced the Convex Condition Graphical Lasso (CoGlasso) algorithm for estimating the precision matrix in a high-dimensional setting with measurement error, assuming exponential and polynomial-type tail conditions. We presented the distributional properties of the estimators, described the estimation techniques, and established theoretical guarantees under the assumptions made. Theorem 1 provided the elementwise maximum norm bound for the estimation error of the precision matrix for four different scenarios with different tail conditions and measurement error types. Lemmas 1, 2, 3, and 4 were original contributions required to prove Theorem 1. We also conducted extensive simulations and provided a real dataset illustration of the proposed method.

Chapter 4:

In this chapter, we proposed a three-step estimation technique for jointly estimating the precision matrix and regression coefficients in a conditional graphical Lasso setting with missing data. We presented the estimation techniques and established theoretical guarantees for all stages of the estimation. Propositions 1, 2, and 3, and eventually Theorem 3, required the use of Lemmas 13, 14, 15, and 16, which were original contributions. We also provided an algorithm for executing the estimation and demonstrated the proposed method using several synthetic simulation scenarios.

Contribution of Authors

Chapter 3:

The Chapter 3 is a joint work done under the supervision of Professor Archer Yang and Professor David A. Stephens. We defined the CoGlasso algorithm to estimate the precision matrix in the presence of noisy data. The initial problem formulation of Chapter 3 was guided by Professor Yang. I was responsible to establish the methodology and derive the theoretical bounds presented in this chapter with the guidance of Professor Yang and Professor Stephens. Specifically, the proofs of Lemmas 1, 2, 3, 4 and Theorem 1 was derived by myself and then proofread and corrected by Professor Yang and Professor Stephens. I also performed the simulation studies and the data illustrations.

Chapter 4:

The Chapter 4 is a joint work done under the supervision of Professor Yang and Professor Stephens. We proposed a three-step method to jointly estimate the precision matrix and regression coefficients in the presence of missing data in a multivariate regression setting. The initial problem formulation of Chapter 4 was guided by Professor Yang and Professor Stephens. I was responsible for deriving the estimation procedures and recovery rates for each stage of the estimation of the proposed stages. Specifically, I derived Lemmas 13, 14, 15, and 16 along with Proposition 1, 2 and 3 which were required to prove the theoretical bound for regression coefficients given in Theorem 3. All the theoretical results

were proofread and corrected by Professor Stephens and Professor Yang. I also performed the simulation studies to illustrate our methods.

Chapter 4 is a collaboration with Professor Celia Greenwood and Yixiao Zeng from Quantitative Life Sciences. However, the theoretical and computational work presented in this chapter is the original contribution of myself, Professor Yang, and Stephens. Yixiao Zeng independently studied the computational aspect of the model and its application to epigenetic data analysis where the implementation is developed as an R package (`https://github.com/yixiao-zeng/missoNet`). It is important to note that there is no overlap between the work of the two teams in this thesis.

List of Abbreviations

ADMM:	Alternating Direction Method of Multipliers
AIC:	Akaike Information Criterion
BIC:	Bayesian Information Criterion
CoCoLasso:	Convex Conditioned Lasso
CoGlasso:	Convex Condition Graphical Lasso
CLIME:	Constrained ℓ_1-Minimization for Inverse Matrix Estimation
CV:	Cross-validation
DAGs:	Directed Acyclic Graphs
EBIC:	Extended Bayesian Information Criterion
EM:	Expectation Maximization
IPW:	Inverse Probability Weighting
ISTA:	Iterative Soft Thresholding Algorithm
KL divergence:	Kullback-Leibler divergence
MCP:	Minimax Concave Penalty
MCAR:	Missing Completely At Random
MRCE:	Multivariate Regression with Covariance Estimation
NC:	Non-convex
RSC:	Restricted Strong Convexity
SCAD:	Smoothly Clipped Absolute Deviation
StARS:	Stability Approach to Regularization Selection
TB:	Tuberculosis

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Estimation of the inverse covariance matrix, also known as the precision matrix, is one of the fundamental problems in modern multivariate statistics. Many fields such as economics, finance, genomics, medical imaging, health science, social networks etc. require precision matrix estimation as a part of the practical research problems. While covariance matrix encodes marginal correlations between the variables, the precision matrix reveals conditional correlations between pairs of variables given the remaining variables (under certain distributional assumptions). The estimation problem becomes challenging when the dimension of the precision matrix is large, specifically when the number of variables p greatly exceeds the sample size n. One of the particular interests in estimating a precision matrix in the literature lies in the key assumption of sparsity, that is, obtaining an estimate of the precision matrix in which some elements are zero. Different penalized maximum likelihood techniques have been proposed in recent years to consistently estimate the precision matrix. When data are fully observed, these type of estimators have become a standard tool for estimating graphical models under sparsity conditions. We give details in Chapter 2.

Most of the theoretical and applied works in estimating the precision matrix in high-dimensional setting are focused on data that are fully observed under the assumption

that they are drawn independently and identically from some underlying distribution. This is often unrealistic since in real world data, we often find covariates that are measured inaccurately or have missing values. For example, sensor network data (Slijepcevic et al., 2002) tend to be both noisy due to measurement error and partially missing due to failures or drop-outs of sensors, gene expression data (Purdom and Holmes, 2005) and high-throughput sequencing (Benjamini and Speed, 2011) also tend to be noisy due to measurement error. Another common example of corrupted data are when we have missing data which can happen due to non-response in many different fields.

In the case of fully observed or clean data, theoretical properties of precision matrix estimation had been studied rigorously in the literature. In Chapter 2, we discuss them in detail. Ravikumar et al. (2011) studied the theoretical properties of estimating both the covariance matrix $\mathbf{\Sigma}^*$ and its inverse $\mathbf{\Theta}^* = (\mathbf{\Sigma}^*)^{-1}$ under given n i.i.d. observations $\{X_1, \ldots, X_n\}$ of a zero mean random vector $X \in \mathbb{R}^p$ for uncontaminated data. In this work no specific distributional assumptions are imposed specifically on X itself, but in terms of the tail behaviour of the maximum deviation of the sample and population covariance matrices. In Chapter 3 we extend this idea in the case when the data are corrupted by measurement error. Specifically, we look at two types of measurement error scenarios, additive and multiplicative, and derive the deviation bounds under different tail conditions, specifically, exponential-type tail and polynomial-type tail. It is well known that if methods developed for clean data are applied on corrupted data, they would lead to misleading inferences. Many unbiased surrogate estimates have been proposed in literature that can take into account of this measurement error issue (Loh and Wainwright, 2012) in a regression setting. The challenges when data are corrupted or not fully observed are that the estimated covariance matrix does not remain positive semi-definite, especially when $p > n$, and the estimated covariance matrix is guaranteed to have negative eigenvalues. As a result, the objective function does not remain convex and consequently becomes unbounded from below. Loh and Wainwright (2012) proposed to

estimate the precision matrix using a nodewise regression (Meinshausen and Bühlmann, 2006) method for graphical models with an additional constraint on the estimator. Fan et al. (2019) generalized this idea to the penalized likelihood formulation of the Gaussian graphical model with an additional constraint on the precision matrix parameter. We discuss each method in detail in Chapter 2 and 3. These methods are non-convex in nature with an additional side constraint and that require an initial guess about the operator norm of the true parameter of interest. Compared to this method, we propose to project $\widehat{\Sigma}$ to its nearest positive semi-definite matrix, thereby guaranteeing the objective function to be convex. Hence we can avoid performing any non-convex analysis. This idea is established in a regression setting with measurement error being present in the data in Datta and Zou (2017). We develop the theoretical deviation bounds for precision matrix estimation in a graphical model setting for different measurement error scenarios and also compare the empirical results with several other existing methods to estimate precision matrix with corrupted data in Chapter 3. Finally, we provide an example of a real data analysis to demonstrate the application of the method developed in Chapter 3.

Next, in Chapter 4, we study multivariate regression with multiple responses regressed on a single set of prediction variables where the responses would contain noisy observations. Here the goal is twofold; estimating the sparse regression coefficient matrix accounting for correlation of the response variables through estimating the precision matrix of the errors. We discuss the existing approaches to handle this type of problem in Chapter 2 when there is no corruption in the data. Specifically, we assume sub-Gaussian tail behaviour of the maximum deviation of the sample and population covariance matrices of the response, predictors and the measurement error variables, respectively. In terms of contamination, we only consider the case when there is missing data in the response variables. In a high-dimensional setting, penalized estimators of the coefficient matrix are obtained with an assumption that the coefficient matrix is elementwise sparse and the precision matrix of the error is also sparse. Similar to Chapter 3, we run into the estimation

problem of the covariance matrix of the error to be non-positive semi-definite. We provide a three stage solution for the estimation and employ the idea of projecting the estimated covariance of the error to its nearest positive semi-definite matrix, thereby preserving the convexity of the objective function. We develop theoretical guarantees under these assumptions in Chapter 4 and also perform simulation studies to demonstrate our method comparing with another existing method.

The thesis is organized as follows. In Chapter 2, we provide a literature review of estimating the precision matrix under clean data and corrupted data assumptions under sparsity assumptions in high-dimensional setting. We also provide a literature review of the joint estimation of conditional graphical model with multiple responses under both clean and unclean data assumptions. In Chapter 3 and 4, we provide the theoretical basis for our work. We also provide some examples of real data analysis under each of the considered setups in Chapter 3 and 4. Finally, we provide a discussion of all our findings in Chapter 5, and in Chapter 6 a general conclusion and future research directions are discussed.

Chapter 2

Literature Review

In this chapter, we provide an extensive summary of precision matrix estimation in sparse high-dimensional settings. First we discuss the classical methods that are available to estimate sparse precision matrices. We also discuss the literature that has studied different convergence rates under different distributional assumptions. Next, we provide a thorough literature review of sparse precision matrix estimation when there is noise present in the data. Finally, we discuss the literature on estimating the coefficient matrix and precision matrix in a multivariate response linear regression model in the presence of noisy data.

2.1 Classical Methods of Precision Matrix Estimation

Assuming multivariate normality of the observations, the sparsity pattern of the precision matrix determines conditional dependence relationships between the variables. More precisely, if we have n independent observations from a p-dimensional zero mean Gaussian random vector $X := (X_1, \ldots, X_p)^\top$, then the density parameterized by the precision matrix $\mathbf{\Theta}^* := (\mathbf{\Sigma}^*)^{-1} \succ 0$ can be written as

$$f(x_1, , \ldots, , x_p; \mathbf{\Theta}^*) = \frac{1}{\sqrt{(2\pi)^p \det{(\mathbf{\Theta}^*)^{-1}}}} \exp\Big\{ -\frac{1}{2} x^\top \mathbf{\Theta}^* x \Big\}. \tag{2.1}$$

The conditional independence relationship can be characterized by an undirected graph $G := (V, E)$, where the vertex set $V := \{1, \ldots, p\}$ corresponds to the p variables in X, and the edge set E describes the conditional independence between any pair X_j and X_k in X ($j, k \in V$). If $X_j \perp\!\!\!\perp X_k | X_{\setminus\{j,k\}}$ that X_j is conditionally independent of X_k given $X_{\setminus\{j,k\}}$, where $X_{\setminus\{j,k\}} := \{X_i : i \neq j, k\}$, we say that X is Markov with respect to G. The goal of covariance selection is to identify the edges in the set E. Under the Gaussian assumption for X, it is a well-known result that the zero pattern of the precision matrix $\mathbf{\Theta}^* := (\mathbf{\Sigma}^*)^{-1}$ corresponds the edge structure E of the underlying graph. We say that X_j and X_k are conditionally independent given the remaining variables precisely if and only if $\mathbf{\Theta}^*_{jk} = 0$ (Lauritzen, 1996).

As discussed in Pourahmadi (2011), as the number of parameters grows rapidly with the number of variables, the problem relies heavily on regularization. When the sample size n is larger than p, the sample covariance matrix $\widehat{\mathbf{\Sigma}}$ is the maximum likelihood estimator of the $p \times p$ covariance matrix $\mathbf{\Sigma}$ and it optimally converges to $\mathbf{\Sigma}$ at the rate $n^{-1/2}$. However, when $p \gg n$, the sample covariance estimate behaves poorly since the eigenstructure of the matrix gets distorted in the sense that the largest sample eigenvalue will be biased upward and the smallest sample eigenvalue would be biased downward (Johnstone, 2001; Johnstone and Lu, 2009). Imposing regularization therefore has become a standard way to improve the estimator.

In this section, we discuss classical methods that are developed in literature to estimate precision matrix in a sparse setting. When the sample is drawn from a multivariate normal distribution, one of the most widely used approach is *neighbourhood selection* (Meinshausen and Bühlmann, 2006). The neighbourhood $\mathcal{N}_j$ of a node $j \in V$ consists of all nodes $k \in V \setminus \{j\}$ such that $(j, k) \in E$. This regression-based approach provides a sparse estimate of the precision matrix or a Gaussian graphical model by fitting separate Lasso (Tibshirani, 1996) regression to each variable, using the others as predictors. Let $\mathbf{X}_j$ be the jth column of $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $\mathbf{X}_{-j} \in \mathbb{R}^{n \times (p-1)}$. For each variable j, we solve the following optimization

problem

$$\widehat{\theta}(\lambda_n) = \underset{\theta \in \mathbb{R}^{p-1}}{\arg\min} \left(\frac{1}{2n} \|\mathbf{X}_j - \mathbf{X}_{-j}\theta\|^2 + \lambda_n \|\theta\|_1 \right) \tag{2.2}$$

where $\lambda_n > 0$. Then we obtain the neighbourhood estimate $\widehat{\mathcal{N}_j} = \{k \in V \backslash \{j\} | \widehat{\theta}_k \neq 0\}$. This step returns the neighbourhood estimate of each variable. These estimators might be inconsistent, meaning that for a given pair of distinct vertices (j, k), it may be the case that $k \in \widehat{\mathcal{N}_j}$ whereas $j \notin \widehat{\mathcal{N}_k}$. To resolve this, we need to combine the estimates to form an edge estimate $\widehat{E}$ using the OR rule or the AND rule. The OR rule declares that $(j, k) \in \widehat{E}_{\text{OR}}$ if either $k \in \widehat{\mathcal{N}_j}$ or $j \in \widehat{\mathcal{N}_k}$ and the AND rule declares that $(j, k) \in \widehat{E}_{\text{AND}}$ if either $k \in \widehat{\mathcal{N}_j}$ and $j \in \widehat{\mathcal{N}_k}$. This procedure consistently estimates the precision matrix even in the case when the number of variables grow as rapidly as the sample size. However, it does not guarantee to produce a positive definite estimate $\widehat{\mathbf{\Theta}}$. Wainwright (2019) detailed graph selection consistency under the incoherence condition on the population which enforces the requirement that there should be no edge variable that is not included in the graph that is highly correlated with variables within the true edge-set.

Since the idea of Meinshausen and Bühlmann (2006) is simple, it has inspired several other improved sparse estimators of the precision matrix using a penalized likelihood approach with a Lasso penalty on the off-diagonal elements (Banerjee et al., 2008; Friedman et al., 2008; Peng et al., 2009; Rocha et al., 2008; Rothman et al., 2008; Yuan and Lin, 2007). They consider maximizing the penalized log-likelihood over a non-negative definite matrix $\mathbf{\Theta}$

$$\log\det(\mathbf{\Theta}) - \text{tr}(S\mathbf{\Theta}) - \lambda \|\!|\mathbf{\Theta}|\!\|_1 \tag{2.3}$$

where tr is the trace operator, λ is the penalty parameter and $\|\!|\mathbf{\Theta}|\!\|_1 = \sum_{ij} |\theta_{ij}|$ is the ℓ_1-norm, that is, the sum of the absolute values of the elements of the positive definite matrix $\mathbf{\Theta}$. Some authors have omitted the diagonal entries from the penalty and only take the sum of the off-diagonal elements. The objective function in (2.3) is convex. Banerjee et al. (2008) uses a block coordinate descent to solve this problem and Friedman et al. (2008) proposed

a coordinate descent approach. Among these methods, Friedman et al. (2008)'s graphical Lasso is a remarkably fast algorithm that provides a sparse covariance estimator and is guaranteed to be positive definite. Witten et al. (2011) presented a necessary and sufficient condition that uses a block diagonal screening rule to speed up computations considerably. These conditions were also discovered by (Mazumder and Hastie, 2012) independently. The `R` package `glasso` (version 1.7) currently implements this block diagonal screening rule for their Algorithm 2 (Witten et al., 2011).

Rothman et al. (2008) studied convergence rates under the Frobenius norm loss and showed that the rate depends on how sparse the true precision matrix is. Particularly, they showed that consistent estimates can be achieved in Frobenius and spectral norm at the rate $\mathcal{O}(\sqrt{((s+p)\log p)/n})$, where n, p and s are the number of observations, number of nodes and the number of true edges, respectively. They used a fast iterative algorithm to compute the estimator which depends on the Cholesky decomposition of the inverse but produces permutation-invariant estimator.

Yuan (2010) proposed a method for estimating $\mathbf{\Theta}^*$ by replacing the Lasso selection by a Dantzig selector, where they first estimated the ratio between the off-diagonal elements ω_{ij} and the corresponding diagonal element ω_{ii} for each row i, and then estimated the diagonal elements ω_{ii} given the estimated ratios. They also obtained the error bounds on $|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_1$ when the columns of $\mathbf{\Theta}^*$ are bounded in ℓ_1 for sub-Gaussian distributions.

The Lasso penalty produces biases in the estimators asymptotically due to the linear increase of the penalty on regression coefficients, even in a simple regression setting (Fan and Li, 2001). Lam and Fan (2009) studied the theoretical properties of sparse precision matrices estimation and found that the bias presented in the Lasso penalty also arises for sparse precision matrix estimation. They studied the estimation of the precision matrix based on regularizers that are more general than the ℓ_1-norm. For the ℓ_1 regularization case, the obtained the same Frobenius and spectral norm rates as Rothman et al. (2008) and showed that it succeeds in recovering the zero-pattern of $\mathbf{\Theta}^*$ under some scaling of

the number of observations n and the number of true edges s. In general, to tackle the bias issue for ℓ_1 regularization, non-convex penalties are considered under the same normal likelihood model, for example, Smoothly Clipped Absolute Deviation (SCAD) penalty (Fan et al., 2009; Fan and Li, 2001) and adaptive Lasso penalty (Zou, 2006).

We closely followed but then significantly extended the theoretical development in Ravikumar et al. (2011) while developing our proposed method and therefore these conditions are explained in more detail in Chapter 3. Ravikumar et al. (2011) showed that even in the case of non-Gaussian X this estimator is meaningful since it corresponds to minimizing an ℓ_1-penalized log-determinant Bregman divergence which does not require X to be multivariate Gaussian. A function is defined to be of Bregman type if it is strictly convex, continuously differentiable and has bounded level sets (Bregman, 1967; Censor et al., 1997). A Bregman divergence of the form

$$D_g(A \parallel B) = g(A) - g(B) - \langle \nabla g(B), A - B \rangle$$

is induced by functions satisfying these conditions. Since g has to be strictly convex, $D_g(A \parallel B) \geq 0$ for all A and B, with equality holding if and only if $A = B$. The log-determinant barrier function is a Bregman function and the Bregman divergence is given by

$$D_g(A \parallel B) = -\log\det(A) + \log\det(B) - \langle\!\langle B^{-1}, A - B \rangle\!\rangle$$

for any strictly positive definite A and B. Since estimating the precision matrix is essentially conducted by minimizing

$$\min_{\mathbf{\Theta} \succ 0} \{ \langle\!\langle \mathbf{\Theta}, \mathbf{\Sigma}^* \rangle\!\rangle - \log\det(\mathbf{\Theta}) \}$$

with a possible addition of a off-diagonal ℓ_1-regularization term on $\mathbf{\Theta}$ defined as $|||\mathbf{\Theta}|||_{1,\text{off}} = \sum_{i \neq j} |\Theta_{ij}|, i, j = 1, , \ldots, , p$, and the true covariance matrix $\mathbf{\Sigma}^*$ replaced by its empirical estimate such as the sample covariance matrix. Given the regularization constant $\lambda_n > 0$,

the precision matrix can be solved by the ℓ_1-regularized log-determinant

$$\widehat{\mathbf{\Theta}} = \underset{\mathbf{\Theta} \succ 0}{\arg\min} \{ \langle\!\langle \mathbf{\Theta}, \mathbf{\Sigma}^* \rangle\!\rangle - \log\det(\mathbf{\Theta}) + \lambda_n |\!|\!| \mathbf{\Theta} |\!|\!|_{1,\text{off}} \}. \tag{2.4}$$

They show that since this objective function corresponds to the Bregman divergence, it can be used without assuming that X is multivariate Gaussian. When the data are generated from a multivariate Gaussian distribution, the divergence coincides with the ℓ_1-regularized maximum likelihood. As discussed earlier, therefore, the precision matrix becomes more interpretable in terms of conditional independence in that case.

Ravikumar et al. (2010) analyzed the ℓ_1-regularized logistic regression method for Ising model selection. In another work, Ravikumar et al. (2011) obtained the convergence rates in the elementwise ℓ_∞ norm and spectral norm, under more restrictive conditions, such as mutual incoherence or irrepresentable conditions for more general non-Gaussian distributions and under a variety of tail conditions. In this work, Ravikumar et al. (2011) analyzed the performance of the precision matrix estimator under high-dimensional scaling, where the number of nodes in the graph p, the number of edges s, and the maximum node degree (maximum number of non-zeros per row) d, are allowed to grow as a function of the number of observations n. They identified key quantities that measure model complexity such as the ℓ_∞-operator norm of the true covariance matrix $\mathbf{\Sigma}^*$, the Hessian of the log-determinant of the objective function, $\mathbf{\Gamma}^* = (\mathbf{\Theta}^*)^{-1} \otimes (\mathbf{\Theta}^*)^{-1}$, the sub-matrix $\mathbf{\Gamma}^*_{SS}$ where S indexes the graph edges, a mutual incoherence or irrepresentability measure on the Hessian matrix $\mathbf{\Gamma}^*$ which is similar to the condition imposed on $\mathbf{\Sigma}^*$ in case of Lasso (Wainwright, 2009; Zhao and Yu, 2006) and the rate of decay on the probabilities of the deviation bound between the estimated and true covariance matrix.

Their work establishes consistency of the precision matrix estimator $\widehat{\mathbf{\Theta}}$ in an element-wise maximum-norm sense. They showed that the rate depends on the tail behaviour of the entries of the deviation between the estimated and the true covariance matrix. Con-

vergence rates in Frobenius and spectral norms are derived for the special cases when X is sub-Gaussian and when X has bounded $4m$th moment. Specifically, for the sub-Gaussian case, they showed consistency of the estimator under the spectral norm at rate $|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_2 = \mathcal{O}(\sqrt{\min\{d^2 \log p, (s+p)\log p\}})$, with high probability when d, s and p are the maximal degree, number of edges and number of nodes of the graph, respectively. When X has bounded $4m$th moment, they obtained consistency of the estimator under the spectral norm at rate $|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_2 = \mathcal{O}(dp^{1/2m}/\sqrt{n})\}$, which turned out to be slower than exponential-type tail behaviour indicating that the logarithm dependence on the model size p is linked to particular tail behaviour of the distribution of X. They also compared their convergence rates with past research (Lam and Fan, 2009; Rothman et al., 2008) and found equivalent results when maximal node degree $d^2 \geq s$ and improvements when $d = o(\sqrt{s})$. Finally, they showed that the estimator $\widehat{\mathbf{\Theta}}$ correctly specifies the zero pattern of the precision matrix $\mathbf{\Theta}^*$, with probability converging to one.

Cai et al. (2011) introduce a method of constrained ℓ_1-minimization for inverse matrix estimation (CLIME). They studied the estimation of the precision matrix $\mathbf{\Theta}^*$ which is not restricted to a specific sparsity pattern which can be used to recover a wide class of matrices in both theory and application. In particular, they showed that when the population distribution has either exponential-type of polynomial-type tails, the rate of convergence between the estimator and the true s-sparse precision matrix under the spectral norm is $s\sqrt{\log p/n}$. For the specific case of Gaussian graphical models, they compared their work with Ravikumar et al. (2011) that assumes incoherence condition, which is stringent and difficult to check in practice. Cai et al. (2011) established that similar theoretical results can be obtained without assuming the incoherence condition.

In the Bayesian literature, the graphical Lasso has also been studied (Marlin et al., 2012; Marlin and Murphy, 2009). The Bayesian graphical Lasso was developed by Wang (2012), where they imposed a Laplace prior on the off-diagonal entries and an exponential prior on the diagonal entries of the precision matrix independently. They defined the

graphical Lasso estimator to be equivalent to the maximum a posteriori estimation of their chosen model. They have also explored the distributional properties of the graphical Lasso prior distributions. Another work with similar flavour tackling this problem had been derived by Khondker et al. (2013) with difference in algorithmic techniques from Wang (2012). Banerjee and Ghosal (2015) proposed an adjustment with a mixture of a point mass and a Laplace prior to induce exact sparsity, and also derived the optimal posterior contraction rate with respect to the Frobenius norm. In another work, Banerjee and Ghosal (2014) assumed a banding structure on the precision matrix and derived the posterior contraction rate with a G-Wishart prior. Some useful computational methods to make efficient posterior calculation have been proposed by Lenkoski and Dobra (2011) and Mohammadi and Wit (2015).

2.1.1 Tuning Parameter Selection

The appropriate choice of the tuning parameter is critical in order to ensure that the oracle property of the penalized estimator is satisfied. In the context of precision matrix, one desirable property of the estimator is the oracle property as defined in Fan and Li (2001). They demonstrated the oracle properties of precision matrix estimator in Fan et al. (2009). The oracle property consists of two conditions, namely, *sparsity* which means that the true zero entries of the precision matrix are estimated as zero with probability tending to one, and *asymptotic normality*, which implies that the estimators of the non-zero entries of the precision matrix have the same limiting distribution as the maximum likelihood estimator, knowing the true sparsity pattern. The sparsity condition is referred to as *sparsistency* by Lam and Fan (2009).

It is challenging to select the tuning parameter λ, which controls sparsity in $\boldsymbol{\Theta}$ in a high-dimensional setting. For penalized likelihood methods, cross-validation (CV) for the selection of the tuning parameter which is based on a resampling scheme is widely used. Cross-validation requires fitting the model based on different subsets of the observations

multiple times, which increases the computational complexity of this approach. The other common approach to select optimal tuning parameter is based on information criteria. In this section, we discuss the existing standard methods for selecting the tuning parameter λ in precision matrix estimation. Lian (2011) studied the choice of tuning parameter selection when estimating sparse precision matrices using the penalized likelihood approach. They showed that generalized approximate cross-validation (Craven and Wahba, 1978) for tuning parameter selection is more computationally efficient than the traditional methods. For consistency in the selection of nonzero entries in the precision matrix, they used a Bayesian information criterion and showed that it produces consistent model selection in the Gaussian model.

Cross-Validation:

Cross-validation (CV) is a nonparametric methods for estimating prediction error for selecting the tuning parameter in their non-concave penalized likelihood methods. In a K-fold cross-validation, the data are split into training data and validation data, where the training data are used to train the model and it is tested on the validation data. First, it involves randomly dividing the data into K equally sized parts or "folds" and denote the samples in the kth fold by N_k for $k = 1, , \ldots, , K$. Typical choices of K are 5 or 10. For each fold, the model is fitted to the $K - 1$ parts of the data which constitute the training data set and the cross-validation score is calculated on the kth part of the data. This process is repeated K times, with each of the K parts used exactly once as the validation data, and the K estimates of cross-validation score are then combined. The case $K = n$ is called leave-one-out cross-validation. For the ith observation, the fit is computed using all the data except the ith.

In a graphical model setting to estimate the precision matrix, the observed log-likelihood is used as loss function. We can define the observed log-likelihood of an observation X_i given a precision matrix estimate $\widehat{\mathbf{\Theta}}$ as $\ell(X_i; \widehat{\mathbf{\Theta}}) = \log f(X_i, \widehat{\mathbf{\Theta}})$ and calculate the cross-

validation score which we maximize as

$$CV(\lambda) = \sum_{k} \sum_{i \in N_k} \ell(X_i; \widehat{\mathbf{\Theta}}_{\lambda}^{(k)}). \tag{2.5}$$

Then we find the best $\widehat{\lambda}$ that maximizes $CV(\lambda)$. Finally, using the chosen $\widehat{\lambda}$, a final estimator of the precision matrix is calculated using all the data.

The Akaike Information Criterion:

Another existing method for choosing the tuning parameter in penalized likelihood approaches is the Akaike information criterion (AIC) (Akaike, 1973). It was derived as an estimator of the Kullback-Leibler (KL) divergence and it aims to minimize the KL divergence between the true distribution and the estimate from a candidate model. The model selection rule has the form of "in-sample performance plus penalty" and is defined as in precision matrix estimation context

$$AIC(\lambda) = -2\ell_n(\widehat{\mathbf{\Theta}}_{\lambda}) + 2\sum_{i<j} \mathbb{I}(\widehat{\theta}_{ij,\lambda} \neq 0)$$

where $\ell_n(\widehat{\mathbf{\Theta}}_{\lambda})$ is the multivariate Gaussian log-likelihood, evaluated at $\widehat{\mathbf{\Theta}}_{\lambda}$ which is the penalized maximum likelihood estimator of $\mathbf{\Theta}$ for a specific given λ. $\mathbb{I}(\cdot)$ is an indicator function which counts the number of non-zero elements among the $p(p-1)/2$ off-diagonal entries in the upper half of the matrix. The optimal value of the tuning parameter in this case is taken to be the minimizer of the criterion.

Bayesian Information Criterion:

Another model selection criterion is the Bayesian Information Criterion (BIC) developed by Schwarz (1978). Yuan and Lin (2007) used BIC to select the tuning parameter with the ℓ_1 penalty in the estimation of the precision matrix. The BIC arises from the Bayesian

approach to model selection; choosing the model with minimum BIC is equivalent to choosing the model with largest (approximate) posterior probability. In the context of precision matrix estimation, it is defined as

$$BIC(\lambda) = -2\ell_n(\widehat{\mathbf{\Theta}}_\lambda) + \log n \sum_{i<j} \mathbb{I}(\widehat{\theta}_{ij,\lambda} \neq 0) \tag{2.6}$$

where $\ell_n(\widehat{\mathbf{\Theta}}_\lambda)$ is defined similarly as in the definition of AIC. The optimal value of the tuning parameter is taken to be the minimizer of the criterion. BIC uses a stronger penalty than AIC and as n goes to infinity, the BIC will select the true model with probability one if it is one of the models being considered.

For penalized regression estimation of $\mathbf{\Theta}$, Gao et al. (2012) studied the selection of tuning parameter using BIC. They showed consistency of BIC in model selection for using SCAD penalty or adaptive Lasso penalties in precision matrix estimation problem for a fixed p. This refers to sparsistency, that is, BIC with SCAD penalty identifies the sparsity pattern of the true precision matrix with probability approaching to one when n is large. They also showed that a modified BIC with an extra penalty on the dimension p of the precision matrix is consistent when the true edges are included in a bounded subset, if p tends to infinity at a certain rate with the sample size. The modified BIC proposed by Gao et al. (2012) is equivalent to the extended BIC (EBIC) selection criteria proposed by Foygel and Drton (2010) when $\gamma = 1$. Foygel and Drton (2010) adapted EBIC for precision matrix estimation problems from Chen and Chen (2008) who studied it for Gaussian linear models. For some $\gamma > 0$, EBIC is defined as

$$EBIC(\lambda) = -2\ell_n(\widehat{\mathbf{\Theta}}_\lambda) + \Big\{ \log n + 4\gamma \log p \Big\} \sum_{i<j} \mathbb{I}(\widehat{\theta}_{ij,\lambda} \neq 0).$$

Chen and Chen (2008) showed that the traditional BIC is likely to be inconsistent when p is of a larger order than $\sqrt{n}$.

The work by Gao et al. (2012) compared empirical performance of BIC and EBIC to cross-validation and through simulation studies showed that BIC performs better for sparse precision matrix estimation. One should choose a tuning parameter selection procedure based on one's statistical goal. BIC and EBIC tends to perform better due to their selection consistency properties when the goal is to correctly identify the zeros and non-zeroes of the precision matrix. If the goal is to achieve better prediction performance, cross-validation and AIC are better options since they are both estimators of the KL divergence and are asymptotically equivalent under certain assumptions.

Stability Approach to Regularization Selection (StARS):

The previously mentioned tuning parameter selection methods such as K-fold cross-validation, AIC, and BIC work well for low-dimensional problems with good theoretical properties, but they are not best suited for high-dimensional settings. Liu et al. (2010) proposed the method named Stability Approach to Regularization Selection (StARS) which is a stability-based method for choosing the regularization parameter in high-dimensional inference for undirected graphs. In this method, the least amount of regularization is used which simultaneously results into a sparse precision matrix and makes the graph reproducible under random sampling. Specifically, the process starts with large regularization which corresponds to an empty and highly stable graph and gradually decreases the amount of regularization until there is small dissonance between the graphs across the subsamples. The authors showed that under mild condition, StARS achieve sparsistency in terms of graph estimation. That is, the procedure selects all true edges with high probability even when the graph size diverges with the sample size.

2.2 Literature Review for Precision Matrix Estimation with Corrupted Data

In the previous section we have discussed existing methods that are available to estimate conditional dependence graphs and precision matrix for fully observed data under sparsity conditions. Penalized likelihood estimation is a common approach to tackle such problems. In this section, we review the literature for precision matrix estimation in the presence of measurement error, specifically, when the data are contaminated with additive error or multiplicative error. Errors-in-variables models have been extensively studied in regression settings (Carroll et al. (2006); Hwang (1986); Iturria et al. (1999); Xu and You (2007) and references therein). However, these works were not tackling the high-dimensional scenarios where the number of variables p is much larger than sample size n.

One of the special cases of multiplicative measurement error model is the case of missing data (Little and Rubin, 2019). One simple ad hoc approach would be to use only the complete cases which results into a substantial decrease in sample size. Another ad hoc method would impute the missing values by the corresponding mean and use traditional models to solve for the precision matrix, in a graphical model setup. More systematic approaches based on likelihoods are also popular in terms of imputing missing data (Little and Rubin, 2019; Schafer, 1997). Städler and Bühlmann (2012) developed an Expectation Maximization (EM)-based method for sparse inverse covariance matrix estimation in the missing data regime in the multivariate normal case, and used this result to derive an algorithm for sparse linear regression with missing data. They showed that estimation of mean values and covariance matrices becomes difficult when the data are incomplete and no explicit maximization of the likelihood is possible. Their algorithm maximizes a ℓ_1-penalized observed log-likelihood, where the missing data are imputed using EM algorithm. Loh and Wainwright (2012) pointed out that the EM approach proposed by Städler and Bühlmann (2012) becomes possibly non-convex with missing or noisy data

which will lead to optimization problems. As a consequence it becomes difficult to establish theoretical guarantees for the algorithmic counterpart.

Among other notable studies for precision matrix estimation for data with missing values are Kolar and Xing (2012) and Lounici (2014). Lounici (2014) established theoretical results for non-asymptotic bounds for the estimation of covariance matrix involving the Frobenius and spectral norms which are valid for any setting of the sample size, probability of a missing observation and the dimensionality of the covariance matrix. Kolar and Xing (2012) proposed a two stage method that directly estimates a large dimensional precision matrix from data with missing values. To get an estimate of the precision matrix they form an unbiased inverse probability weighting (IPW) estimator of the covariance matrix from available data and then use it as a plug-in estimator in the penalized maximum likelihood objective function for a multivariate Gaussian distribution. The elements of the covariance matrix are calculated in such a way that takes into account of the missing values naturally while using only the observed samples. They also provided rates of convergence for this estimator in the spectral norm, Frobenius norm and elementwise maximum norm. They compared their results with the EM-based method (Städler and Bühlmann, 2012) and showed that it performs favourably.

As mentioned earlier, in a high-dimensional setting, undirected graphs can be estimated using penalized methods defined in (2.4). The true covariance matrix $\mathbf{\Sigma}^*$ is unknown and typically replaced by the sample covariance matrix estimator $\widehat{\mathbf{\Sigma}} = X^\top X/n$ as an input. When the data are not corrupted, this estimator is at least positive semi-definite and the optimization problem remains convex under ℓ_1 regularization. In this setting, it can be shown that for $\lambda > 0$ a unique optimum $\widehat{\mathbf{\Theta}}$ exists with bounded eigenvalues and that the iterates for any descent algorithm will also have bounded eigenvalues (Hsieh et al., 2014). In case of noisy or missing data, the most natural choice of the sample covariance matrix is no longer positive semi-definite and is guaranteed to have negative eigenvalues,

therefore, making the objective functions non-convex. Moreover, the objective function is unbounded from below when $\widehat{\mathbf{\Sigma}}$ has a negative eigenvalue.

The noise-corrected non-convex approach in a regression setting was proposed and analyzed by Loh and Wainwright (2012, 2017). They discussed unbiased surrogate estimators of the sample covariance matrix in both multiplicative and additive noise cases, one of which has also been studied by Xu and You (2007) in case of additive noise in $n > p$ case. The unbiased surrogate estimate is not positive semi-definite and therefore Loh and Wainwright (2012) optimized a non-convex objective function and studied the statistical error associated with any global optimum. We discuss the surrogate estimates for each type of measurement error in detail in Chapter 3. Loh and Wainwright (2012) proposed an algorithm based on projected gradient descent which optimizes a Lasso type objective function with an additional side constraint on the ℓ_1-norm of the regression coefficients. They proved that the algorithm will converge in polynomial time to a small neighbourhood of the set of all global minimizers. In this work, they provided non-asymptotic bounds that hold with high probability. They also showed an application to graphical model where they estimate the precision matrix using nodewise regression (Meinshausen and Bühlmann, 2006) incorporating an additional constraint on the regression coefficients of the objective function. They generalized this work and obtained theoretical results for regularized M-estimators where both loss and penalty functions are allowed to be non-convex (Loh and Wainwright, 2013, 2017). They established that the graphical Lasso using non-convex penalties can be modified to accommodate noisy or missing data by using the unbiased surrogate estimates for the sample covariance estimate. A related corrected form of the Dantzig selector was proposed by Rosenbaum and Tsybakov (2010) in case of sparse regression models.

An alternative approach to the unboundedness of the objective function with non-positive semi-definite input is to project the input matrix $\widehat{\mathbf{\Sigma}}$ to the positive semi-definite cone and then use that as a plug-in estimate for the objective function in (2.4). Specifically,

we can project $\widehat{\mathbf{\Sigma}}$ to its nearest positive semi-definite matrix $\mathbf{\Sigma}$ as,

$$\widetilde{\mathbf{\Sigma}} = \underset{\mathbf{\Sigma} \succeq 0}{\arg\min} \, |||\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}|||_{\max}$$

where $|||\cdot|||_{\max}$ is the elementwise maximum norm. $\widetilde{\mathbf{\Sigma}}$ is known as the *Convex Conditioned Lasso (CoCoLasso)* estimate as proposed by Datta and Zou (2017). Since the input projected matrix is positive semi-definite, consequently the objective function is convex. The projection can be implemented using an alternating direction method of multipliers (ADMM) method (Boyd et al., 2011) as shown by Datta and Zou (2017). This method can handle a general class of corrupted datasets including the cases of additive or multiplicative measurement error and random missing data. It is well known that Lasso enjoys theoretical and computational benefits of convexity, and this also holds for CoCoLasso as well. In Datta and Zou (2017), they derived statistical error bounds for the CoCoLasso estimate and established asymptotic sign-consistent selection properties of CoCoLasso. Their method is advantageous compared to Loh and Wainwright (2012) because the latter did not provide sign-consistency results for the non-convex approach. Another advantage of this method is that it does not require any prior knowledge of the parameters, unlike Loh and Wainwright (2012), since their approach assumes a constraint on the parameter. They also proposed a calibrated cross-validation method for tuning the regularization parameter in their regression setup. Our proposed methodology and theoretical development fundamentally relies on this idea and will be demonstrated in Chapter 3 in detail.

Fan et al. (2019) generalizes the idea of Loh and Wainwright (2012) for sparse precision matrix estimation with corrupted data and developed an ADMM algorithm for efficient estimation. Their approach proposes using non-convex regularizers such as SCAD (Fan and Li, 2001) and minimax concave penalty (MCP) (Zhang, 2010) along with a ℓ_1-regularizer. They compared their proposed method empirically with other existing methods including the projection method (Datta and Zou, 2017) and argued that when the penalties are

non-convex the unboundedness of the objective function still remains and their method deems useful even for such projected methods. In their paper, they consider the objective function

$$\widehat{\boldsymbol{\Theta}} \in \underset{\boldsymbol{\Theta}\succeq 0, |||\boldsymbol{\Theta}|||_2 \leq R}{\arg\min} \{\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Theta}) - \log\det(\boldsymbol{\Theta}) + g_\lambda(\boldsymbol{\Theta})\} \tag{2.7}$$

where $|||\boldsymbol{\Theta}|||_2$ is the spectral norm of the true precision matrix, $\widehat{\boldsymbol{\Sigma}}$ is a surrogate unbiased estimator accounting for the type of measurement error and g_λ is a separable (entry-wise) sparsity inducing penalty function. $\widehat{\boldsymbol{\Sigma}}$ may not be positive semi-definite in the presence of measurement error, resulting in an unbounded objective function. In their proposed non-convex approach solved with the ADMM algorithm, they focus on the estimator of the precision matrix using the operator norm as a side constraint. In this chapter, we compare our proposed method empirically with the method proposed by Fan et al. (2019).

There are some Bayesian approaches that tackle the problem of precision matrix estimation in the presence of measurement error. Byrd et al. (2021) proposed a Bayesian estimator to estimate sparse precision matrices that corrects for measurement error. With the assumption that the variance of the measurement error is known, they treated the unobservable outcomes as missing data and proposed a method to impute them and iteratively estimate the precision matrix. They combined the imputation-regularized optimization algorithm (Liang et al., 2018) and Bayesian regularization for graphical models with unequal shrinkage (Byrd et al., 2021) to formulate a new procedure and prove its consistency. Their method had desirable results compared to other naive approaches. Shi et al. (2021) established a fully Bayesian framework to handle measurement error and established a general result which provides sufficient conditions under which the posterior contraction rates that hold in the no-measurement-error case carry over to the measurement-error case.

Recently, methods have been developed for a more general type of missing dependence structure unlike the simpler cases where every variable of each sample is independently subject to missingness with equal probability. For high-dimensional precision matrix

estimation, Park et al. (2021) studied the theoretical properties of the deviation of the IPW estimators that correct for bias due to missingness under general missing dependency. They provided optimal convergence rates of the estimator based on the elementwise maximum norm, even when the assumptions such as known mean and/or missing probabilities are relaxed.

Öllerer and Croux (2015) proposed different high-dimensional precision matrix estimators that are robust to cellwise contamination. They proved that replacing the sample covariance matrix in the graphical Lasso with an elementwise robust covariance matrix leads to an elementwise robust, sparse precision matrix estimator computable in high-dimensions. Loh and Tan (2018) studied cellwise contamination for sparse precision matrix estimation from a statistical consistency point of view. They provided high-dimensional error bounds for the precision matrix estimators that reveal the interplay between the dimensionality of the problem and the degree of contamination permitted in the observed distribution.

2.3 Literature Review for Joint Estimation of Regression Coefficients and Precision Matrix Estimation

In the previous section we discussed the literature for estimating sparse linear regression with a single response variable in a high-dimensional setting when the covariates were corrupted (Datta and Zou, 2017; Loh and Wainwright, 2012). Imagine that instead of having one response variable, now we have multiple responses. Let us first consider the case when we have fully observed data. Specifically, suppose that we have n independent and identically distributed observations from some joint distribution of Y and X. In matrix notation, we can rewrite (4.1) as a model of n stacked observations

$$\mathbf{Y} = \mathbf{X}\mathbf{B}^* + \boldsymbol{\varepsilon},$$

where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top \in \mathbb{R}^{n\times p}$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^\top \in \mathbb{R}^{n\times q}$ denote the data matrices, in which each row $\mathbf{x}_i \in \mathbb{R}^p$ and $\mathbf{y}_i \in \mathbb{R}^q$ corresponds to an observation drawn i.i.d. from a distribution for $\mathbf{X}$ and $\mathbf{Y}$ respectively, and $\boldsymbol{\varepsilon} = [\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n]^\top \in \mathbb{R}^{n\times q}$ denotes the matrix of random noises. We aim to estimate the true coefficient matrix $\mathbf{B}^* \in \mathbb{R}^{p\times q}$ and $\boldsymbol{\Sigma}^*_{\varepsilon\varepsilon}$ where $\boldsymbol{\Sigma}^*_{\varepsilon\varepsilon} \in \mathbb{R}^{q\times q}$ represents the covariance structure of Y conditional on X.

This problem can be formulated as the sparse multivariate regression with correlated errors and can be solved by ℓ_1-penalization methods. The model has both high-dimensional regression coefficient matrix and high-dimensional covariance matrix. In the literature such models have been studied to estimate multivariate regression with correlated errors. Rothman et al. (2010) has proposed to jointly estimate $\mathbf{B}^*$ and $\boldsymbol{\Theta}^*_{\varepsilon\varepsilon} := (\boldsymbol{\Sigma}^*_{\varepsilon\varepsilon})^{-1}$ by minimizing the negative log-likelihood with ℓ_1 penalization as follows:

$$(\widehat{\boldsymbol{\Theta}}, \widehat{\mathbf{B}}) = \underset{\boldsymbol{\Theta}\succeq 0, \mathbf{B}}{\arg\min}\, \mathrm{tr}\left[\frac{1}{2n}(\mathbf{Y} - \mathbf{X}\mathbf{B})^\top(\mathbf{Y} - \mathbf{X}\mathbf{B})\boldsymbol{\Theta}\right] - \frac{1}{2}\log\det(\boldsymbol{\Theta}) + \lambda_{\boldsymbol{\Theta}}|||\boldsymbol{\Theta}|||_{1,\mathrm{off}} + \lambda_{\mathbf{B}}|||\mathbf{B}|||_{1,1}$$

where $|||\boldsymbol{\Theta}|||_{1,\mathrm{off}} = \sum_{j'\neq j}|\Theta_{jj'}|$, $|||\mathbf{B}|||_{1,1} = \sum_{j,k}|B_{jk}|$, and $\lambda_{\boldsymbol{\Theta}}, \lambda_{\mathbf{B}} \geq 0$ are tuning parameters controlling the sparsity in $\widehat{\boldsymbol{\Theta}}$ and $\widehat{\mathbf{B}}$, respectively. They called this method multivariate regression with covariance estimation (MRCE). Their approach involved a penalized likelihood and they proposed an efficient algorithm and a fast approximation by simultaneously estimating the regression coefficients and the covariance structure. Their work was computational in nature and no theoretical results were provided. They showed that the optimization problem is only convex if $\mathbf{B}^*$ is estimated with a fixed $\boldsymbol{\Theta}_{\varepsilon\varepsilon}$ and vise versa. When the components of the error vector $\boldsymbol{\varepsilon}$ are strongly correlated they showed that employing a one-step method to estimate $\mathbf{B}^*$ is improved by incorporating an estimate of $\boldsymbol{\Theta}_{\varepsilon\varepsilon}$.

Yin and Li (2011) developed a coordinate descent algorithm that iteratively updates the regression coefficients and the precision matrix based on ℓ_1-penalization. They provided asymptotic results on estimation bounds and consistency. In another study (Yin and

Li, 2013), they proposed a two-stage estimation procedure to first identify the relevant covariates that affect the means by a joint ℓ_1 penalization. Then they use the estimated regression coefficients to estimate the mean values in a multivariate sub-Gaussian model in order to estimate the sparse precision matrix through a ℓ_1-penalized log-determinant Bregman divergence. They also established convergence rates in elementwise maximum norm, Frobenius norm and spectral norm when both p and q are larger than n.

Cai et al. (2013) took a similar approach to estimate covariate adjusted precision matrix, but without making a multivariate normal assumption on the error distribution. They provided the rates of convergence and the estimation bounds for the estimates of both the regression coefficient matrix and the precision matrix in various matrix norms, allowing both p and q to diverge with n.

Another work by Wang (2015), proposed a method that decomposes the multivariate regression problem into a series of penalized conditional log-likelihood of each response conditional on the covariates and other responses. The use the adaptive Lasso penalty (Zou, 2006) to facilitate the sparse estimation of both the sparse multivariate regression coefficient matrix and the precision matrix. They showed that the proposed estimators possess asymptotic consistency and normality in diverging dimensions.

For the multivariate regression case, in a situation when we observe only strictly increasing transformations of the continuous responses and covariates, Zhao and Genest (2019) proposed an estimation of the joint dependence between all the observed variables characterized by an elliptical copula and used non-parametric estimators of the input matrix for the covariates. The coefficient matrix was assumed to be either elementwise sparse or row-sparse along with sparsity assumption on the precision matrix. Their method follows a one-step procedure similar to Rothman et al. (2010) in three stages and also considers cases when the estimated covariance structure for the covariates are not positive semi-definite, thereby the objective function at that step to be non-convex. They used the projection method proposed by Datta and Zou (2017) to convert the optimization

problem to be convex for further analysis. They also established the theoretical properties of their estimators.

Our theoretical analysis adopts the framework by Zhao and Genest (2019) assuming that the responses, covariates and the errors are sub-Gaussian. We assume that the responses are partially observed and contains missing data, but the covariates are fully observed. In Chapter 4 we show that due to the presence of missing data, the input matrix may not be positive semi-definite and the objective function might become unbounded from below. We assume that the true regression coefficients are elementwise sparse and propose a three stage estimation of the regression coefficients and the precision matrix. Since the estimated error covariance does not remain positive semi-definite, we first replace the empirical sample covariance for the error with an unbiased surrogate estimate for missing data and then project the estimator onto the nearest positive semi-definite cone (Datta and Zou, 2017). As a result, the overall objective function becomes convex. We also establish the theoretical guarantees in terms of elementwise maximum norm. Our theoretical results will have some resemblance to their work which will be shown in Chapter 4.

2.4 Summary

In this chapter, we reviewed the existing classical methods in the literature for estimating a sparse precision matrix in a high-dimensional setup. We focused mainly on the works that are based on penalized likelihood framework. We discussed various tuning parameter selection procedures along with the computational algorithm for the classical methods. Next, we studied the literature that are developed to estimate the sparse precision matrix in the presence of measurement error that includes presence of additive error and missing data. Finally, we discussed literature that studied how to jointly estimate the sparse

precision matrix and the regression coefficients from a multivariate regression setup in the presence of missing data.

Chapter 3

Precision Matrix Estimation in the Presence of Corrupted Data

In this chapter we develop estimators of the precision matrix when the observed data are corrupted by measurement error. We provide the key theoretical results concerning consistency and give rates of convergence. We also provide simulation studies and a real data example to demonstrate the proposed method.

3.1 Introduction

Given n independent observations of a p-dimensional random vector $X := (X_1, \ldots, X_p)^\top$, we wish to estimate the conditional independence relationships between $X_1, \ldots, X_p$, which can be characterized by an undirected graph $G := (V, E)$, where the vertex set $V := \{1, \ldots, p\}$ corresponds to the p variables in X, and the edge set E describes the conditional independence between any pair X_j and X_k in X ($j, k \in V$). Denote by $X_j \perp\!\!\!\perp X_k | X_{\setminus\{j,k\}}$ that X_j is conditionally independent of X_k given $X_{\setminus\{j,k\}}$, where $X_{\setminus\{j,k\}} := \{X_i : i \neq j, k\}$. We say that X is Markov with respect to G if

$$X_j \perp\!\!\!\perp X_k | X_{\setminus\{j,k\}} \quad \text{for all } (j,k) \notin E. \tag{3.1}$$

Denote the inverse covariance or precision matrix $\mathbf{\Theta}^* := (\mathbf{\Sigma}^*)^{-1}$. Under the Gaussian assumption of X, it is a well-known result that the zero pattern of $\mathbf{\Theta}^*$ corresponds the edge structure E of the underlying graph, i.e. (3.1) holds if and only if $\mathbf{\Theta}^*_{jk} = 0$. This estimator is sensible even for non-Gaussian X since it corresponds to minimizing an ℓ_1-penalized log-determinant Bregman divergence which does not necessarily require X to be multivariate Gaussian (Ravikumar et al., 2011).

Suppose that we are given an $n \times p$ data matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$, in which each row $\mathbf{x}_i \in \mathbb{R}^p$ corresponds to an observation drawn i.i.d. from a distribution. When $n \gg p$, we can estimate $\mathbf{\Sigma}^*$ using the sample covariance estimator

$$\mathbf{S} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}, \tag{3.2}$$

When $n < p$, the sample covariance estimate will be singular. Therefore, it is common to consider a regularized approach such as the graphical Lasso estimator

$$\widehat{\mathbf{\Theta}} = \underset{\Theta \succeq 0}{\arg\min}\, \mathrm{tr}(\mathbf{S}\mathbf{\Theta}) - \log\det(\mathbf{\Theta}) + \lambda_n \|\mathbf{\Theta}\|_{1,\mathrm{off}}. \tag{3.3}$$

When the data are contaminated by measurement errors, we observe a corrupted matrix $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^\top$ instead of the original matrix $\mathbf{X}$. We can view $\mathbf{Z}$ as some function of the true matrix $\mathbf{X}$ and the measurement error matrix. The measurement error process can be modeled in various ways. If the random error is additive, we observe $\mathbf{z}_i = \mathbf{x}_i + \mathbf{w}_i$, where $\mathbf{w}_i$ is a random error independent of $\mathbf{x}_i$. If the measurement error is multiplicative, we observe $\mathbf{z}_i = \mathbf{x}_i \odot \mathbf{w}_i$, where $\odot$ is the elementwise multiplication operator and $\mathbf{w}_i$ is the multiplicative error independent of $\mathbf{x}_i$. The missing data setup can be viewed as a special case of multiplicative errors, where w_{ij}, i.e. the jth component of $\mathbf{w}_i$, follows a Bernoulli$(1 - \pi_j)$ distribution and w_{ij}'s are independent for $j \in \{1, \ldots, p\}$.

If the measurement error issue is simply ignored, we would consider directly minimizing the objective function

$$\mathrm{tr}(\mathbf{S}_Z\mathbf{\Theta}) - \log\det(\mathbf{\Theta}) + \lambda_n\|\mathbf{\Theta}\|_{1,\text{off}}$$

where $\mathbf{S}_Z = (1/n)\mathbf{Z}^\top\mathbf{Z}$ is the sample covariance matrix of the corrupted data. Alternatively, Loh and Wainwright (2012) use $\mathbf{Z}$ to construct unbiased estimates $\widehat{\mathbf{\Sigma}}$ for $\mathbf{\Sigma}^*$. For example, consider the additive measurement errors $\mathbf{w}_i$'s are i.i.d. with mean zero and variance-covariance $\sigma_W^2\mathbf{I}_p$. Here σ_W^2 is known. Loh and Wainwright (2012) proposed the following unbiased estimate for $\mathbf{S}$ when the data are corrupted

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{Z}^\top\mathbf{Z} - \sigma_W^2\mathbf{I}_p = \mathbf{S}_Z - \sigma_W^2\mathbf{I}_p. \tag{3.4}$$

We can see that it is unbiased since

$$\mathbb{E}\left[\frac{1}{n}\mathbf{Z}^\top\mathbf{Z}\right] = \frac{1}{n}\mathbf{X}^\top\mathbf{X} + \sigma_W^2\mathbf{I}_p.$$

Note that if $\sigma_W^2\mathbf{I}_p = \mathbf{0}$, it reduces to the clean data case. It is natural to consider minimizing

$$\mathrm{tr}(\widehat{\mathbf{\Sigma}}\mathbf{\Theta}) - \log\det(\mathbf{\Theta}) + \lambda_n\|\mathbf{\Theta}\|_{1,\text{off}} \tag{3.5}$$

to get an estimate of $\mathbf{\Theta}$.

However, unlike $\mathbf{S} = (1/n)\mathbf{X}^\top\mathbf{X}$ which is always positive semi-definite, the estimate $\widehat{\mathbf{\Sigma}}$ is not necessarily positive semi-definite. This could lead to a non-convex objective function in (3.5). Moreover, when $\widehat{\mathbf{\Sigma}}$ has negative eigenvalues, the objective function (3.5) is unbounded from below. In order to overcome these issues, Loh and Wainwright (2012) proposed an additional constraint on the estimator $\|\beta\|_1 \leq b_0\sqrt{s}$ for the regression setting, where b_0 is some constant and the value of s is given. They used this method to estimate the

inverse covariance for Gaussian graphical models by considering the node-wise regression problem

$$X_j = X_{-j}\beta^{(j)} + \varepsilon^{(j)},$$

where X_j denotes the vector X with jth entry removed, $\varepsilon^{(j)}$ is a vector of i.i.d. Gaussian and $\varepsilon^{(j)} \perp\!\!\!\perp X_{-j}$. Then the inverse covariance can be estimated using the relationship $\mathbf{\Theta}_{j,-j} = -(\mathbf{\Sigma}_{jj} - \mathbf{\Sigma}_{j,-j}\beta^{(j)})^{-1}\beta^{(j)}$.

Fan et al. (2019) generalized this idea to the global likelihood formulation for the Gaussian graphical model. They consider minimizing the following objective function

$$\widehat{\mathbf{\Theta}} = \underset{\Theta \succeq 0, |||\mathbf{\Theta}|||_2 \leq R}{\arg\min} \ \mathrm{tr}(\widehat{\mathbf{\Sigma}}\mathbf{\Theta}) - \log\det(\mathbf{\Theta}) + \lambda_n \|\mathbf{\Theta}\|_{1,\mathrm{off}}, \tag{3.6}$$

subject to the additional constraint on the operator norm of the true precision matrix. They showed that if the value of R is properly chosen, an ADMM algorithm can converge to the global minimum of (3.6).

Compared to these methods, we propose to project $\widehat{\mathbf{\Sigma}}$ to its nearest positive semi-definite matrix, thereby guaranteeing the estimator to be convex. Hence we can avoid performing any non-convex analysis. Since there are no additional constraints to satisfy in the convex analysis, another advantage of our method is that it does not require the knowledge of an initial estimate of $\|\beta\|_1$ or $|||\mathbf{\Theta}|||_2$ to obtain a bound for $\|\beta\|_1$ or $|||\mathbf{\Theta}|||_2$, respectively.

Notation and Definitions. For a matrix $\mathbf{A}$, we denote by $\mathbf{A} \succeq 0$ when $\mathbf{A}$ is positive semi-definite. Let $|||\mathbf{A}|||_1$ be the operator norm induced by ℓ_1 norm for vectors, which can be computed by $|||\mathbf{A}|||_1 = \max_j \sum_i \mid a_{ij} \mid$, i.e. the maximum absolute column sum of the matrix. Denoted by $|||\mathbf{A}|||_2$, the operator norm that can be computed as the greatest singular value of $\mathbf{A}$, i.e. $|||\mathbf{A}|||_2 = \max_j \sigma_j(\mathbf{A})$. Let $|||\mathbf{A}|||_\infty$ be the operator norm induced by an ℓ_∞ norm, which can be computed by $|||\mathbf{A}|||_\infty = \max_i \sum_j \mid a_{ij} \mid$, i.e. the maximum absolute row sum of

the matrix. Let $|||\mathbf{A}|||_{1,1} = \sum_{i,j} \mid a_{ij} \mid$ be the elementwise ℓ_1-norm, $|||\mathbf{A}|||_F = \sqrt{\sum_{i,j} \mid a_{ij} \mid^2}$ be the Frobenius norm, and $|||\mathbf{A}|||_{\max} = \max_{i,j} \mid a_{ij} \mid$ be the elementwise maximum norm. Let $\Lambda_{\min}(\mathbf{A})$ and $\Lambda_{\max}(\mathbf{A})$ denote the smallest and largest eigenvalues of $\mathbf{A}$. For two matrices $\mathbf{A} = (a_{ij})$ and $\mathbf{B} = (b_{ij})$, we define $\mathbf{A} \odot \mathbf{B} = (a_{ij}b_{ij})$ as their elementwise product, and $\mathbf{A} \oslash \mathbf{B} = (a_{ij}/b_{ij})$ as their elementwise division. We assume that all variables are centered so that the intercept term is not included in the model and the design matrix X has normalized columns, that is, $(1/n)\sum_{i=1}^n x_{ij}^2 = 1$ for every $j = 1, \ldots, p$.

A *sub-Gaussian* random variable Z with the parameter $\tau > 0$ satisfies the tail probability bound $\Pr(|Z| \geq t) \leq 2\exp(-t^2/2\tau^2)$ for all $t \geq 0$; a $4m$*th moment-bounded* random variable Z with the parameter $K_m > 0$ satisfies the condition $\mathbb{E}(Z^{4m}) \leq K_m$ with $m \in \mathbb{Z}^+$.

3.2 Methodology

From the earlier discussion, we know that the estimate $\widehat{\mathbf{\Sigma}}$ is often not positive semi-definite. To overcome this technical difficulty, by following (Datta and Zou, 2017), we can project $\widehat{\mathbf{\Sigma}}$ to its nearest positive semi-definite matrix. Specifically,

$$\widetilde{\mathbf{\Sigma}} = \underset{\mathbf{\Sigma} \succeq 0}{\arg\min} |||\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}|||_{\max}. \tag{3.7}$$

Then we define the ***C**onvex c**o**nditioned **G**raphical **Lasso** (CoGlasso) estimate as*

$$\widehat{\mathbf{\Theta}} = \underset{\Theta \succeq 0}{\arg\min} \operatorname{tr}(\widetilde{\mathbf{\Sigma}}\mathbf{\Theta}) - \log\det(\mathbf{\Theta}) + \lambda_n |||\mathbf{\Theta}|||_{1,\text{off}}. \tag{3.8}$$

We aim to derive the elementwise ℓ_∞-norm for the statistical error of the CoGlasso estimate from the truth. To do so, we need to bound the statistical error between the projected covariance matrix $\widetilde{\mathbf{\Sigma}}$ for the unclean data defined in (3.7) and the true covariance

matrix $\mathbf{\Sigma}^*$. We can decompose the error as follows:

$$|||\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*|||_{\max} = |||\widetilde{\mathbf{\Sigma}} - \mathbf{S} + \mathbf{S} - \mathbf{\Sigma}^*|||_{\max},$$

where $\mathbf{S}$ is the estimated sample covariance matrix if there is no corruption in the data. Using triangular inequality, we can write

$$|||\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*|||_{\max} \leq |||\widetilde{\mathbf{\Sigma}} - \mathbf{S}|||_{\max} + |||\mathbf{S} - \mathbf{\Sigma}^*|||_{\max}. \tag{3.9}$$

In the following section, we define conditions that are required to characterize the closeness between two random matrices in terms of the elementwise maximum norm that would essentially provide us the probabilistic bounds for the terms in (3.9).

3.3 Tail Conditions

We define the following *closeness condition* (Ravikumar et al., 2011) between two random matrices.

Definition 1. [Closeness condition] Two (random) matrices $\mathbf{\Sigma}^{(1)}$ and $\mathbf{\Sigma}^{(2)}$ satisfy the closeness condition if there exists a constant $v_* \in (0, \infty]$ and a function $f : \mathbb{N} \times (0, \infty) \to (0, \infty)$ such that for any $j, k = 1, \ldots, p$ the following probability bound holds:

$$\Pr(|||\mathbf{\Sigma}^{(1)} - \mathbf{\Sigma}^{(2)}|||_{\max} \geq \delta) \leq \frac{p^2}{f(n, \delta)} \qquad \text{for all } \delta \in \Big(0, \frac{1}{v_*}\Big]. \tag{3.10}$$

When $v_* = 0$ the inequality holds for any $\delta \in (0, \infty)$. We will consider two types of tail functions, namely,

a) Exponential-tail function: when $f(n, \delta) = C \exp(cn\delta^a)$, for some positive constant C and c and exponent $a > 0$.

b) Polynomial-tail function: when $f(n, \delta) = c_* n^m \delta^{2m}$, for some positive integer $m \in \mathbb{N}$ and scalar $c_* > 0$.

Example 1. By Lemma 1 of Ravikumar et al. (2011): When zero mean and normalized random vector $(X_1, \ldots, X_p)$ is sub-Gaussian with parameter σ_X, for n i.i.d. samples, the associated sample covariance $\mathbf{S}$ obtained from the clean data and the true covariance $\mathbf{\Sigma}^*$ satisfies the closeness condition with exponential-tail function $f(n, \delta) = (1/4)\exp(cn\delta^2)$ with $C = 1/4$ and $c = [64(1 + 4\sigma_X^2)^2]^{-1}$.

Example 2. By Lemma 2 of Ravikumar et al. (2011): When zero mean and normalized random vector $(X_1, \ldots, X_p)$ has $4m$th bounded moments, for n i.i.d. samples, the sample covariance $\mathbf{S}$ and the true covariance $\mathbf{\Sigma}^*$ satisfies the closeness condition with polynomial-tail function $f(n, \delta) = c_* n^m \delta^{2m}$, with $c_* = [2^{2m} C_m (K_m + 2)]^{-1}$ and C_m as a constant depending only on m.

As n increases, we can expect that the elementwise tail probability bound $1/f(n, \delta)$ would decrease, or equivalently the tail function $f(n, \delta)$ would increase. Therefore, f is required to monotonically increase in n, so that for each fixed $\delta > 0$, the inverse function can be defined as

$$\bar{n}_f(\delta, r) := \arg\max \{n \mid f(n, \delta) \leq r\}, \tag{3.11}$$

which is the largest n such that $f(n, \delta) \leq r$, where $r \in [1, \infty)$. Similarly, we expect that f is monotonically increasing in δ, so that for each fixed n, the inverse function in the second argument can be defined as

$$\bar{\delta}_f(n, r) := \arg\max \{\delta \mid f(n, \delta) \leq r\}, \tag{3.12}$$

which is the largest δ such that $f(n, \delta) \leq r$. Now, if we can find a n such that $n > \bar{n}_f(\delta, r)$, that would imply that $f(n, \delta) > r$, for some $\delta > 0$. Consequently, since f is a monotone

function, this would imply that $\bar{\delta}_f(n, r) \leq \delta$. Therefore, we can write

$$n > \bar{n}_f(\delta, r) \qquad \text{for some } \delta > 0 \qquad \Longrightarrow \qquad \bar{\delta}_f(n, r) \leq \delta. \tag{3.13}$$

The inverse functions $\bar{n}_f$ and $\bar{\delta}_f$ are important because they help to describe the behaviour of the estimators.

In case of Example 1, if $\mathbf{X}$ is multivariate Gaussian, then the deviation of the sample covariance matrix has an exponential-type tail function with $a = 2$. Therefore, it can be shown by some calculations that the associated inverse functions take the following forms:

$$\bar{\delta}_f(n, r) = \sqrt{\frac{\log(4r)}{cn}}, \qquad \text{and} \qquad \bar{n}_f(\delta, r) = \frac{\log(4r)}{c\delta^2}.$$

In Example 2, for the polynomial-type tail function it can be shown that the inverse tail functions take the forms

$$\bar{\delta}_f(n, r) = \frac{(r/c_*)^{1/2m}}{\sqrt{n}}, \qquad \text{and} \qquad \bar{n}_f(\delta, r) = \frac{(r/c_*)^{1/m}}{\delta^2}.$$

By applying the closeness condition to $\mathbf{S}$ and $\mathbf{\Sigma}^*$, it can be shown that we can bound the second term in (3.9) as demonstrated in Example 1 and 2 for exponential and polynomial-type tails. The details of the proofs can be found in Lemma 1 and 2 in Ravikumar et al. (2011). Therefore, we only need to provide probabilistic bounds for the first term in (3.9). Deriving the bounds for these components separately will lead us to deriving a bound for $\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*$.

In the following sections we start with deriving a bound for the first quantity $|||\widehat{\mathbf{\Sigma}} - \mathbf{S}|||_{\max}$ in terms of a closeness condition for cases when the underlying random variables are (i) sub-Gaussian, and (ii) have bounded moments. Then we provide a probability bound to ensure that $\widetilde{\mathbf{\Sigma}}$ approximates $\widehat{\mathbf{\Sigma}}$ and put all the components together in Theorem 1.

3.4 Two Types of Measurement Errors

Given n independent observations of a p-dimensional random vector $X := (X_1, \ldots, X_p)^\top$, we want to estimate the conditional dependence relationships between $X_1, \ldots, X_p$. With the presence of contamination, we do not directly observe X, we instead use a surrogate estimator $\widehat{\mathbf{\Sigma}}$ to estimate the covariance matrix based on the contaminated data $\mathbf{Z}$.

3.4.1 Surrogate Estimators

Additive error. Following Loh and Wainwright (2012) and Xu and You (2007), we assume the observed matrix can be written $\mathbf{Z} = \mathbf{X} + \mathbf{W}$, where the rows of $\mathbf{X}$ are i.i.d. with zero mean, finite covariance $\mathbf{\Sigma}^*$. Here $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_n)^\top$ is a matrix of additive measurement errors, with rows being i.i.d. with zero mean, finite known covariance $\mathbf{\Sigma}_W$. Also assume that any row of $\mathbf{X}$ is independent to any row of $\mathbf{W}$. One can find an unbiased estimator of $\mathbf{\Sigma}^*$ as

$$\widehat{\mathbf{\Sigma}}_{\text{addi}} = \frac{1}{n}\mathbf{Z}^\top\mathbf{Z} - \mathbf{\Sigma}_W.$$

In the following two lemmas, we show that the surrogate estimators $\widehat{\mathbf{\Sigma}}_{\text{addi}}$ is sufficiently "close" to the covariance estimator $\mathbf{S}$. We consider two different distribution assumptions for $\mathbf{X}$ and $\mathbf{W}$: sub-Gaussian and moment-bounded. The proofs of the lemmas are provided in Section 3.9.

Lemma 1. *[Additive, sub-Gaussian errors.] Suppose that the rows of* $\mathbf{X}$ *and* $\mathbf{W}$ *are i.i.d. sub-Gaussian with parameter* σ_X^2 *and* σ_W^2*, respectively. The associated surrogate estimator* $\widehat{\mathbf{\Sigma}}_{\text{addi}}$ *and covariance for clean data* $\mathbf{S}$ *satisfy the closeness condition* (3.10) *with the function* $f(n, \delta) = C\exp(cn\delta^2\xi^{-1})$*, where* C *and* c *are universal constants and* $\xi = \max(\sigma_W^4, \sigma_W^2, \sigma_X^2\sigma_W^2, \sigma_X\sigma_W)$ *and* δ_0 *are positive functions depending on* σ_X^2 *and* σ_W^2 *such that for every* $\delta \leq \delta_0$ *the bound holds. When* δ *is sufficiently small, specifically for* $\delta \leq \min(\sigma_X\sigma_W, \sigma_W^2)$*,* $\widehat{\mathbf{\Sigma}}_{\text{addi}}$ *and* $\mathbf{S}$ *satisfy the closeness condition with* $\xi = \max(\sigma_W^4, \sigma_X^2\sigma_W^2)$*.*

Lemma 2. *[Additive, moment-bounded errors.] Consider that the rows of* $\mathbf{X}$ *and* $\mathbf{W}$ *are i.i.d.* $4m$*th moment-bounded with parameter* $K_{m,X}$ *and* $K_{m,W}$*, respectively.* $\widehat{\mathbf{\Sigma}}_{\text{addi}}$ *and* $\mathbf{S}$ *satisfy the closeness condition with the function* $f(n,\delta) = c_* n^m \delta^{2m}$*, where* c_* *is a universal constants*

$$c_* = \left[C_m 2^{4m}\left\{6^m + 2(1+K_{m,X})(1+K_{m,W}) + K_{m,W} + (1+2^m)\right\}\right]^{-1},$$

and the bound holds true for every $\delta \geq 0$*.*

Multiplicative error and missing data. We can also consider the case when the errors are multiplicative. We assume the observed matrix is $\mathbf{Z} = \mathbf{X} \odot \mathbf{W}$ where $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_n)^\top$ is a matrix of multiplicative errors where each row $\mathbf{w}_i \in \mathbb{R}^p$ of $\mathbf{W}$ is independent and identically distributed with known mean $\mathbb{E}(W) = \mu_W \in \mathbb{R}^p$ and covariance $\mathbf{\Sigma}_W$. In addition, μ_W and $\mathbf{\Sigma}_W$ are assumed to have positive entries. Under these assumptions, Loh and Wainwright (2012) proposed to use the unbiased estimators

$$\widehat{\mathbf{\Sigma}}_{\text{mult}} = \frac{1}{n}\mathbf{Z}^\top\mathbf{Z} \oslash \mathbb{E}(WW^\top) = \frac{1}{n}\mathbf{Z}^\top\mathbf{Z} \oslash \left(\mathbf{\Sigma}_W + \mu_W\mu_W^\top\right). \tag{3.14}$$

Data that are missing at random can be viewed as a special case of the multiplicative error model. Assume that x_{ij}, the jth component of $\mathbf{x}_i$, is missing at random with probability π_j. In other words, for each observation $\mathbf{z}_i$, we independently observe the jth component $z_{ij} = x_{ij}$ with probability $1 - \pi_j$, and $z_{ij} = 0$ with probability π_j. This can be modeled by introducing Bernoulli random variables $w_{ij} = \mathbb{I}(x_{ij} \text{ is not missing}) \sim \text{Bernoulli}(1-\pi_j)$ as the (i,j)th entry of $\mathbf{W}$ in the previous multiplicative error model. If the value of π_j is unknown, it can be estimated by using the proportion of missing entries in jth column of $\mathbf{X}$. Followed by (3.14), the estimator becomes

$$\widehat{\mathbf{\Sigma}}_{\text{miss}} = \frac{1}{n}\mathbf{Z}^\top\mathbf{Z} \oslash \mathbb{E}(WW^\top),$$

with

$$\mathbb{E}[WW^\top]_{ij} = \begin{cases} (1-\pi_i)(1-\pi_j), & i \neq j, \\ (1-\pi_i), & i = j. \end{cases}$$

Lemma 3. *[Multiplicative, sub-Gaussian errors.]* $\widehat{\mathbf{\Sigma}}_{\text{mult}}$ *and* $\mathbf{S}$ *satisfy the closeness condition with the function* $f(n,\delta) = C\exp(cn\delta^2\xi^{-1})$ *and* $\xi = \max(\sigma_W^4\sigma_X^4, \sigma_W\sigma_X, \sigma_X^4, \sigma_X^2)$ *under the sub-Gaussian assumptions for the rows of* $\mathbf{X}$ *and* $\mathbf{W}$*, respectively. When* δ *is sufficiently small, specifically for* $\delta \leq \min(\sigma_X^2\sigma_W^2, \sigma_X^2)$*,* $\widehat{\mathbf{\Sigma}}_{\text{mult}}$ *and* $\mathbf{S}$ *satisfy the closeness condition with* $\xi = \max(\sigma_X^4\sigma_W^4, \sigma_X^4)$*.*

Lemma 4. *[Multiplicative, moment-bounded errors.] Consider that the rows of* $\mathbf{X}$ *and* $\mathbf{W}$ *are i.i.d.* $4m$*th moment-bounded with parameter* $K_{m,X}$ *and* $K_{m,W}$*, respectively.* $\widehat{\mathbf{\Sigma}}_{\text{mult}}$ *and* $\mathbf{S}$ *satisfy the closeness condition with the function* $f(n,\delta) = c_* n^m \delta^{2m}$ *and the parameter* c_* *where*

$$c_* = \left[C_m 2^{2m} \left\{ \frac{1}{m_{\min}^{2m}} \left(2 + K_{m,X}K_{m,W} \right) + K_{m,X} \right\} \right]^{-1}.$$

3.5 Consistency Bounds

3.5.1 Rates in Elementwise ℓ_∞-norm

Next, we derive the consistency bounds on the deviation of $\widehat{\mathbf{\Theta}}$ from the true precision matrix $\mathbf{\Theta}^*$. To do so, we require bounds on the deviation between the projected sample covariance $\widetilde{\mathbf{\Sigma}}$ and the true covariance $\mathbf{\Sigma}^*$.

The proofs of consistency of $\widehat{\mathbf{\Theta}}$ and its error in elementwise ℓ_∞-norm is extended from the proofs of Graphical Lasso (Ravikumar et al., 2011) to incorporate measurement error and missing data cases. The following results depend on the quantities defined in Ravikumar et al. (2011),

$$\kappa_{\mathbf{\Sigma}^*} = |||\mathbf{\Sigma}^*|||_\infty = \left(\max_{i=1,\ldots,p} \sum_{j=1}^{p} \mathbf{\Sigma}_{ij}^* \right) \tag{3.15}$$

corresponding to the ℓ_∞-operator norm of the true covariance matrix $\mathbf{\Sigma}^*$, the inverse of the sub-block of the Hessian $\mathbf{\Gamma}^*$, defined as

$$\mathbf{\Gamma}^*_{SS} = (\mathbf{\Theta}^{*-1} \otimes \mathbf{\Theta}^{*-1})_{SS} \in \mathbb{R}^{(s+p)\times(s+p)} \tag{3.16}$$

where s denote the number of edges and p denote the number of nodes in the graph and the parameter

$$\kappa_{\mathbf{\Gamma}^*} = |||(\mathbf{\Gamma}^*_{SS})^{-1}|||_\infty. \tag{3.17}$$

Here, S is an augmented set that includes all the off-diagonal entries of the true precision matrix and the diagonal elements, that is, the true edges as well as the diagonal elements. Therefore, the cardinality of S, $|S| = s + p$. We denote the complement of S as S^c corresponding to all the pairs for which the true precision matrix have zero entries. We assume that the Hessian satisfies the following type of mutual coherence or irrepresentability condition:

(C1) Mutual Incoherence Condition. There exists some $\alpha \in (0, 1]$ such that

$$\max_{e \in S^c} |||\mathbf{\Gamma}^*_{eS}(\mathbf{\Gamma}^*_{SS})^{-1}|||_{1,1} \leq 1 - \alpha \tag{3.18}$$

The intuition behind this condition is that it controls the influence that the non-edge terms, indexed by S^c can have on the terms representing edges that are indexed by S. In other words, this assumption enforces the requirement that there should be no edge variable that is not included in the graph that is highly correlated with variables within the true edge-set.

The proofs of estimation consistency and elementwise ℓ_∞-norm are based on the technique called the primal-dual witness method which was used previously for the analysis of Lasso (Wainwright, 2009) and graphical Lasso (Ravikumar et al., 2011). The method requires to construct a pair $(\widetilde{\mathbf{\Theta}}, \widetilde{\mathbf{Z}})$ of symmetric matrices, where $\widetilde{\mathbf{\Theta}} \succ 0$ as a primal optimal solution and $\widetilde{\mathbf{Z}}$ as the corresponding dual optimum. This pair satisfies the optimality

conditions associated with the convex problem (3.8) with high probability. When the primal-dual witness method succeeds, the estimator $\widehat{\mathbf{\Theta}}$ inherits various optimality properties in terms of its distance to the truth $\mathbf{\Theta}^*$ from $\widetilde{\mathbf{\Theta}}$. The matrix $\widetilde{\mathbf{Z}}$ must belong to the sub-differential of the norm $|||\cdot|||_{1,\text{off}}$, evaluated at $\widetilde{\mathbf{\Theta}}$, such that

$$\widetilde{\mathbf{Z}}_{ij} = \begin{cases} 0, & i = j, \\ \text{sgn}(\widetilde{\mathbf{\Theta}}_{ij}), & i \neq j \text{ and } \widetilde{\mathbf{\Theta}}_{ij} \neq 0 \\ \in [-1, +1], & i \neq j \text{ and } \widetilde{\mathbf{\Theta}}_{ij} = 0. \end{cases}$$

The primal-dual witness condition requires that the uniqueness of the solution of the ℓ_1-regularized log-determinant problem (stated in Lemma 8) where the projected sample covariance $\widetilde{\mathbf{\Sigma}}$ is used as an input to take care of the measurement errors. We also need to verify the strict dual feasibility condition in step (d) for the primal-dual witness condition to hold and it is stated in Lemma 9. Some additional useful notations to prove Lemma 9 which will also be used in Theorem 1 are defined as follows.

Let $\mathbf{W} \in \mathbb{R}^{p \times p}$ be the effective noise in the projected sample covariance matrix $\widetilde{\mathbf{\Sigma}}$,

$$\mathbf{W} = \widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^* = \widetilde{\mathbf{\Sigma}} - (\mathbf{\Theta}^*)^{-1}. \tag{3.19}$$

Next, we use $\mathbf{\Delta} = \widetilde{\mathbf{\Theta}} - \mathbf{\Theta}^*$ to measure the discrepancy between the primal witness $\widetilde{\mathbf{\Theta}}$ and the truth $\mathbf{\Theta}^*$. Note that, by definition of $\widetilde{\mathbf{\Theta}}$, $\mathbf{\Delta}_{S^c} = 0$.

Finally, let $R(\mathbf{\Delta})$ denote the difference of the gradient of the log-determinant function, i.e., $\nabla\{-\log\det\widetilde{\mathbf{\Theta}}\} = \widetilde{\mathbf{\Theta}}^{-1}$, from its first-order Taylor expansion around $\mathbf{\Theta}^*$. Therefore, the remainder term takes the form

$$R(\mathbf{\Delta}) = \widetilde{\mathbf{\Theta}}^{-1} - (\mathbf{\Theta}^*)^{-1} + (\mathbf{\Theta}^*)^{-1}\mathbf{\Delta}(\mathbf{\Theta}^*)^{-1} \tag{3.20}$$

The Lemmas 8 and 9 are stated and proved in in Section 3.9. With the Lemmas and required terminologies in place, the primal-dual witness condition can be defined as follows:

Definition 2. [Primal-dual Witness Condition.] Based on Lemma 8, we can construct the primal-dual witness solution $(\widetilde{\mathbf{\Theta}}, \widetilde{\mathbf{Z}})$ as follows.

a) We determine the matrix $\widetilde{\mathbf{\Theta}}$ by solving the restricted log-determinant problem

$$\widetilde{\mathbf{\Theta}} = \underset{\mathbf{\Theta} \succ 0, \mathbf{\Theta} = \mathbf{\Theta}^{\top}, \mathbf{\Theta}_{S^c} = 0}{\arg\min} \left\{ \langle\!\langle \mathbf{\Theta}, \widetilde{\mathbf{\Sigma}} \rangle\!\rangle - \log\det(\mathbf{\Theta}) + \lambda_n |\!|\!| \mathbf{\Theta} |\!|\!|_{1,\text{off}} \right\}.$$

By construction, we have $\widetilde{\mathbf{\Theta}} \succ 0$ and $\widetilde{\mathbf{\Theta}}_{S^c} = 0$.

b) We choose $\widetilde{\mathbf{Z}}$ as a member of the sub-differential of the regularizer $|\!|\!| \cdot |\!|\!|_{1,\text{off}}$, evaluated at $\widetilde{\mathbf{\Theta}}$.

c) For each $(i, j) \in S^c$, we replace $\widetilde{\mathbf{Z}}_{ij}$ with the quantity

$$\widetilde{\mathbf{Z}}_{ij} = \frac{1}{\lambda_n} \left\{ -\widetilde{\mathbf{\Sigma}}_{ij} + \widetilde{\mathbf{\Theta}}_{ij}^{-1} \right\},$$

ensuring that the constructed matrices $(\widetilde{\mathbf{\Theta}}, \widetilde{\mathbf{Z}})$ satisfy the optimality conditions (3.48).

d) We verify the *strict dual feasibility condition*

$$|\widetilde{\mathbf{Z}}_{ij}| < 1 \qquad \text{for all } (i, j) \in S^c.$$

Steps (a)-(c) are necessary conditions to obtain a pair $(\widetilde{\mathbf{\Theta}}, \widetilde{\mathbf{Z}})$ that satisfy the optimality condition, but they do not guarantee that $\widetilde{\mathbf{Z}}$ is an element of the sub-differential $\partial |\!|\!| \widehat{\mathbf{\Theta}} |\!|\!|_{1,\text{off}}$. By construction, $\widetilde{\mathbf{Z}}$ in S satisfy the sub-differential conditions, since $\widetilde{\mathbf{Z}}_S$ is a member of the sub-differential $[\partial |\!|\!| \widehat{\mathbf{\Theta}} |\!|\!|_{1,\text{off}}]_S$. In addition, the strict feasibility condition is a necessary condition to ensures two things; the first being that $\widetilde{\mathbf{Z}}_{S^c}$ is indeed within the sub-differential

and the second being that no false inclusion condition holds for $\widetilde{\mathbf{\Theta}}$. In a linear regression setup, the strict dual feasibility condition is also used to ensure the uniqueness of the solution. However, in the graphical model setup this condition is not required to ensure uniqueness as we have noticed in Lemma 8. But, we still require this condition $|\widetilde{\mathbf{Z}}_{S^c}| < 1$ to be satisfied because otherwise $\widetilde{\mathbf{\Theta}}_{S^c}$ might still be non-zero if $|\widetilde{\mathbf{Z}}_{S^c}| = 1$. This is because when $\widehat{\mathbf{\Theta}}_{S^c} = 0$, it implies no false inclusion, that is, the graph includes no false edges, which is equivalent to saying that the solution has its support set $\widehat{E}$ contained within the true support set E. On the other hand, if $\widehat{\mathbf{\Theta}}_{S^c} \neq 0$ would imply that no false inclusion condition is violated, that is, the support set of the solution $\widehat{E}$ would not be contained within the true support set E.

With the required terms and conditions defined above, next we state the main theorem to bound the deviation of $\widehat{\mathbf{\Theta}}$ from the true precision matrix $\mathbf{\Theta}^*$ in terms of elementwise norm. In addition, Theorem 1 depends on Lemmas 10 and 11 which are stated and proved in Section 3.9. In the statement of the following theorem, the choice of the regularization parameter λ_n is specified in terms of a user-defined parameter $\gamma > 2$. With growing γ, the rate of convergence in probability gets faster, but consequently a restriction is imposed on the sample size. The rates in Theorem 1 differ from the classical graphical Lasso results presented in Ravikumar et al. (2011) in terms of the quantities $\bar{\delta}_{f_*}(n, p^\gamma)$ and $\bar{n}_{f_*}(\delta, p^\gamma)$, which varies in our setup depending on the distributional assumptions on X and W and the type of measurement errors under consideration. The specific forms of $\bar{\delta}_{f_*}(n, p^\gamma)$ and $\bar{n}_{f_*}(\delta, p^\gamma)$ for each scenario are given in Remarks.

Theorem 1. *Consider a distribution satisfying the incoherence condition* (3.18) *with parameter* $\alpha \in (0, 1]$*, and (random) estimators obtained from that distribution satisfying the closeness conditions* (1)*. Let* $\widehat{\mathbf{\Theta}}$ *be the unique solution of the CoGlasso program with regularization parameter* $\lambda_n = (8/\alpha)\bar{\delta}_{f_*}(n, p^\gamma)$ *for some* $\gamma > 2$*. Then if the sample size is lower bounded as*

$$n > \bar{n}_f \left(\frac{1}{\max\{v_*, 6(1 + 8/\alpha)^2 d \max\{\kappa_{\mathbf{\Sigma}^*}\kappa_{\mathbf{\Gamma}^*}, \kappa^3_{\mathbf{\Sigma}^*}\kappa^2_{\mathbf{\Gamma}^*}\}\}}, p^\gamma \right) \tag{3.21}$$

Then with probability at least $1-(1/p^{\gamma-2}) \to 1$*, we have the following:*

a) The estimate $\widehat{\mathbf{\Theta}}$ *satisfies the elementwise* ℓ_∞*-norm bound*

$$|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_{\max} \leq \left\{2\kappa_{\mathbf{\Gamma}^*}\Big(1+(8/\alpha)\Big)\right\}\bar{\delta}_{f_*}(n,p^\gamma). \tag{3.22}$$

b) It specifies an edge set $E(\widehat{\mathbf{\Theta}})$ *that is a subset of the true edge set* $E(\mathbf{\Theta}^*)$*, and includes all edges* (i,j) *with* $|\ \mathbf{\Theta}^*_{ij}\ | > \left\{2\kappa_{\mathbf{\Gamma}^*}\Big(1+8/\alpha\Big)\right\}\bar{\delta}_{f_*}(n,p^\gamma)$.

We can simplify the probability bounds for each particular case and give expressions for the tail function and their associated inverse functions for $\bar{\delta}_{f_*}(n,p^\gamma) \leq \delta$. Recall that, from the Definition 1, if the rows of $\mathbf{X}$ and $\mathbf{W}$ are multivariate Gaussian, then the deviation bound for the sample covariance matrix has an exponential-type tail function with $a=2$. Remark 1 and 3 are simplified for the particular case of exponential-type tail function for the deviation bounds when the rows of $\mathbf{X}$ and $\mathbf{W}$ are multivariate Gaussian.

Remark 1. The rows of $\mathbf{X}$ and $\mathbf{W}$ are both sub-Gaussian and the measurement error is additive: we have for a particular case when the rows of both $\mathbf{X}$ and $\mathbf{W}$ are multivariate Gaussian

$$\Pr(|||\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*|||_{\max} \geq \delta) \leq \frac{p^2(4+C)}{4C}\exp\left\{-cn\delta^2\min\left(\frac{1}{16\sigma_X^2\sigma_W^2}, \frac{1}{16\sigma_W^4}, \frac{1}{256c(1+4\sigma_X^2)^2}\right)\right\}$$

with the tail function defined as

$$f_*(n,\delta) = \frac{4C}{4+C}\exp\left\{cn\delta^2\min\left(\frac{1}{16\sigma_X^2\sigma_W^2}, \frac{1}{16\sigma_W^4}, \frac{1}{256c(1+4\sigma_X^2)^2}\right)\right\}$$

and associated inverse functions taking the forms

$$\bar{\delta}_{f_*}(n,p^\gamma) = \sqrt{\frac{\log\left(\dfrac{(4+C)p^\gamma}{4C}\right)}{cn\min\left(\dfrac{1}{16\sigma_X^2\sigma_W^2}, \dfrac{1}{16\sigma_W^4}, \dfrac{1}{256c(1+4\sigma_X^2)^2}\right)}},$$

and

$$\bar{n}_{f_*}(\delta, p^\gamma) = \frac{\log\left(\frac{(4+C)p^\gamma}{4C}\right)}{c\delta^2 \min\left(\frac{1}{16\sigma_X^2\sigma_W^2}, \frac{1}{16\sigma_W^4}, \frac{1}{256c(1+4\sigma_X^2)^2}\right)}.$$

Remark 2. The rows of $\mathbf{X}$ and $\mathbf{W}$ both have bounded moments and the measurement error is additive: we have

$$\Pr(|||\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*|||_{\max} \geq \delta) \leq \frac{p^2 2^{4m}}{n^m \delta^{2m}} \left(\frac{1}{c_*} + C_m(K_m + 2)\right)$$

with $c_* = [C_m 2^{4m} \{6^m + 2(1+K_{m,X})(1+K_{m,W}) + K_{m,W} + (1+2^m)\}]^{-1}$ and the tail function defined as

$$f_*(n, \delta) = \left\{\frac{2^{4m}}{n^m \delta^{2m}} \left(\frac{1}{c_*} + C_m(K_m + 2)\right)\right\}^{-1}$$

and associated inverse functions taking the forms

$$\bar{\delta}_{f_*}(n, p^\gamma) = \frac{\left(2^{4m}\left(\frac{1}{c_*} + C_m(K_m+2)\right)p^\gamma\right)^{1/2m}}{\sqrt{n}}$$

and

$$\bar{n}_{f_*}(\delta, p^\gamma) = \frac{\left(2^{4m}\left(\frac{1}{c_*} + C_m(K_m+2)\right)p^\gamma\right)^{1/m}}{\delta^2}.$$

Remark 3. The rows of both $\mathbf{X}$ and $\mathbf{W}$ are sub-Gaussian and the measurement error is multiplicative: we have for a particular case when the rows of both $\mathbf{X}$ and $\mathbf{W}$ are multivariate Gaussian

$$\Pr(|||\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*|||_{\max} \geq \delta) \leq \frac{p^2(4+C)}{4C} \exp\left\{-cn\delta^2 \min\left(\frac{1}{16\sigma_X^2\sigma_W^2}, \frac{1}{16\sigma_X^4}, \frac{1}{256c(1+4\sigma_X^2)^2}\right)\right\}$$

with tail function defined as

$$f_*(n, \delta) = \frac{4C}{4+C} \exp\left\{cn\delta^2 \min\left(\frac{1}{16\sigma_X^2\sigma_W^2}, \frac{1}{16\sigma_X^4}, \frac{1}{256c(1+4\sigma_X^2)^2}\right)\right\}$$

and associated inverse functions taking the forms

$$\bar{\delta}_{f_*}(n, p^\gamma) = \sqrt{\frac{\log\left(\frac{(4+C)p^\gamma}{4C}\right)}{cn\min\left(\frac{1}{16\sigma_X^2\sigma_W^2}, \frac{1}{16\sigma_X^4}, \frac{1}{256c(1+4\sigma_X^2)^2}\right)}}$$

and

$$\bar{n}_{f_*}(\delta, p^\gamma) = \frac{\log\left(\frac{(4+C)p^\gamma}{4C}\right)}{c\delta^2\min\left(\frac{1}{16\sigma_X^2\sigma_W^2}, \frac{1}{16\sigma_X^4}, \frac{1}{256c(1+4\sigma_X^2)^2}\right)}.$$

Remark 4. The rows of $\mathbf{X}$ and $\mathbf{W}$ both have bounded moments and the measurement error is multiplicative: we have

$$\Pr(|||\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*|||_{\max} \geq \delta) \leq \frac{p^2 2^{4m}}{n^m \delta^{2m}}\left(\frac{1}{c_*} + C_m(K_m + 2)\right)$$

with $c_* = [C_m 2^{2m}\{\frac{1}{m_{\min}^{2m}}(2 + K_{m,X}K_{m,W}) + K_{m,X}\}]^{-1}$ and with tail function defined as

$$f_*(n, \delta) = \left\{\frac{2^{4m}}{n^m\delta^{2m}}(\frac{1}{c_*} + C_m(K_m + 2))\right\}^{-1}$$

and associated inverse functions taking the exact forms as the bounded moment additive error case with c_* defined as mentioned.

Proof. In this proof, first we want to bound the quantity $|||\mathbf{W}|||_{\max} = |||\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*|||_{\max}$. Recall that using the definition of the closeness condition for $\widetilde{\mathbf{\Sigma}}$ and $\mathbf{\Sigma}^*$ of the decay function f, we have the probability bound

$$\Pr(|||\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*|||_{\max} \geq \delta) \leq \frac{p^2}{f_*(n, \delta)} \qquad \text{for all } \delta \in \left(0, \frac{1}{v_*}\right].$$

Setting $\delta = \bar{\delta}_{f_*}(n, p^\gamma)$,we get

$$\Pr(|||\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*|||_{\max} \geq \bar{\delta}_{f_*}(n, p^\gamma)) \leq \frac{p^2}{f(n, \bar{\delta}_{f_*}(n, p^\gamma))} = \frac{1}{p^{\gamma-2}}.$$

The last equality follows by the definition of the inverse function $\bar{\delta}_f$ for a fixed n, and therefore $f(n, \bar{\delta}_{f_*}(n, p^\gamma)) = \bar{\delta}_{f_*}^{-1}\left\{\bar{\delta}_{f_*}(n, p^\gamma)\right\} = p^\gamma$. Therefore, we need to condition on the event $|||\mathbf{W}|||_{\max} \leq \bar{\delta}_{f_*}(n, p^\gamma)$ for further analysis.

Now, let us denote $\mathcal{A}$ as the event that $|||\mathbf{W}|||_{\max} \leq \bar{\delta}_{f_*}(n, p^\gamma)$. We verify that the assumption

$$\max\left\{|||\mathbf{W}|||_{\max}, |||R(\boldsymbol{\Delta})|||_{\max}\right\} \leq \frac{\alpha\lambda_n}{8}$$

of Lemma 9 holds. Recall the choice of regularization parameter $\lambda_n = (8/\alpha)\bar{\delta}_{f_*}(n, p^\gamma))$, we have $|||\mathbf{W}|||_{\max} \leq \bar{\delta}_{f_*}(n, p^\gamma) \leq \alpha\lambda_n/8$. Hence $|||\mathbf{W}|||_{\max} \leq \alpha\lambda_n/8$. Next we show that the condition also holds for $|||R(\boldsymbol{\Delta})|||_{\max}$. To do that we show that the condition required for Lemma 11 holds under the specified conditions on n and λ_n. From our choice of regularization constant $\lambda_n = (8/\alpha)\bar{\delta}_{f_*}(n, p^\gamma)$,

$$2\kappa_{\boldsymbol{\Gamma}^*}\left\{|||\mathbf{W}|||_{\max} + \lambda_n\right\} \leq 2\kappa_{\boldsymbol{\Gamma}^*}\left\{\bar{\delta}_{f_*}(n, p^\gamma) + (8/\alpha)\bar{\delta}_{f_*}(n, p^\gamma)\right\} = 2\kappa_{\boldsymbol{\Gamma}^*}\Big(1 + (8/\alpha)\Big)\bar{\delta}_f(n, p^\gamma),$$

for all $\bar{\delta}_{f_*}(n, p^\gamma) < 1/v_*$. From the monotonicity of the inverse tail function (3.13), we have that if $n > \bar{n}_{f_*}(\delta, p^\gamma)$ for some $\delta > 0$, then $\bar{\delta}_{f_*}(n, p^\gamma) \leq \delta$. Therefore, by using the lower bound on the sample size for all $\bar{\delta}_{f_*}(n, p^\gamma) < 1/v_*$, we get

$$\bar{\delta}_{f_*}(n, p^\gamma) \leq \delta \leq \frac{1}{6(1 + (8/\alpha))^2 d \max\left\{\kappa_{\boldsymbol{\Sigma}^*}\kappa_{\boldsymbol{\Gamma}^*}, \kappa^3_{\boldsymbol{\Sigma}^*}\kappa^2_{\boldsymbol{\Gamma}^*}\right\}}.$$

Consequently, we have

$$2\kappa_{\boldsymbol{\Gamma}^*}\left\{|||\mathbf{W}|||_{\max} + \lambda_n\right\} \leq 2\kappa_{\boldsymbol{\Gamma}^*}\Big(1 + (8/\alpha)\Big)\bar{\delta}_{f_*}(n, p^\gamma)$$

$$
\begin{aligned}
&\leq \frac{2\kappa_{\mathbf{\Gamma}^*}\Big(1+(8/\alpha)\Big)}{6(1+(8/\alpha))^2 d \max\left\{\kappa_{\mathbf{\Sigma}^*}\kappa_{\mathbf{\Gamma}^*}, \kappa^3_{\mathbf{\Sigma}^*}\kappa^2_{\mathbf{\Gamma}^*}\right\}} \\
&\leq \frac{1}{3d\Big(1+(8/\alpha)\Big)\max\left\{\kappa_{\mathbf{\Sigma}^*}, \kappa^3_{\mathbf{\Sigma}^*}\kappa_{\mathbf{\Gamma}^*}\right\}} \\
&\leq \frac{1}{3d\max\left\{\kappa_{\mathbf{\Sigma}^*}, \kappa^3_{\mathbf{\Sigma}^*}\kappa_{\mathbf{\Gamma}^*}\right\}} \\
&= \min\left\{\frac{1}{3\kappa_{\mathbf{\Sigma}^*}d}, \frac{1}{3\kappa^3_{\mathbf{\Sigma}^*}\kappa_{\mathbf{\Gamma}^*}d}\right\} \qquad (3.23)
\end{aligned}
$$

Therefore, the assumptions of Lemma 11 are satisfied and we can apply this lemma to conclude that

$$
|||\mathbf{\Delta}|||_{\max} = |||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_{\max} \leq 2\kappa_{\mathbf{\Gamma}^*}\left\{|||\mathbf{W}|||_{\max} + \lambda_n\right\} \leq 2\kappa_{\mathbf{\Gamma}^*}\Big(1+(8/\alpha)\Big)\bar{\delta}_{f_*}(n, p^\gamma). \qquad (3.24)
$$

Using inequalities (3.23) and (3.24), we see that the assumption $|||\mathbf{\Delta}|||_{\max} \leq 1/(3\kappa_{\mathbf{\Sigma}^*}d)$ of Lemma 10 holds. Plugging in the upper bound for $|||\mathbf{\Delta}|||_{\max}$ to the result of Lemma 10 we get

$$
\begin{aligned}
|||R(\mathbf{\Delta})|||_{\max} &\leq \frac{3}{2}d|||\mathbf{\Delta}|||^2_{\max}\kappa^3_{\mathbf{\Sigma}^*} \\
&\leq \frac{3}{2}d\kappa^3_{\mathbf{\Sigma}^*}4\kappa^2_{\mathbf{\Gamma}^*}\Big(1+(8/\alpha)\Big)^2\Big(\bar{\delta}_{f_*}(n, p^\gamma)\Big)^2 \\
&\leq \left\{6d\kappa^2_{\mathbf{\Gamma}^*}\kappa^3_{\mathbf{\Sigma}^*}\Big(1+(8/\alpha)\Big)^2\bar{\delta}_{f_*}(n, p^\gamma)\right\}\bar{\delta}_{f_*}(n, p^\gamma) \\
&\leq \left\{6d\kappa^2_{\mathbf{\Gamma}^*}\kappa^3_{\mathbf{\Sigma}^*}\Big(1+(8/\alpha)\Big)^2\bar{\delta}_{f_*}(n, p^\gamma)\right\}\frac{\alpha\lambda_n}{8} \\
&\leq \frac{\alpha\lambda_n}{8},
\end{aligned}
$$

where the final inequality follows from the condition on the sample size (3.21), and the monotonicity property (3.13), since

$$
\bar{\delta}_{f_*}(n, p^\gamma) \leq \delta \leq \frac{1}{6d\kappa^2_{\mathbf{\Gamma}^*}\kappa^3_{\mathbf{\Sigma}^*}(1+(8/\alpha))^2},
$$

so,

$$6d\kappa_{\mathbf{\Gamma}^*}^2\kappa_{\mathbf{\Sigma}^*}^3(1+(8/\alpha))^2\bar{\delta}_{f_*}(n,p^\gamma)\leq 1.$$

Hence, it is shown that the assumption (3.50) of Lemma 9 holds and we can conclude that the primal-dual witness construction succeeds with high probability. This would imply that the witness matrix $\widetilde{\mathbf{\Theta}}$ is equal to the solution $\widehat{\mathbf{\Theta}}$ to the original log-determinant problem (3.8) with high probability. Then the estimator $\widehat{\mathbf{\Theta}}$ satisfies the ℓ_∞-bound (3.24) of $\widetilde{\mathbf{\Theta}}$ as claimed in Theorem 1(a). Part (a) guarantees that $\widehat{\mathbf{\Theta}}$ is uniformly close to $\mathbf{\Theta}^*$ in an elementwise sense. We also have $\widehat{\mathbf{\Theta}}_{S^c}=\widetilde{\mathbf{\Theta}}_{S^c}=0$, as claimed in Theorem 1(b). Since the above was conditioned on the event $\mathcal{A}$, these statements hold with probability $\Pr(\mathcal{A})\geq 1-(1/p^{\gamma-2})$.

In the final part of the proof we derive specific forms for the function $\bar{\delta}_{f_*}(n,p^\gamma)$ for the event $\mathcal{A}$ for different distributional assumptions of $\mathbf{X}$ and $\mathbf{W}$. Recall from (3.9) that using triangular inequality, we have

$$|||\widetilde{\mathbf{\Sigma}}-\mathbf{\Sigma}^*|||_{\max}\leq |||\widetilde{\mathbf{\Sigma}}-\mathbf{S}|||_{\max}+|||\mathbf{S}-\mathbf{\Sigma}^*|||_{\max}.$$

Next, we show that we can approximate $|||\widetilde{\mathbf{\Sigma}}-\mathbf{S}|||_{\max}$ by $|||\widehat{\mathbf{\Sigma}}-\mathbf{S}|||_{\max}$. By the definition of $\widetilde{\mathbf{\Sigma}}$, $|||\widetilde{\mathbf{\Sigma}}-\widehat{\mathbf{\Sigma}}|||_{\max}\leq |||\mathbf{S}-\widehat{\mathbf{\Sigma}}|||_{\max}$. Combining this with the triangular inequality, we have

$$|||\widetilde{\mathbf{\Sigma}}-\mathbf{S}|||_{\max}\leq |||\widetilde{\mathbf{\Sigma}}-\widehat{\mathbf{\Sigma}}|||_{\max}+|||\widehat{\mathbf{\Sigma}}-\mathbf{S}|||_{\max}\leq 2|||\widehat{\mathbf{\Sigma}}-\mathbf{S}|||_{\max}.$$

Thus

$$\Pr(|||\widetilde{\mathbf{\Sigma}}-\mathbf{S}|||_{\max}\geq\delta)\leq\Pr(|||\widehat{\mathbf{\Sigma}}-\mathbf{S}|||_{\max}\geq\delta/2). \tag{3.25}$$

When the rows of $\mathbf{X}$ and $\mathbf{W}$ are both sub-Gaussian and the measurement error is additive, we have

$$\Pr(|||\widetilde{\mathbf{\Sigma}}-\mathbf{S}|||_{\max}\geq\delta)\leq\frac{p^2}{C}\exp\left\{-cn\min\left(\frac{\delta^2}{4\sigma_X^2\sigma_W^2},\frac{\delta^2}{4\sigma_W^4},\frac{\delta}{2\sigma_X\sigma_W},\frac{\delta}{2\sigma_W^2}\right)\right\}.$$

By (1) and (3.9) we have

$$\Pr\Big\{ |\!|\!| \mathbf{S} - \mathbf{\Sigma}^* |\!|\!|_{\max} \geq \delta \Big\} \leq \frac{p^2}{4} \exp\left\{ -\frac{n\delta^2}{64(1+4\sigma_X^2)^2} \right\}.$$

To put the pieces together we can write, for $\delta > 0$,

$$\begin{aligned}
&\Pr(|\!|\!| \widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^* |\!|\!|_{\max} \geq \delta) \\
&\leq \Pr(|\!|\!| \widetilde{\mathbf{\Sigma}} - \mathbf{S} |\!|\!|_{\max} + |\!|\!| \mathbf{S} - \mathbf{\Sigma}^* |\!|\!|_{\max} \geq \delta) \\
&\overset{\text{(i)}}{\leq} \Pr\left\{ \max\left(|\!|\!| \widetilde{\mathbf{\Sigma}} - \mathbf{S} |\!|\!|_{\max}, |\!|\!| \mathbf{S} - \mathbf{\Sigma}^* |\!|\!|_{\max} \right) \geq \delta/2 \right\} \\
&= \Pr\left\{ (|\!|\!| \widetilde{\mathbf{\Sigma}} - \mathbf{S} |\!|\!|_{\max} \geq \delta/2) \cup (|\!|\!| \mathbf{S} - \mathbf{\Sigma}^* |\!|\!|_{\max} \geq \delta/2) \right\} \\
&\leq \Pr(|\!|\!| \widetilde{\mathbf{\Sigma}} - \mathbf{S} |\!|\!|_{\max} \geq \delta/2) + \Pr(|\!|\!| \mathbf{S} - \mathbf{\Sigma}^* |\!|\!|_{\max} \geq \delta/2) \\
&\leq \frac{p^2}{C} \exp\left\{ -cn \min\left(\frac{\delta^2}{16\sigma_X^2\sigma_W^2}, \frac{\delta^2}{16\sigma_W^4}, \frac{\delta}{4\sigma_X\sigma_W}, \frac{\delta}{4\sigma_W^2} \right) \right\} \\
&\quad + \frac{p^2}{4} \exp\left\{ -\frac{n\delta^2}{256(1+4\sigma_X^2)^2} \right\} \\
&\leq \frac{p^2}{C} \exp\left\{ -cn \min\left(\frac{\delta^2}{16\sigma_X^2\sigma_W^2}, \frac{\delta^2}{16\sigma_W^4}, \frac{\delta}{4\sigma_X\sigma_W}, \frac{\delta}{4\sigma_W^2}, \frac{\delta^2}{256c(1+4\sigma_X^2)^2} \right) \right\} \\
&\quad + \frac{p^2}{4} \exp\left\{ -cn \min\left(\frac{\delta^2}{16\sigma_X^2\sigma_W^2}, \frac{\delta^2}{16\sigma_W^4}, \frac{\delta}{4\sigma_X\sigma_W}, \frac{\delta}{4\sigma_W^2}, \frac{\delta^2}{256c(1+4\sigma_X^2)^2} \right) \right\} \\
&\leq p^2(\frac{1}{C} + \frac{1}{4}) \exp\left\{ -cn \min\left(\frac{\delta^2}{16\sigma_X^2\sigma_W^2}, \frac{\delta^2}{16\sigma_W^4}, \frac{\delta}{4\sigma_X\sigma_W}, \frac{\delta}{4\sigma_W^2}, \frac{\delta^2}{256c(1+4\sigma_X^2)^2} \right) \right\} \\
&\leq \frac{p^2(4+C)}{4C} \exp\left\{ -cn \min\left(\frac{\delta^2}{16\sigma_X^2\sigma_W^2}, \frac{\delta^2}{16\sigma_W^4}, \frac{\delta}{4\sigma_X\sigma_W}, \frac{\delta}{4\sigma_W^2}, \frac{\delta^2}{256c(1+4\sigma_X^2)^2} \right) \right\}
\end{aligned}$$

Inequality (i) is due to the relationship $A + B + \mid A - B \mid = 2\max(A, B)$, which implies $A + B \leq 2\max(A, B)$. If we consider the particular case when the rows of both $\mathbf{X}$ and $\mathbf{W}$ are multivariate Gaussian, the above expression simplifies to

$$\Pr(|\!|\!| \widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^* |\!|\!|_{\max} \geq \delta) \leq \frac{p^2(4+C)}{4C} \exp\left\{ -cn\delta^2 \min\left(\frac{1}{16\sigma_X^2\sigma_W^2}, \frac{1}{16\sigma_W^4}, \frac{1}{256c(1+4\sigma_X^2)^2} \right) \right\}. \tag{3.26}$$

Let us define the tail function obtained from this bound as

$$f_*(n,\delta) = \frac{4C}{4+C}\exp\left\{cn\delta^2 \min\left(\frac{1}{16\sigma_X^2\sigma_W^2}, \frac{1}{16\sigma_W^4}, \frac{1}{256c(1+4\sigma_X^2)^2}\right)\right\}.$$

Setting $\bar{\delta}_{f_*}(n,p^\gamma) \leq \delta$, some further calculations show that the associated inverse functions defined in (3.11) and (3.12) take the form

$$\bar{\delta}_{f_*}(n,p^\gamma) = \sqrt{\frac{\log\left(\frac{(4+C)p^\gamma}{4C}\right)}{cn\min\left(\frac{1}{16\sigma_X^2\sigma_W^2}, \frac{1}{16\sigma_W^4}, \frac{1}{256c(1+4\sigma_X^2)^2}\right)}},$$

and

$$\bar{n}_{f_*}(\delta,p^\gamma) = \frac{\log\left(\frac{(4+C)p^\gamma}{4C}\right)}{c\delta^2\min\left(\frac{1}{16\sigma_X^2\sigma_W^2}, \frac{1}{16\sigma_W^4}, \frac{1}{256c(1+4\sigma_X^2)^2}\right)}.$$

When the rows of $\mathbf{X}$ and $\mathbf{W}$ both have bounded moments and the measurement error is additive, we have,

$$\Pr(|||\widetilde{\mathbf{\Sigma}} - \mathbf{S}|||_{\max} \geq \delta) \leq \frac{p^2 2^{2m}}{c_* n^m \delta^{2m}}$$

with $c_* = [C_m 2^{4m}\{6^m + 2(1+K_{m,X})(1+K_{m,W}) + K_{m,W} + (1+2^m)\}]^{-1}$. By (1) and (3.9) we have

$$\Pr\Big\{|||\mathbf{S} - \mathbf{\Sigma}^*|||_{\max} \geq \delta\Big\} \leq \frac{p^2 2^{2m} C_m(K_m+2)}{n^m\delta^{2m}}.$$

To put the pieces together, for $\delta > 0$,

$$\begin{aligned}\Pr(|||\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*|||_{\max} \geq \delta) &\leq \frac{p^2 2^{4m}}{c_* n^m\delta^{2m}} + \frac{p^2 2^{4m} C_m(K_m+2)}{n^m\delta^{2m}} \\ &\leq \frac{p^2 2^{4m}}{n^m\delta^{2m}}\left(\frac{1}{c_*} + C_m(K_m+2)\right) \qquad (3.27)\end{aligned}$$

with $f_*(n,\delta) = \{(2^{4m})/(n^m\delta^{2m})(\frac{1}{c_*} + C_m(K_m+2))\}^{-1}$. Therefore, the associated inverse functions for $\bar{\delta}_{f_*}(n,p^\gamma) \leq \delta$,

$$\bar{\delta}_{f_*}(n,p^\gamma) = \frac{\left(2^{4m}\left(\frac{1}{c_*} + C_m(K_m+2)\right)p^\gamma\right)^{1/2m}}{\sqrt{n}}$$

and

$$\bar{n}_{f_*}(\delta,p^\gamma) = \frac{\left(2^{4m}\left(\frac{1}{c_*} + C_m(K_m+2)\right)p^\gamma\right)^{1/m}}{\delta^2}.$$

When the rows of $\mathbf{X}$ and $\mathbf{W}$ are both sub-Gaussian and the measurement error is multiplicative, we have,

$$\Pr(|||\widetilde{\mathbf{\Sigma}} - \mathbf{S}|||_{\max} \geq \delta) \leq \frac{p^2}{C}\exp\left\{-cn\min\left(\frac{\delta^2}{4\sigma_X^4\sigma_W^4}, \frac{\delta^2}{4\sigma_X^4}, \frac{\delta}{2\sigma_X^2}, \frac{\sqrt{\delta}}{\sqrt{2\sigma_X^2\sigma_W^2}}\right)\right\}.$$

For $\delta > 0$,

$$\begin{aligned}
&\Pr(|||\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*|||_{\max} \geq \delta) \\
&\leq \frac{p^2}{C}\exp\left\{-cn\min\left(\frac{\delta^2}{16\sigma_X^4\sigma_W^4}, \frac{\delta^2}{16\sigma_X^4}, \frac{\delta}{4\sigma_X^2}, \frac{\sqrt{\delta}}{2\sqrt{\sigma_X^2\sigma_W^2}}\right)\right\} \\
&\qquad + \frac{p^2}{4}\exp\left\{-\frac{n\delta^2}{256(1+4\sigma_X^2)^2}\right\} \\
&\leq \frac{p^2(4+C)}{4C}\exp\left\{-cn\min\left(\frac{\delta^2}{16\sigma_X^4\sigma_W^4}, \frac{\delta^2}{16\sigma_X^4}, \frac{\delta}{4\sigma_X^2}, \frac{\sqrt{\delta}}{2\sqrt{\sigma_X^2\sigma_W^2}}, \frac{\delta^2}{256c(1+4\sigma_X^2)^2}\right)\right\}
\end{aligned}$$

If we consider the particular case when the rows of both $\mathbf{X}$ and $\mathbf{W}$ are multivariate Gaussian, the above expression simplifies to

$$\Pr(|||\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*|||_{\max} \geq \delta) \leq \frac{p^2(4+C)}{4C}\exp\left\{-cn\delta^2\min\left(\frac{1}{16\sigma_X^2\sigma_W^2}, \frac{1}{16\sigma_X^4}, \frac{1}{256c(1+4\sigma_X^2)^2}\right)\right\} \tag{3.28}$$

with tail function defined as

$$f_*(n,\delta) = \frac{4C}{4+C}\exp\left\{cn\delta^2 \min\left(\frac{1}{16\sigma_X^2\sigma_W^2}, \frac{1}{16\sigma_X^4}, \frac{1}{256c(1+4\sigma_X^2)^2}\right)\right\}.$$

Setting $\bar{\delta}_{f_*}(n,p^\gamma) \leq \delta$, the associated inverse functions take the form

$$\bar{\delta}_{f_*}(n,p^\gamma) = \sqrt{\frac{\log\left(\frac{(4+C)p^\gamma}{4C}\right)}{cn\min\left(\frac{1}{16\sigma_X^2\sigma_W^2}, \frac{1}{16\sigma_X^4}, \frac{1}{256c(1+4\sigma_X^2)^2}\right)}}$$

and

$$\bar{n}_{f_*}(\delta,p^\gamma) = \frac{\log\left(\frac{(4+C)p^\gamma}{4C})\right)}{c\delta^2\min\left(\frac{1}{16\sigma_X^2\sigma_W^2}, \frac{1}{16\sigma_X^4}, \frac{1}{256c(1+4\sigma_X^2)^2}\right)}.$$

When the rows of $\mathbf{X}$ and $\mathbf{W}$ both have bounded moments and the measurement error is multiplicative, we get for $\delta > 0$,

$$\Pr(|||\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*|||_{\max} \geq \delta) \leq \frac{p^2 2^{4m}}{n^m\delta^{2m}}\left(\frac{1}{c_*} + C_m(K_m+2)\right) \tag{3.29}$$

with $c_* = [C_m 2^{2m}\{\frac{1}{m_{\min}^{2m}}(2+K_{m,X}K_{m,W}) + K_{m,X}\}]^{-1}$. We get

$$f_*(n,\delta) = \left[\frac{2^{4m}}{n^m\delta^{2m}}\Big(\frac{1}{c_*} + C_m(K_m+2)\Big)\right]^{-1}.$$

Therefore, the functional form of the associated inverse functions are exactly the same as the bounded moments additive error case with c_* defined as mentioned. □

3.5.2 Model Selection Consistency

The following theorem provides sufficient conditions to link the sample size n and the minimum value

$$\theta_{\min} := \min_{(i,j)\in E(\mathbf{\Theta}^*)} |\ \mathbf{\Theta}_{ij}^*\ | \tag{3.30}$$

that allows us to study model selection consistency. Let us define the event that the estimator $\widehat{\mathbf{\Theta}}$ has the same edge set as $\mathbf{\Theta}^*$, that is, CoGlasso recovers the full edge set correctly and recovers the correct signs on these edges as well.

$$\mathcal{M}(\widehat{\mathbf{\Theta}}; \mathbf{\Theta}^*) := \left\{ \operatorname{sgn}(\widehat{\mathbf{\Theta}}_{ij}) = \operatorname{sgn}(\mathbf{\Theta}^*_{ij}) \qquad \forall (i,j) \in E(\mathbf{\Theta}^*) \right\}. \tag{3.31}$$

Theorem 2. *Under the same conditions as Theorem 1, suppose that the sample size satisfies the lower bound*

$$n > \bar{n}_f \left(\frac{1}{\max\left\{ 2\kappa_{\mathbf{\Gamma}^*}\Big(1 + (8/\alpha)\Big)\theta_{\min}^{-1}, v_*, 6(1+(8/\alpha))d \max\left\{ \kappa_{\mathbf{\Sigma}^*}\kappa_{\mathbf{\Gamma}^*}, \kappa^3_{\mathbf{\Sigma}^*}\kappa^2_{\mathbf{\Gamma}^*} \right\} \right\}}, p^\gamma \right). \tag{3.32}$$

Then the estimator is model selection consistent with high probability as $p \to \infty$, specifically

$$\Pr(\mathcal{M}(\widehat{\mathbf{\Theta}}; \mathbf{\Theta}^*)) \geq 1 - \frac{1}{p^{\gamma-2}}. \tag{3.33}$$

If we compare the restrictions imposed on the lower bounds on the sample sizes of Theorem 1 and 2, we can see that Theorem 2 differs only in terms of the additional quantity $2\kappa_{\mathbf{\Gamma}^*}(1 + 8/\alpha)/\theta_{\min}$. This quantity acts as a constraint on how quickly the minimum can decay as a function of (n, p). The proof of the theorem is given in Section 3.9.

The following corollary can be established similar to Ravikumar et al. (2011) in terms of the Frobenius and spectral norm. The proof of the corollary is provided in Section 3.9.

Corollary 1. *[Rates in Frobenius and Operator Norm.] Under the same assumptions as Theorem 1, with probability at least $1 - (1/p^{\gamma-2})$, the estimator $\widehat{\mathbf{\Theta}}$ satisfies*

$$|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_F \leq \left\{ 2\kappa_{\mathbf{\Gamma}^*}\Big(1 + (8/\alpha)\Big) \right\} \left(\sqrt{s+p}\right) \bar{\delta}_{f_*}(n, p^\tau) \tag{3.34}$$

and

$$|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_2 \leq \left\{ 2\kappa_{\mathbf{\Gamma}^*}\Big(1 + (8/\alpha)\Big) \right\} \min\left\{ \sqrt{s+p}, d \right\} \bar{\delta}_{f_*}(n, p^\tau). \tag{3.35}$$

Notice that the Corollary differs from the classical graphical Lasso results provided in Ravikumar et al. (2011) in terms of the quantity $\bar{\delta}_{f_*}(n, p^\tau)$, which varies depending on the distributional assumptions on X and W and the type of measurement errors under consideration. Recall that, the expressions of $\bar{\delta}_{f_*}(n, p^\tau)$ for particular cases are derived in the proof of Theorem 1.

3.6 Simulation

To test our method, we would utilize two types of data generating scenarios which would result in indefinite covariance estimators. For the additive measurement error, we would use the Kronecker sum type model used in a graphical model estimation setting by Park et al. (2017) and in a regression setting by Rudelson and Zhou (2017). For multiplicative errors, we would use the missing data model described in Loh and Wainwright (2012).

3.6.1 Additive Model

Following Fan et al. (2019) and Rudelson and Zhou (2017), we use the Kronecker sum type covariance to generate the corrupted and observable data matrix $\mathbf{Z}$ based on the clean but unobservable data matrix $\mathbf{X}$. Let the data matrix that we want to generate be defined as $\mathbf{Z} = \mathbf{X}_0\mathbf{A}^{1/2} + \mathbf{B}^{1/2}\mathbf{W}_0$, where the first component contains the signal and has independent sub-Gaussian row vectors and the second component contains the random noise matrix with independent columns but dependent rows. Here, $\mathbf{A}^{1/2}$ and $\mathbf{B}^{1/2}$ are the unique square root of the positive definite matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$ and $\mathbf{B} \in \mathbb{R}^{n\times n}$, respectively and represent the covariance structures of $\mathbf{X}_0$ and $\mathbf{W}_0$, respectively. We generate $\mathbf{X}_0, \mathbf{W}_0 \in \mathbb{R}^{n\times p}$ as independent mean-zero sub-Gaussian random matrices. Note that $\mathbf{W}_0$ and $\mathbf{X}_0$ are independent.

Our primary interest is to estimate the precision matrix with sparse off-diagonal entries, $\mathbf{\Theta}^* = \mathbf{A}^{-1}$. As shown in Rudelson and Zhou (2017), that the covariance model becomes

unidentifiable if one of the traces of $\mathbf{A}$ or $\mathbf{B}$ is not assumed known. We assume that the trace of $\mathbf{A}$ is a known constant, that is, $\mathrm{tr}(\mathbf{A}) = p$, and we construct and estimator for $\mathrm{tr}(\mathbf{B})$ as

$$\widehat{\mathrm{tr}}(\mathbf{B}) = \frac{1}{p}\Big(|||\mathbf{X}|||_F^2 - n\mathrm{tr}(\mathbf{A})\Big)_+ \quad \text{and define} \quad \widehat{\tau_B} = \frac{1}{n}\widehat{\mathrm{tr}}(\mathbf{B}) \geq 0$$

where $(a)_+ = a \vee 0$ and $|||\mathbf{X}|||_F^2 = \sum_i \sum_j x_{ij}^2$.

In the true model, τ_B is the variance of the noise variable. We normalize $\mathbf{B}$ in the covariance generation process to make sure the assumption $\mathrm{tr}(\mathbf{B}) = n\tau_B$ holds. Therefore, the surrogate covariance estimate for $\mathbf{A}$ is given by as shown in Rudelson and Zhou (2017)

$$\widehat{\mathbf{\Sigma}}_{KS} = \frac{1}{n}\mathbf{Z}^\top\mathbf{Z} - \frac{\widehat{\mathrm{tr}}(\mathbf{B})}{n}\mathbf{I}_p.$$

Note that, when $p > n$, this estimator is guaranteed to not be positive semi-definite.

Covariance models:

To generate data from the above mentioned simulation settings, we consider the following covariance models for $\mathbf{A}$ and $\mathbf{B}$. For the precision matrix, $\mathbf{\Theta}^* = \mathbf{A}^{-1} = (\omega_{ij})$ and $\boldsymbol{\Pi} = \mathbf{B}^{-1} = (\nu_{ij})$. respectively. Following Rudelson and Zhou (2017), we choose $\mathbf{A}$ from the following models:

- AR(1) model: To obtain a chain graph for the precision matrix $\mathbf{A}^{-1}$, we set the form of the covariance, $\mathbf{A} = \left\{\rho^{|i-j|}\right\}_{i,j}$.

We choose $\mathbf{B}$ from the following covariance models. Note that $\tau_B = \mathrm{tr}(\mathbf{B})/n$.

- Erdös-Rényi random graph: We consider a type of Erdös-Rényi random graph for $\boldsymbol{\Pi} = \mathbf{B}^{-1}$. We start by setting $\boldsymbol{\Pi} = c\mathbf{I}_{n\times n}$ where c is a constant. Next, we randomly select $n \log n$ edges and update $\boldsymbol{\Pi}$ as follows: for each new edge (i, j), a weight $w > 0$ is chosen uniformly at random from $[w_{\min}, w_{\max}]$ where $w_{\max} > w_{\min} > 0$; we

subtract w from ν_{ij} and ν_{ji} and increase ν_{ii} and ν_{jj} by w. This maintains the positive definiteness of $\boldsymbol{\Pi}$. Then we rescale $\mathbf{B}$ to have a certain desired trace parameter τ_B.

3.6.2 Missing Data Model

Following Loh and Wainwright (2012) we use the missing data model with missing-completely-at-random (MCAR) observations to estimate our graphical model. Let $\mathbf{X}_0 \in \mathbb{R}^{n\times p}$ be an independent mean-zero sub-Gaussian random matrix. Let $\mathbf{W} \in \{0,1\}^{n\times p}$ with $W_{ij} \sim Bernoulli(1-\pi_j)$. This means that the entries of the jth column of the data matrix is observed with probability π_j. Note that $\mathbf{W}$ is independent of $\mathbf{X}$. We can generate the observed matrix as $\mathbf{X} = \mathbf{X}_0\mathbf{A}^{1/2}$ and the unobserved matrix $\mathbf{Z} = \mathbf{X} \odot \mathbf{W}$ where $\odot$ denotes the Hadamard, or elementwise product. Then the surrogate estimate for $\mathbf{A}$ can be given by the estimator $\widehat{\boldsymbol{\Sigma}}_{\text{miss}}$ defined in (3.14). Since the off-diagonal entries are divided by smaller values, $\widehat{\boldsymbol{\Sigma}}_{\text{miss}}$ will not necessarily be positive semi-definite.

Covariance models:

For the missing data model, we considered two types of precision matrix for $\boldsymbol{\Theta}^*$, based on the chain graph and the Erdös-Rényi random graph. The construction of the matrices are similar to as it is explained above.

Tuning parameter selection

In practice, the parameter λ must be tuned for all the models. Two methods are used, namely, cross validation and BIC criterion to tune the models as decribed in Chapter 2.

- Cross-validation: We performed a five-fold cross-validation method to tune λ. We estimate the precision matrix $\widehat{\boldsymbol{\Theta}}$ from the training set and validate it on the test set. We calculate the cross-validation score given in (2.5) for each fold described in Section 2.1.1. The observed log-likelihood was used as the loss function. Then we

find the best $\widehat{\lambda}$ that maximizes $CV(\lambda)$. Finally, using the chosen $\widehat{\lambda}$, a final estimator of the precision matrix is calculated using all the data. We used used a sequence of equally spaced values for the tuning parameter in the logarithmic scale from $[-2, 2]$ to perform the cross-validation.

- BIC: We used the BIC criterion as defined in (2.6) to tune the penalty parameter λ. The optimal value of the tuning parameter is taken to be the minimizer of the criterion. We used a sequence of equally spaced values for the tuning parameter in the logarithmic scale from $[-2, 2]$ to perform the cross-validation.

Methods comparison

We compared our results with two other methods. The first method is the ADMM algorithm described in Fan et al. (2019) Algorithm 1 solving for the non-convex objective function. The algorithm required the knowledge of an upper bound for the true precision matrix, that is, $|||\mathbf{\Theta}|||_2 \leq R$. In the simulations, we chose the value of R to be two times the magnitude of the spectral norm of the true precision matrix.

The second method is a nodewise regression where we adapted the algorithm for graphical model following Meinshausen and Bühlmann (2006); Yuan and Lin (2007). This method is slightly naive in a sense that we used the `R` package `glmnet` to perform the column by column Lasso regressions in the first step of the algorithm directly on the noisy data. Specifically, we can perform p Lasso-type regressions to obtain estimates $\widehat{\beta}_j$ and form estimates $\widehat{a}_j$, where $\widehat{\beta}_j$ are the estimated coefficients from the Lasso regression and $\widehat{a}_j = -(\widehat{\mathbf{\Sigma}}_{jj} - \widehat{\mathbf{\Sigma}}_{j,-j}\widehat{\beta}_j)^{-1}$ based on the surrogate estimate of $\mathbf{\Sigma}$. Since the unbiased surrogates could be unbounded from below and not be positive semi-definite due to noisy data, we projected the estimate to the nearest positive semi-definite cone. Next, we formed $\tilde{\mathbf{\Theta}}_{j,-j} = \widehat{a}_j\widehat{\beta}_j$ and $\tilde{\mathbf{\Theta}}_{jj} = -\widehat{a}_j$. In the last step, we symmetrize the results to obtain $\widehat{\mathbf{\Theta}} = \arg\min_{S^p} |||\mathbf{\Theta} - \tilde{\mathbf{\Theta}}|||_{\max}$, where S^p is the set of symmetric matrices.

For the nodewise regression method, we only used cross-validation to choose the tuning parameter. For our proposed method CoGlasso and the ADMM approach (Fan et al., 2019), we tuned λ with both cross-validation and BIC criterion.

Performance metrics

- We calculated relative Frobenius, spectral, nuclear and ℓ_1 norm for the statistical error of the estimation, which is defined as $|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_{\text{norm}}/|||\mathbf{\Theta}^*|||_{\text{norm}}$. We also calculated the false positive rate (FPR), false negative rate (FNR) and the true positive rate (TPR) as defined below:

$$FPR = \frac{FP}{TN + FP}; \qquad FNR = \frac{FN}{FN + TP}; \qquad TPR = \frac{TP}{FN + TP}$$

  where TP is the number of true positives (true non-zero edges that are estimated as such), TN is the number of true negatives (true zero edges that are recognized as such), FP is the number of false positives (true zero edges that are estimated as non-zero) and FN is the number of false negatives (true non-zero edges that are estimated as zero). All the results are averaged across 100 replications.

3.6.3 Simulation Results

Additive error case:

We generated the true precision matrix $\mathbf{\Theta}^*$ as a chain graph with parameter ρ such that $\Sigma_X = \mathbf{\Theta}^{*-1} \sim AR(1)$ with $\rho = \{-0.7, -0.5, -0.3\}$ with two different samples sizes, $n = \{160, 240\}$, two different variance parameter for the noise, $\tau_B = \{0.3, 0.5\}$ and the number of parameters $p = 400$. The covariance structure for the noise, $\Sigma_W = \mathbf{\Pi}^{-1}$, that is, it is chosen from an Erdös-Rényi random graph such that $\mathbf{\Pi} \sim ER$ with $n\log(n)$ randomly

selected edges in $\mathbf{\Pi}$ construction with partial correlation randomly chosen from a uniform distribution $Unif(0.6, 0.8)$. Table 3.1-3.12 present the results of the additive error cases.

Table 3.1: *Scenario A:* $p = 400, n = 160, \rho = -0.7, \tau_B = 0.3$

Method	Criteria	Frobenius	Spectral	Nuclear	ℓ_1	FPR	FNR	TPR
CoGlasso	CV	0.639	0.743	0.594	0.976	0.028	0.000	1.000
	BIC	0.673	0.761	0.630	0.915	0.014	0.001	0.999
ADMM	CV	1.201	1.380	1.178	2.594	0.371	0.000	1.000
	BIC	1.791	1.746	1.890	9.383	0.840	0.000	1.000
Nodewise	CV	0.594	1.784	0.539	2.205	0.015	0.001	0.999

Table 3.2: *Scenario B:* $p = 400, n = 160, \rho = -0.7, \tau_B = 0.5$

Method	Criteria	Frobenius	Spectral	Nuclear	ℓ_1	FPR	FNR	TPR
CoGlasso	CV	0.718	0.803	0.675	1.034	0.022	0.003	0.997
	BIC	0.718	0.803	0.675	1.034	0.022	0.003	0.997
ADMM	CV	1.149	1.395	1.095	2.638	0.339	0.002	0.998
	BIC	1.784	1.750	1.884	9.337	0.842	0.000	1.000
Nodewise	CV	0.732	2.469	0.613	2.851	0.018	0.004	0.996

Table 3.3: *Scenario C:* $p = 400, n = 240, \rho = -0.7, \tau_B = 0.3$

Method	Criteria	Frobenius	Spectral	Nuclear	ℓ_1	FPR	FNR	TPR
CoGlasso	CV	0.586	0.695	0.537	0.837	0.013	0.000	1.000
	BIC	0.540	0.668	0.490	0.895	0.029	0.000	1.000
ADMM	CV	1.308	1.394	1.317	2.640	0.466	0.000	1.000
	BIC	1.669	1.645	1.741	7.700	0.777	0.000	1.000
Nodewise	CV	0.482	0.728	0.461	0.978	0.015	0.000	1.000

Table 3.4: *Scenario D:* $p = 400, n = 240, \rho = -0.7, \tau_B = 0.5$

Method	Criteria	Frobenius	Spectral	Nuclear	ℓ_1	FPR	FNR	TPR
CoGlasso	CV	0.678	0.768	0.633	0.905	0.011	0.000	1.000
	BIC	0.642	0.749	0.594	0.975	0.023	0.000	1.000
ADMM	CV	1.251	1.401	1.234	2.654	0.434	0.000	1.000
	BIC	1.656	1.650	1.727	7.711	0.779	0.000	1.000
Nodewise	CV	0.574	0.729	0.546	0.938	0.018	0.000	1.000

Table 3.5: *Scenario E:* $p = 400, n = 160, \rho = -0.5, \tau_B = 0.3$

Method	Criteria	Frobenius	Spectral	Nuclear	ℓ_1	FPR	FNR	TPR
CoGlasso	CV	0.603	0.717	0.553	1.117	0.025	0.053	0.947
	BIC	0.629	0.727	0.579	0.990	0.009	0.109	0.891
ADMM	CV	0.638	1.262	0.560	2.663	0.096	0.009	0.991
	BIC	1.684	1.753	1.694	10.493	0.767	0.000	1.000
Nodewise	CV	0.587	2.073	0.516	2.609	0.018	0.061	0.939

Table 3.6: *Scenario F:* $p = 400, n = 160, \rho = -0.5, \tau_B = 0.5$

Method	Criteria	Frobenius	Spectral	Nuclear	ℓ_1	FPR	FNR	TPR
CoGlasso	CV	0.673	0.775	0.618	1.212	0.025	0.108	0.892
	BIC	0.698	0.784	0.643	1.077	0.007	0.252	0.748
ADMM	CV	0.653	1.318	0.569	2.748	0.088	0.025	0.975
	BIC	1.702	1.765	1.719	10.590	0.795	0.000	1.000
Nodewise	CV	0.684	2.388	0.577	2.885	0.019	0.132	0.868

Table 3.7: *Scenario G:* $p = 400, n = 240, \rho = -0.5, \tau_B = 0.3$

Method	Criteria	Frobenius	Spectral	Nuclear	ℓ_1	FPR	FNR	TPR
CoGlasso	CV	0.546	0.657	0.501	0.896	0.009	0.018	0.982
	BIC	0.546	0.657	0.501	0.896	0.009	0.018	0.982
ADMM	CV	0.479	1.062	0.419	1.956	0.089	0.000	1.000
	BIC	1.583	1.700	1.545	10.503	0.744	0.000	1.000
Nodewise	CV	0.476	0.720	0.446	1.057	0.018	0.007	0.993

Table 3.8: *Scenario H:* $p = 400, n = 240, \rho = -0.5, \tau_B = 0.5$

Method	Criteria	Frobenius	Spectral	Nuclear	ℓ_1	FPR	FNR	TPR
CoGlasso	CV	0.610	0.722	0.558	1.129	0.022	0.034	0.966
	BIC	0.627	0.729	0.576	1.038	0.012	0.064	0.936
ADMM	CV	0.768	1.266	0.665	2.910	0.113	0.004	0.996
	BIC	1.579	1.700	1.560	9.853	0.721	0.000	1.000
Nodewise	CV	0.557	0.754	0.515	1.045	0.021	0.025	0.975

Overall, tuning parameter selection with cross-validation tends to perform well in all of the scenarios. In the case of additive error, CoGlasso clearly performs better than the non-convex approach of the analysis. When the partial correlation in the true precision matrix is set to be stronger, we can see that CoGlasso and nodewise regression perform well, and the ADMM method performs better only in the case of moderate partial correlation in the

Table 3.9: *Scenario I:* $p = 400, n = 160, \rho = -0.3, \tau_B = 0.3$

Method	Criteria	Frobenius	Spectral	Nuclear	ℓ_1	FPR	FNR	TPR
CoGlasso	CV	0.545	0.685	0.472	1.009	0.004	0.753	0.247
	BIC	0.553	0.692	0.477	0.782	0.000	0.992	0.008
ADMM	CV	0.430	0.676	0.383	1.561	0.040	0.356	0.644
	BIC	1.545	1.654	1.483	11.768	0.827	0.014	0.986
Nodewise	CV	0.619	3.227	0.454	3.787	0.010	0.679	0.321

Table 3.10: *Scenario J:* $p = 400, n = 160, \rho = -0.3, \tau_B = 0.5$

Method	Criteria	Frobenius	Spectral	Nuclear	ℓ_1	FPR	FNR	TPR
CoGlasso	CV	0.609	0.742	0.531	1.151	0.006	0.831	0.169
	BIC	0.614	0.739	0.534	0.826	0.000	0.997	0.003
ADMM	CV	0.466	0.734	0.413	1.625	0.035	0.478	0.521
	BIC	1.496	1.639	1.432	10.583	0.758	0.031	0.969
Nodewise	CV	0.636	2.584	0.502	3.094	0.010	0.749	0.251

Table 3.11: *Scenario K:* $p = 400, n = 240, \rho = -0.3, \tau_B = 0.3$

Method	Criteria	Frobenius	Spectral	Nuclear	ℓ_1	FPR	FNR	TPR
CoGlasso	CV	0.488	0.628	0.425	0.904	0.004	0.578	0.422
	BIC	0.491	0.629	0.428	0.860	0.003	0.645	0.355
ADMM	CV	0.378	0.614	0.336	1.479	0.047	0.165	0.835
	BIC	1.404	1.610	1.311	10.435	0.688	0.007	0.993
Nodewise	CV	0.444	0.781	0.390	1.215	0.013	0.463	0.537

Table 3.12: *Scenario L:* $p = 400, n = 240, \rho = -0.3, \tau_B = 0.5$

Method	Criteria	Frobenius	Spectral	Nuclear	ℓ_1	FPR	FNR	TPR
CoGlasso	CV	0.559	0.694	0.484	1.036	0.004	0.723	0.277
	BIC	0.564	0.696	0.487	0.890	0.001	0.875	0.125
ADMM	CV	0.417	0.674	0.370	1.570	0.041	0.272	0.728
	BIC	1.431	1.622	1.345	10.610	0.726	0.012	0.988
Nodewise	CV	0.512	0.778	0.444	1.128	0.013	0.577	0.423

true precision matrix. When partial correlation is weaker, all the methods perform poorly. As expected, with increasing sample size, performance of the methods improve. With a larger effect of additive noise the performance metrics deteriorate, as expected. Nodewise regression performs comparably well along with the CoGlasso, which is expected since we are employing the same projection method in the nodewise regression as in CoGlasso.

Missing data case:

Following the missing data simulation setup as Fan et al. (2019), we used sample sizes $n = \{80, 130, 250, 700\}$, with corresponding probability of being missing, $P_{\text{miss}} = 1 - \pi_j = \{0.1, 0.3, 0.5, 0.7\}$, respectively to keep the effective sample size to be around 62 to 65. In all the settings the effective sample size is calculated as $n \times (1 - P_{\text{miss}})^2$ is kept constant. We kept $p = 400$ for all scenarios and generated the true precision matrix $\mathbf{\Theta}^*$ from two different setups.

Setup 1:

The true precision matrix is generated from a chain graph with strong partial correlation $\rho = -0.7$. The results are shown in Table 3.13 - 3.16.

Table 3.13: *Scenario M:* $p = 400, n = 80, P_{\text{miss}} = 0.1$

Method	Criteria	Frobenius	Spectral	Nuclear	ℓ_1	FPR	FNR	TPR
CoGlasso	CV	0.599	0.695	0.561	0.884	0.027	0.001	0.999
	BIC	0.645	0.724	0.609	0.842	0.011	0.004	0.996
ADMM	CV	0.577	0.687	0.540	1.063	0.097	0.000	1.000
	BIC	2.132	1.942	2.335	13.925	0.996	0.000	1.000
Nodewise	CV	1.017	7.229	0.589	7.701	0.008	0.004	0.996

Table 3.14: *Scenario N:* $p = 400, n = 130, P_{\text{miss}} = 0.3$

Method	Criteria	Frobenius	Spectral	Nuclear	ℓ_1	FPR	FNR	TPR
CoGlasso	CV	0.701	0.781	0.659	0.940	0.027	0.003	0.997
	BIC	0.728	0.796	0.689	0.892	0.011	0.011	0.989
ADMM	CV	0.631	0.745	0.589	1.061	0.075	0.002	0.998
	BIC	1.934	1.883	2.033	13.824	0.942	0.000	1.000
Nodewise	CV	0.665	1.665	0.620	2.057	0.013	0.014	0.986

Table 3.15: *Scenario O:* $p = 400, n = 250, P_{\text{miss}} = 0.5$

Method	Criteria	Frobenius	Spectral	Nuclear	ℓ_1	FPR	FNR	TPR
CoGlasso	CV	0.774	0.833	0.734	0.917	0.012	0.020	0.980
	BIC	0.775	0.834	0.736	0.914	0.010	0.021	0.979
ADMM	CV	0.621	0.758	0.576	1.164	0.085	0.012	0.988
	BIC	1.604	1.750	1.614	11.261	0.736	0.001	0.999
Nodewise	CV	0.740	0.813	0.703	0.897	0.017	0.023	0.977

Table 3.16: *Scenario P:* $p = 400, n = 700, P_{\text{miss}} = 0.7$

Method	Criteria	Frobenius	Spectral	Nuclear	ℓ_1	FPR	FNR	TPR
CoGlasso	CV	0.805	0.857	0.766	0.916	0.008	0.032	0.968
	BIC	2.533	3.177	2.419	19.441	0.737	0.003	0.997
ADMM	CV	0.666	0.799	0.618	1.095	0.056	0.025	0.975
	BIC	1.360	1.649	1.321	9.451	0.630	0.005	0.995
Nodewise	CV	0.798	0.850	0.760	0.910	0.024	0.019	0.981

Setup 2:

The true precision matrix is generated from an Erdös-Rényi random graph with $p * 0.1$ randomly selected edges with partial correlation randomly chosen from a uniform distribution $Unif(0.6, 0.8)$. The results are shown in Table 3.17 - 3.20.

Table 3.17: *Scenario Q:* $p = 400, n = 80, P_{\text{miss}} = 0.1$

Method	Criteria	Frobenius	Spectral	Nuclear	ℓ_1	FPR	FNR	TPR
CoGlasso	CV	0.673	0.761	0.400	0.811	0.005	0.395	0.605
	BIC	0.695	0.762	0.413	0.811	0.000	0.894	0.105
ADMM	CV	0.725	0.852	0.433	0.888	0.506	0.149	0.851
	BIC	8.705	1.919	11.959	11.417	0.998	0.000	1.000
Nodewise	CV	1.340	3.518	0.545	2.614	0.004	0.004	0.996

Table 3.18: *Scenario R:* $p = 400, n = 130, P_{\text{miss}} = 0.3$

Method	Criteria	Frobenius	Spectral	Nuclear	ℓ_1	FPR	FNR	TPR
CoGlasso	CV	0.731	0.824	0.435	0.863	0.003	0.477	0.523
	BIC	0.741	0.824	0.443	0.864	0.000	0.848	0.152
ADMM	CV	0.725	0.853	0.434	0.889	0.525	0.153	0.847
	BIC	7.901	1.919	10.055	11.492	0.973	0.000	1.000
Nodewise	CV	0.762	1.328	0.414	1.150	0.004	0.011	0.989

Table 3.19: *Scenario S:* $p = 400, n = 250, P_{\text{miss}} = 0.5$

Method	Criteria	Frobenius	Spectral	Nuclear	ℓ_1	FPR	FNR	TPR
CoGlasso	CV	0.769	0.858	0.484	0.891	0.000	0.702	0.298
	BIC	0.770	0.858	0.485	0.891	0.000	0.796	0.204
ADMM	CV	0.728	0.855	0.440	0.890	0.514	0.192	0.808
	BIC	6.487	1.919	7.568	9.665	0.830	0.001	0.999
Nodewise	CV	0.714	0.806	0.449	0.840	0.003	0.014	0.986

Table 3.20: *Scenario T:* $p = 400, n = 700, P_{\text{miss}} = 0.7$

Method	Criteria	Frobenius	Spectral	Nuclear	ℓ_1	FPR	FNR	TPR
CoGlasso	CV	0.785	0.874	0.512	0.904	0.000	0.858	0.142
	BIC	7.177	2.781	7.822	12.395	0.809	0.003	0.997
ADMM	CV	0.733	0.859	0.446	0.894	0.400	0.230	0.770
	BIC	5.786	1.919	6.343	9.016	0.818	0.002	0.998
Nodewise	CV	0.760	0.850	0.495	0.880	0.003	0.027	0.973

In the missing data scenario, all the methods perform well when the underlying structure of the precision matrix is a chain graph even in an increasing degree of missingness introduced in the data. However, when we assume the underlying structure of the precision matrix to have Erdös-Rényi graph structure, only nodewise regression performs well. Even with a small proportion of missing data present in the data, false positive rate and the false negative rate are not comparable to the chain graph scenario.

We report some additional simulation results to check the scalings predicted by the theory. Based on Theorem 1, we can show that in the additive error case with sub-Gaussian tails, the elementwise maximum norm should decay at the rate $\mathcal{O}(\sqrt{\log p/n})$. In Figure 3.1 and 3.2, we plotted the elementwise maximum norm error against the original sample size n and the rescaled sample size $n/\log p$ and showed that the curves align in the presence of additive noise and missing data, respectively. We generate a chain structured graph where all the nodes are arranged in a linear chain with each node having degree 2 (except the two ends). We generated the precision matrix with the diagonal entries of $\mathbf{\Theta}^*$ are set equal to 1, and all entries corresponding to links in the chain are set equal to 0.1. We generated the matrix $\mathbf{X}$ as a zero-mean sub-Gaussian with covariance $\mathbf{\Sigma}_X = (\mathbf{\Theta}^*)^{-1}$. For the additive

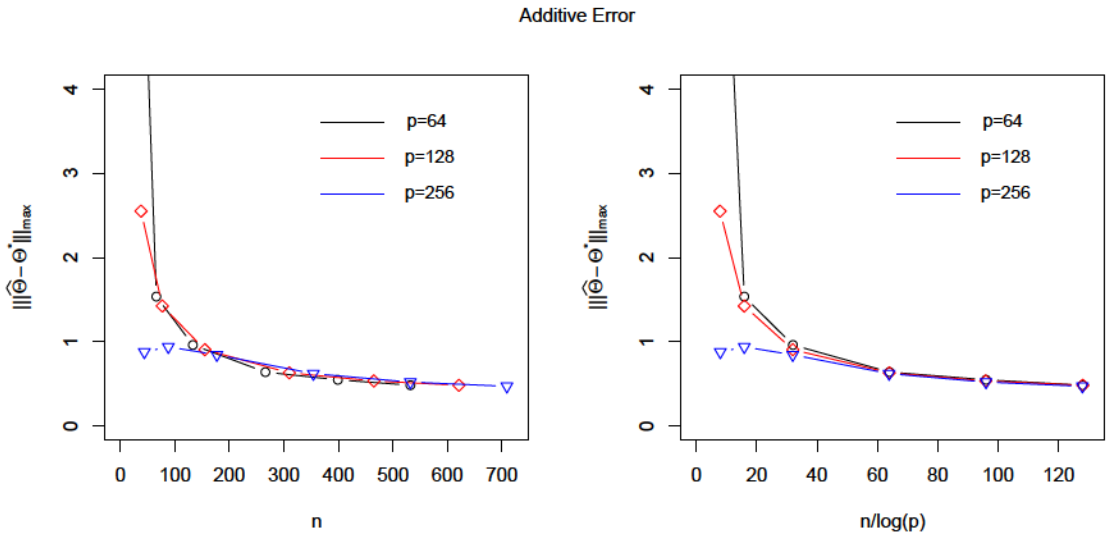

Figure 3.1: *Plots of the error* $\|\|\widehat{\Theta} - \Theta^*\|\|_{\max}$ *against the sample size* n *(left) and rescaled sample size* $n/\log p$ *(right) in the case of a chain structured precision matrix when the error is additive. Each point represents an average of 100 trials.*

noise case, we generated the corrupted matrix $\mathbf{Z} = \mathbf{X} + \mathbf{W}$ with $\Sigma_W = (0.2)^2 I$ or in the missing data case we generated $\mathbf{Z}$ with 20% missing data in each column of $\mathbf{X}$. We see good agreement with the theoretical predictions.

3.7 Real Data Analysis

In this section we present an example of estimating the precision matrix for real gene expression data which was collected to distinguish active tuberculosis (TB) patients from latently infected and healthy individuals. We performed our method on a real gene expression data set obtained from Singhania et al. (2018). Using microarray analysis, Berry et al. (2010) first identified a whole blood 393 transcript signature for active TB. Later a confirmatory analysis conducted by Singhania et al. (2018) found a 373-genes signature of active tuberculosis using RNA-Seq that discriminates active tuberculosis from latently infected and healthy individuals.

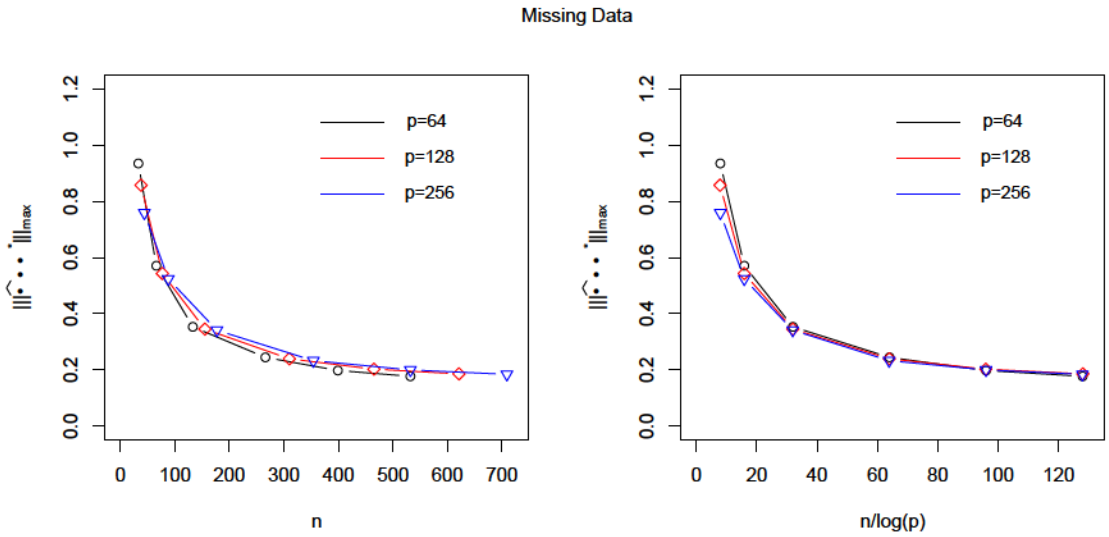

Figure 3.2: *Plots of the error* $\|\|\widehat{\Theta} - \Theta^*\|\|_{\max}$ *against the sample size* n *(left) and rescaled sample size* $n/\log p$ *(right) in the case of a chain structured precision matrix when the error is multiplicative. Each point represents an average of 100 trials.*

The dataset includes gene expression of 54 patients (21 active TB patients, 21 latent TB patients and 12 healthy controls) and contains 14150 genetic markers from a cohort studied in London. The gene expression is preprocessed into log(counts) per million (i.e. log-cpm) and is available in the `R` package `dearseq`. For the sake of demonstrating our method, we randomly selected 200 genetic markers. First we performed a graphical Lasso algorithm using the `R` package `glasso` and estimated the precision matrix. Then we randomly deleted 10% observations from each genetic marker to mimic a missing completely at random scenario and applied the CoGlasso algorithm for missing data. For the sake of our analysis, we scaled the columns of the data so that $(1/n)\sum_{i=1}^{n} x_{ij}^2 = 1$ for every column.

Since this is an illustration and performed on randomly selected genetic markers, the partial correlations obtained could not be validated with the existing biological information. However, it is shown to demonstrate the methodology established in this Chapter. The plot on the top in Figure 3.3 appear to be sparser than the one created with no missing data (bottom). We performed a 5-fold cross-validation to tune the regularization parameter in

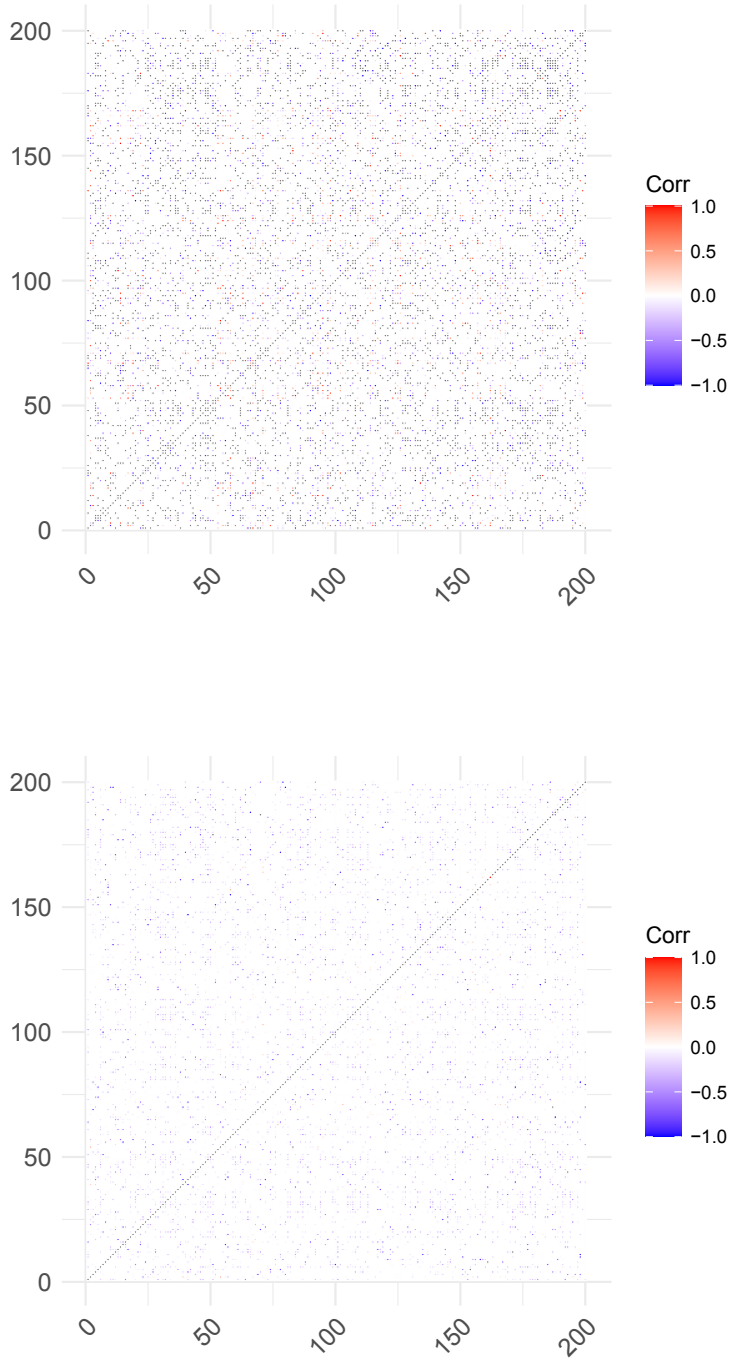

Figure 3.3: *Plot of the partial correlation matrix for complete data for 200 randomly selected genetic markers using graphical Lasso method (top). Plot of partial correlation matrix with 10% missing data for 200 randomly selected genetic markers using CoGlasso method (bottom).*

both the cases. We used used a sequence of equally spaced values for the tuning parameter in the logarithmic scale from $[-2, 2]$ to perform the cross-validation. For the clean graphical Lasso the regularization parameter was chosen to be 0.01. For the missing data case, the regularization parameter was chosen to be 0.1 based on a 5-fold cross-validation, therefore, resulting into a sparser graph.

3.8 Discussion and Conclusion

In this chapter, we have studied the estimation of precision matrix in the presence of additive and multiplicative noise in a high-dimensional setting. We have derived theoretical deviation bounds for the estimated precision matrix from the truth in elementwise maximum norm for two types of tail deviation conditions in two types of measurement error settings. In terms of the upper bound on statistical convergence rates, our method shares many theoretical properties with classical graphical Lasso in the clean case and provides specific rates in element-wise ℓ_∞-norm for different distributional assumptions on the signal and noise variables under different types formulation of measurement errors. It is easy to implement and guarantees a convex solution to the problem.

We have performed simulation studies for different types of measurement error introduced in the model. In terms of additive error, all the methods performed comparably. Specifically, CoGlasso tends to give smaller false positive rate compared to the ADMM method when the partial correlation is stronger in the additive case. When the partial correlation is weaker, the performance of all the methods worsened, particularly, the false negative rate of CoGlasso performed poorly. Similar trends were visible in terms of the relative Frobenius, spectral, nuclear and ℓ_1 norm. With stronger to moderate partial correlation among the nodes in the chain graph, nodewise regression performed similarly to CoGlasso.

In case of missing data, in a simpler covariance structure such as a chain graph, CoGlasso performs comparatively better than the non-convex approach, especially in terms of false positive rates. With more complex form of covariance structure such as Erdös-Rényi graph, both the methods tend to perform poorly with increasing missingness introduced into the data. However, even with a deteriorating false negative rate, CoGlasso provided better false positive rates compared to the ADMM method. The nodewise regression tend to outperform both the methods.

In terms of tuning the regularization parameter, cross-validation method performed better than the BIC criterion. The convergence of the ADMM algorithm depended on the assumption of the side-constraint and added difficulty to the convergence of the estimator. Compared to that, CoGlasso did not depend on any prior information and therefore, was straightforward to implement.

3.9 Technical Details

Lemma 5. *Let $Z = (Z_1, Z_2, \ldots, Z_n)^\top$ where Z_i's are independent sub-Gaussian random variables with sub-Gaussian parameter at most τ^2. If Z is a sub-Gaussian random vector then $Z - \mathbb{E}(Z)$ is also sub-Gaussian; the weighted sums of the centered Z_i's are also sub-Gaussian and satisfy the probabilistic bound*

$$\Pr\left\{ \mid \mathbf{v}^\top (Z - \mathbb{E}(Z)) \mid > t \right\} \leq 2 \exp\left(-\frac{ct^2}{\tau^2 \|\mathbf{v}\|_2^2} \right) \qquad \forall t > 0 \tag{3.36}$$

where c is an universal constant. If Z is centered and the weights $\|\mathbf{v}\|_2^2 = 1$, then the bound can be simplified to

$$\Pr\Big(\Big| \sum_{i=1}^{n} v_i \left(Z_i - \mathbb{E}(Z_i) \right) \Big| \geq t \Big) \leq 2 \exp\left(-\frac{ct^2}{\tau^2} \right)$$

Proof. By Lemma 2.6.8 (Vershynin, 2018) and general Hoeffding inequality from Theorem 2.6.3 Vershynin (2018). □

Lemma 6. *(Datta and Zou, 2017) Let $Z_i = (X_i, Y_i)^\top$ for $i = 1, \ldots, n$ denote independent and identically distributed two-dimensional vectors with zero mean, covariance $\mathbf{\Sigma}_{XY} = (\sigma_{XY})_{2\times 2}$ and sub-Gaussian parameter τ^2. Then there exist absolute constants C and c such that, we have*

$$\Pr\left\{\frac{1}{n}\Big|\sum_{i=1}^{n} v_i(X_iY_i - \sigma_{XY})\Big| \geq \delta\right\} \leq \frac{1}{C}\exp\left[-cn\min\Big(\frac{\delta^2}{\tau^4\|\mathbf{v}\|_2^2}, \frac{\delta}{\tau^2\|\mathbf{v}\|_\infty}\Big)\right]. \tag{3.37}$$

When δ is small enough, that is, when $\delta \leq \tau^2\|\mathbf{v}\|_2^2/\|\mathbf{v}\|_\infty$, we can simplify the probabilistic bound as

$$\Pr\left\{\frac{1}{n}\Big|\sum_{i=1}^{n} v_i(X_iY_i - \sigma_{XY})\Big| \geq \delta\right\} \leq \frac{1}{C}\exp\left[-\frac{cn\delta^2}{\tau^4\|\mathbf{v}\|_2^2}\right].$$

Proof. We have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} v_i(X_iY_i - \sigma_{XY}) \\
=&\frac{1}{4n}\sum_{i=1}^{n} v_i\left\{\Big(X_i + Y_i\Big)^2 - \Big(\sigma_{XX} + \sigma_{YY} + 2\sigma_{XY}\Big)\right\} \\
&\qquad - \frac{1}{4n}\sum_{i=1}^{n} v_i\left\{\Big(X_i - Y_i\Big)^2 - \Big(\sigma_{XX} + \sigma_{YY} - 2\sigma_{XY}\Big)\right\} \\
=&\frac{1}{2n}\sum_{i=1}^{n} v_i\left\{\Big(\frac{1}{\sqrt{2}}\Big(X_i + Y_i\Big)\Big)^2 - \frac{1}{2}\Big(\sigma_{XX} + \sigma_{YY} + 2\sigma_{XY}\Big)\right\} \\
&\qquad - \frac{1}{2n}\sum_{i=1}^{n} v_i\left\{\Big(\frac{1}{\sqrt{2}}\Big(X_i - Y_i\Big)\Big)^2 - \frac{1}{2}\Big(\sigma_{XX} + \sigma_{YY} - 2\sigma_{XY}\Big)\right\} \\
=&\frac{1}{2n}\sum_{i=1}^{n} v_i\left\{\Big(a_1^\top Z_i\Big)^2 - \mathbb{E}\left\{\Big(a_1^\top Z_i\Big)^2\right\}\right\} - \frac{1}{2n}\sum_{i=1}^{n} v_i\left\{\Big(a_2^\top Z_i\Big)^2 - \mathbb{E}\left\{\Big(a_2^\top Z_i\Big)^2\right\}\right\}.
\end{aligned}$$

Here, $a_1 = (1/\sqrt{2}, 1/\sqrt{2})^\top$ and $a_2 = (1/\sqrt{2}, -1/\sqrt{2})^\top$. So, $(a_1^\top Z_i)^2 = (1/2)X_i^2 + (1/2)Y_i^2 + X_iY_i$ and $\mathbb{E}[(a_1^\top Z_i)^2] = (1/2)(\sigma_{XX} + \sigma_{YY} + 2\sigma_{XY})$. Similarly, $(a_2^\top Z_i)^2 = (1/2)X_i^2 + (1/2)Y_i^2 - X_iY_i$ and $\mathbb{E}[(a_2^\top Z_i)^2] = (1/2)(\sigma_{XX} + \sigma_{YY} - 2\sigma_{XY})$. As $\|a_k\|_2^2 = 1$, $a_k^\top Z_i$ is sub-Gaussian with parameter at most τ^2 for $k = 1, 2$, followed by Lemma 5. Next, we use the relationship between sub-Gaussian and sub-exponential random variables from Lemma 5.14 of Vershynin (2010), which states that a random variable Z is sub-Gaussian if and only if Z^2 is sub-

exponential. We also use the Remark 5.18 from Vershynin (2010) which states that if Z is sub-exponential, then so is $Z - \mathbb{E}(Z)$. Hence, we see that for $k = 1, 2$, $(a_k^\top Z_i)^2 - \mathbb{E}\{(a_k^\top Z_i)^2\}$ is sub-exponential with parameter at most $c\tau^2$ where c is an absolute constant. Therefore, $v_i\{(a_1^\top Z_i)^2 - \mathbb{E}\{(a_1^\top Z_i)^2\}\}$ is sub-exponential with parameter at most $c\tau^2\|\mathbf{v}\|_\infty$. Since we have a linear combination of sub-exponential random variables, we can directly apply Proposition 5.16 and Corollary 5.17 from Vershynin (2010) which provide a tail bound for the sums of independent centered sub-exponential random variables. We have

$$
\begin{aligned}
&\Pr\left[\frac{1}{n}\Big|\sum_{i=1}^{n} v_i(X_iY_i - \sigma_{XY})\Big| \geq \delta\right] \\
&\leq \Pr\left[\frac{1}{2n}\Big|\sum_{i=1}^{n} v_i\left\{\left(a_1^\top Z_i\right)^2 - \mathbb{E}\left\{\left(a_1^\top Z_i\right)^2\right\}\right\}\Big| + \frac{1}{2n}\Big|\sum_{i=1}^{n} v_i\left\{\left(a_2^\top Z_i\right)^2 - \mathbb{E}\left\{\left(a_2^\top Z_i\right)^2\right\}\right\}\Big| \geq \delta\right] \\
&\leq \frac{1}{C}\exp\left[-cn\min\Big(\frac{\delta^2}{\tau^4\|\mathbf{v}\|_2^2}, \frac{\delta}{\tau^2\|\mathbf{v}\|_\infty}\Big)\right].
\end{aligned}
$$

Now, when δ is sufficiently small, that is, when

$$
\frac{\delta^2}{\tau^4\|\mathbf{v}\|_2^2} \leq \frac{\delta}{\tau^2\|\mathbf{v}\|_\infty} \quad \text{implying} \quad \delta \leq \frac{\tau^2\|\mathbf{v}\|_2^2}{\|\mathbf{v}\|_\infty},
$$

the simplified probabilistic bound would be

$$
\Pr\left\{\frac{1}{n}\Big|\sum_{i=1}^{n} v_i(X_iY_i - \sigma_{XY})\Big| \geq \delta\right\} \leq \frac{1}{C}\exp\left[-\frac{cn\delta^2}{\tau^4\|\mathbf{v}\|_2^2}\right].
$$

□

Lemma 7. *Let $\mathbf{\Sigma}^{(1)}$ and $\mathbf{\Sigma}^{(2)}$ be two $p \times p$ dimensional (random) matrices. The tail bound for the elementwise-max norm of the deviation, for some $\delta > 0$, can be upper bounded in terms of their elementwise absolute deviation, that is*

$$
\Pr(|||\mathbf{\Sigma}^{(1)} - \mathbf{\Sigma}^{(2)}|||_{\max} \geq \delta) \leq p^2 \max_{i,j} \Pr\Big(\mid \mathbf{\Sigma}_{ij}^{(1)} - \mathbf{\Sigma}_{ij}^{(2)} \mid \geq \delta\Big).
$$

Proof. We can write the elementwise-max norm as,

$$\begin{aligned}
\Pr(|||\boldsymbol{\Sigma}^{(1)} - \boldsymbol{\Sigma}^{(2)}|||_{\max} \geq \delta) &\leq \Pr\Big(\max_{i,j} \mid \boldsymbol{\Sigma}_{ij}^{(1)} - \boldsymbol{\Sigma}_{ij}^{(2)} \mid \geq \delta \Big) \\
&\leq \Pr\Big(\cup_{i,j} \mid \boldsymbol{\Sigma}_{ij}^{(1)} - \boldsymbol{\Sigma}_{ij}^{(2)} \mid \geq \delta \Big) \\
&\overset{\text{(i)}}{\leq} \sum_{i=1}^{p} \sum_{j=1}^{p} \Pr\Big(\mid \boldsymbol{\Sigma}_{ij}^{(1)} - \boldsymbol{\Sigma}_{ij}^{(2)} \mid \geq \delta \Big) \\
&\leq p^2 \Pr\Big(\mid \boldsymbol{\Sigma}_{ij}^{(1)} - \boldsymbol{\Sigma}_{ij}^{(2)} \mid \geq \delta \Big).
\end{aligned}$$

Inequality (i) is due to Boole's inequality. □

Proof of Lemma 1

Proof. Recall that, for an additive measurement error model, we assume the observed matrix is $\mathbf{Z} = \mathbf{X} + \mathbf{W}$. Let $\boldsymbol{\Sigma}_W = (\sigma_{W,jk})_{p\times p}$, the covariance matrix of the measurement error $\mathbf{W}$ for the additive model. Given $\mathbf{S}$ as the sample covariance matrix for the data without any corruption, we have

$$\begin{aligned}
\widehat{\boldsymbol{\Sigma}}_{\text{addi}} - \mathbf{S} =& \frac{1}{n}\mathbf{Z}^\top\mathbf{Z} - \boldsymbol{\Sigma}_W - \mathbf{S} \\
=& \frac{1}{n}(\mathbf{X}+\mathbf{W})^\top(\mathbf{X}+\mathbf{W})^\top - \boldsymbol{\Sigma}_W - \frac{1}{n}\mathbf{X}^\top\mathbf{X} \\
=& \frac{1}{n}\mathbf{X}^\top\mathbf{X} + \frac{1}{n}\mathbf{X}^\top\mathbf{W} + \frac{1}{n}\mathbf{W}^\top\mathbf{X} + \frac{1}{n}\mathbf{W}^\top\mathbf{W} - \boldsymbol{\Sigma}_W - \frac{1}{n}\mathbf{X}^\top\mathbf{X} \\
=& \frac{1}{n}\mathbf{X}^\top\mathbf{W} + \frac{1}{n}\mathbf{W}^\top\mathbf{X} + \frac{1}{n}\mathbf{W}^\top\mathbf{W} - \boldsymbol{\Sigma}_W. \qquad (3.38)
\end{aligned}$$

Let X_j and W_k be the jth and the kth column of $\mathbf{X}$ and $\mathbf{W}$, respectively, where $j, k = 1, \ldots, p$. Therefore by (3.38), we have,

$$\widehat{\boldsymbol{\Sigma}}_{\text{addi},jk} - \mathbf{S}_{jk} = \frac{1}{n}W_j^\top X_k + \frac{1}{n}W_k^\top X_j + \frac{1}{n}W_j^\top W_k - \sigma_{W,jk}.$$

In order to bound this, we would now bound each term in (3.38) separately. We have that $W_j = (W_{1j}, W_{2j}, \ldots, W_{nj})^\top$ and $X_k = (X_{1k}, X_{2k}, \ldots, X_{nk})^\top$ are independent, zero-centered sub-Gaussian random variables where each entry has a parameter at most σ_W^2 and σ_X^2, respectively, for all j and k. We also assumed that X_k and W_j are independent.

For any random variable Y and $\alpha > 0$, let us define the quasi-norm

$$\|Y\|_{\psi_\alpha} = \inf \left\{ t > 0 : \mathbb{E}\left[\exp\left(\frac{|Y|^\alpha}{t^\alpha} \right) \leq 2 \right] \right\}.$$

Here we define $\inf \varnothing = \infty$. This is a generalization of sub-Gaussianity and sub-exponentiality. The random variable with finite exponential Orlicz norm $\|\cdot\|_{\psi_\alpha}$ corresponds to the α-sub-exponential tail decay family which can be defined as

$$\Pr(|Y| \geq t) \leq \frac{1}{C} \exp(-ct^\alpha), \qquad \forall t \geq 0 \tag{3.39}$$

where C and c are constants. We have two special cases of Orlicz norms: $\alpha = 1$ corresponds to the family of sub-exponential distributions and $\alpha = 2$ corresponds to the family of sub-Gaussian distributions.

According to the Lemma 2.7.7 in Vershynin (2018) if W_j and X_k are sub-Gaussian random variables then their product would be a sub-exponential random variable. Therefore, for two generic random variables X and W the following inequality in terms of the Orlicz norm would hold

$$\|XW\|_{\psi_1} \leq \|X\|_{\psi_2} \|W\|_{\psi_2}. \tag{3.40}$$

We can also observe from Lemma 5.5 of Vershynin (2010) that there exist universal constants m_1, m_2, M_1 and M_2 such that $m_1\|X\|_{\psi_2}^2 \leq \sigma_X^2 \leq M_1\|X\|_{\psi_2}^2$ and $m_2\|W\|_{\psi_2}^2 \leq \sigma_W^2 \leq M_2\|W\|_{\psi_2}^2$ hold. Since the inner product of X_k and W_j is sub-exponential in the first term in (3.38), therefore, following Corollary 5.17 from Vershynin (2010), the sum of independent, cen-

tered sub-exponential random variables have the following probabilistic bound:

$$
\begin{aligned}
&\Pr\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} W_{ij}X_{ik}\Big| \geq \delta\Big) \\
&\leq \frac{1}{C}\exp\left[-cn\min\left(\frac{\delta^2}{\{\max_i \|W_{ij}X_{ik}\|_{\psi_1}\}^2}, \frac{\delta}{\max_i \|W_{ij}X_{ik}\|_{\psi_1}}\right)\right] \\
&\leq \frac{1}{C}\exp\left[-cn\min\left(\frac{\delta^2}{\{\max_i \|X_{ik}\|_{\psi_2}\|W_{ij}\|_{\psi_2}\}^2}, \frac{\delta}{\max_i \|X_{ik}\|_{\psi_2}\|W_{ij}\|_{\psi_2}}\right)\right] \\
&= \frac{1}{C}\exp\left[-cn\min\left(\frac{\delta^2}{\max_i \|X_{ik}\|^2_{\psi_2}\|W_{ij}\|^2_{\psi_2}}, \frac{\delta}{\max_i \|X_{ik}\|_{\psi_2}\|W_{ij}\|_{\psi_2}}\right)\right] \\
&\leq \frac{1}{C}\exp\left[-cn\min\left(\frac{\delta^2}{\sigma_X^2\sigma_W^2}, \frac{\delta}{\sigma_X\sigma_W}\right)\right]
\end{aligned}
$$

where c and C are universal constants. When $\delta \leq \sigma_X\sigma_W$, the bound can be simplified to

$$
\Pr\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} W_{ij}X_{ik}\Big| \geq \delta\Big) \leq \frac{1}{C}\exp\Big[-\frac{cn\delta^2}{\sigma_X^2\sigma_W^2}\Big].
$$

A similar bound can be formed for the second term in (3.38) as

$$
\Pr\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} W_{ik}X_{ij}\Big| \geq \delta\Big) \leq \frac{1}{C}\exp\left[-cn\min\left(\frac{\delta^2}{\sigma_X^2\sigma_W^2}, \frac{\delta}{\sigma_X\sigma_W}\right)\right].
$$

When $\delta \leq \sigma_X\sigma_W$, the bound can be simplified to

$$
\Pr\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} W_{ik}X_{ij}\Big| \geq \delta\Big) \leq \frac{1}{C}\exp\Big[-\frac{cn\delta^2}{\sigma_X^2\sigma_W^2}\Big].
$$

Next, to bound the third term in (3.38) let us inspect the correlated sub-Gaussian sequences, $Z_i = (W_{ij}, W_{ik})^\top$. Here Z_i's are independent and identically distributed vectors with zero mean and covariance $\mathbf{\Sigma}_W = (\sigma_{W,jk})_{p\times p}$ and sub-Gaussian parameter σ_W^2. We can directly apply Lemma 6 to bound $(W_{ij}W_{ik})/n - \sigma_{W,jk}$ for $v_i = 1$

$$
\Pr\Big(\frac{1}{n}\Big|\sum_{i=1}^{n}\Big(W_{ij}W_{ik} - \sigma_{W,jk}\Big)\Big| \geq \delta\Big) \leq \frac{1}{C}\exp\left[-cn\min\Big(\frac{\delta^2}{\sigma_W^4}, \frac{\delta}{\sigma_W^2}\Big)\right].
$$

When $\delta \leq \sigma_W^2$, the bound can be simplified to

$$\Pr\Big(\frac{1}{n}\Big|\sum_{i=1}^{n}(W_{ij}W_{ik} - \sigma_{W,jk})\Big| \geq \delta\Big) \leq \frac{1}{C}\exp\Big[-\frac{cn\delta^2}{\sigma_W^4}\Big].$$

To upper bound for the elementwise max norm we can apply Lemma 7. Hence, putting these three pieces together, we see that $\widehat{\mathbf{\Sigma}}_{\text{addi}}$ and $\mathbf{S}$ satisfy the closeness condition in (3.10) with $\xi = \max(\sigma_W^4, \sigma_W^2, \sigma_X^2\sigma_W^2, \sigma_X\sigma_W)$. For a sufficiently small δ, specifically for $\delta \leq \min(\sigma_X\sigma_W, \sigma_W^2)$, $\widehat{\mathbf{\Sigma}}_{\text{addi}}$ and $\mathbf{S}$ satisfy the closeness condition in (3.10) with $\xi = \max(\sigma_W^4, \sigma_X^2\sigma_W^2)$. □

Proof of Lemma 2

Proof. Recall that, for an additive measurement error model, we assume the observed matrix is $\mathbf{Z} = \mathbf{X} + \mathbf{W}$. Let $\mathbf{\Sigma}_{jk} = (\sigma_{X,jk})_{p\times p}$ and $\mathbf{\Sigma}_W = (\sigma_{W,jk})_{p\times p}$ be the covariance matrices of $\mathbf{X}$ and the measurement error $\mathbf{W}$, for the additive model, respectively. Let X_j and W_k be the random variables corresponding to the jth and the kth column, respectively, where $j, k = 1, \ldots, p$. Here, we assume that each column X_j and W_k are each independently and identically distributed with bounded moments

$$\mathbb{E}\left[X_j^{4m}\right] \leq K_{m,X} \qquad \text{and} \qquad \mathbb{E}\left[W_k^{4m}\right] \leq K_{m,W}.$$

Here m is a positive integer and $K_{m,X}, K_{m,W} \in \mathbb{R}^+$. We also assumed that X_j and W_k for any j and k are independent. Given $\mathbf{S}$ as the sample covariance matrix for the data without any corruption, we have

$$\widehat{\mathbf{\Sigma}}_{\text{addi}} - \mathbf{S} = \frac{1}{n}\mathbf{X}^\top\mathbf{W} + \frac{1}{n}\mathbf{W}^\top\mathbf{X} + \frac{1}{n}\mathbf{W}^\top\mathbf{W} - \mathbf{\Sigma}_W.$$

The jkth element of this matrix can be written as

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_{\text{addi},jk} - \mathbf{S}_{jk} &= \frac{1}{n}\sum_{i=1}^{n} X_{ij}W_{ik} + \frac{1}{n}\sum_{i=1}^{n} W_{ij}X_{ik} + \frac{1}{n}\sum_{i=1}^{n} W_{ij}W_{ik} - \sigma_{W,jk} \\ &= \sum_{i=1}^{n}\left\{\frac{1}{n}X_{ij}W_{ik} + \frac{1}{n}W_{ij}X_{ik} + \frac{1}{n}W_{ij}W_{ik} - \frac{1}{n}\sigma_{W,jk}\right\}\end{aligned}$$

Let us define the random variable $T_{jk}^{(i)}$ as

$$T_{jk}^{(i)} = \frac{1}{n}X_{ij}W_{ik} + \frac{1}{n}W_{ij}X_{ik} + \frac{1}{n}W_{ij}W_{ik} - \frac{1}{n}\sigma_{W,jk}$$

and note that they have mean zero. By applying Chebyshev's inequality, we obtain,

$$\begin{aligned}\Pr\left[\left|\sum_{i=1}^{n} T_{jk}^{(i)}\right| > \delta\right] &= \Pr\left[\left(\sum_{i=1}^{n} T_{jk}^{(i)}\right)^{2m} > \delta^{2m}\right] \\ &\leq \frac{\mathbb{E}\left[\left(\sum_{i=1}^{n} T_{jk}^{(i)}\right)^{2m}\right]}{\delta^{2m}}. \end{aligned} \tag{3.41}$$

Now, applying Rosenthal's inequality (Rosenthal, 1970) to obtain that there exists a constant C_m, depending only on m, such that

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i=1}^{n} T_{jk}^{(i)}\right)^{2m}\right] &\leq C_m \max\left(\sum_{i=1}^{n}\mathbb{E}[(T_{jk}^{(i)})^{2m}], \left(\sum_{i=1}^{n}\mathbb{E}[(T_{jk}^{(i)})^{2}]\right)^{m}\right) \\ &\leq C_m\left(\sum_{i=1}^{n}\mathbb{E}[(T_{jk}^{(i)})^{2m}] + \left(\sum_{i=1}^{n}\mathbb{E}[(T_{jk}^{(i)})^{2}]\right)^{m}\right). \end{aligned} \tag{3.42}$$

Turning to each individual expectation, we have

$$\begin{aligned}&\mathbb{E}[(T_{jk}^{(i)})^{2m}] \\ &= \mathbb{E}\left[\left(\frac{1}{n}X_{ij}W_{ik} + \frac{1}{n}W_{ij}X_{ik} + \frac{1}{n}W_{ij}W_{ik} - \frac{1}{n}\sigma_{W,jk}\right)^{2m}\right]\end{aligned}$$

$$
\begin{aligned}
&\overset{\text{(i)}}{\leq} 2^{4m}\left[\mathbb{E}\left\{\left(\frac{1}{n}X_{ij}W_{ik}\right)^{2m}\right\}+\mathbb{E}\left\{\left(\frac{1}{n}W_{ij}X_{ik}\right)^{2m}\right\}+\mathbb{E}\left\{\left(\frac{1}{n}W_{ij}W_{ik}\right)^{2m}\right\}\right.\\
&\qquad\qquad\left.+\frac{1}{n^{2m}}\sigma_{W,jk}^{2m}\right]\\
&\overset{\text{(ii)}}{=} 2^{4m}\left[\frac{1}{n^{2m}}\mathbb{E}[X_{ij}^{2m}]\mathbb{E}[W_{ik}^{2m}]+\frac{1}{n^{2m}}\mathbb{E}[X_{ik}^{2m}]\mathbb{E}[W_{ij}^{2m}]+\frac{1}{n^{2m}}\mathbb{E}\left\{(W_{ij}W_{ik})^{2m}\right\}+\frac{1}{n^{2m}}\sigma_{W,jk}^{2m}\right]\\
&\overset{\text{(iii)}}{\leq} \frac{2^{4m}}{n^{2m}}\left[\left\{1+\mathbb{E}(X_{ij}^{4m})\right\}\left\{1+\mathbb{E}(W_{ik}^{4m})\right\}+\left\{1+\mathbb{E}(X_{ik}^{4m})\right\}\left\{1+\mathbb{E}(W_{ij}^{4m})\right\}\right.\\
&\qquad\left.+\sqrt{\mathbb{E}(W_{ij}^{4m})\mathbb{E}(W_{ik}^{4m})}+\sigma_{W,jk}^{2m}\right]\\
&\overset{\text{(iv)}}{\leq} \frac{2^{4m}}{n^{2m}}\left[(1+K_{m,X})(1+K_{m,W})+(1+K_{m,X})(1+K_{m,W})+K_{m,W}+\sigma_{W,jk}^{2m}\right]\\
&= \frac{2^{4m}}{n^{2m}}\left[2(1+K_{m,X})(1+K_{m,W})+K_{m,W}+\sigma_{W,jk}^{2m}\right]
\end{aligned}
$$

where inequality (i) follows because of the relationship

$$(\sum_{i=1}^{k}a_i)^n \leq k^{n-1}(\sum_{i=1}^{k}a_i^n) \leq k^n(\sum_{i=1}^{k}a_i^n).$$

Specifically since

$$(a+b+c+d)^{2m} \leq 2^{4m-2}\left\{a^{2m}+b^{2m}+c^{2m}+d^{2m}\right\} \leq 2^{4m}\left\{a^{2m}+b^{2m}+c^{2m}+d^{2m}\right\}.$$

The first and the second term in equality (ii) is due to independence between X and W. The first and the second term in inequality (iii) follows from the relationship $\mathbb{E}(|X|^k) \leq 1+\mathbb{E}(|X|^n)$ for some positive integer k and n where $k<n$ for a generic random variable X. The third term in inequality (iii) follows from the Cauchy-Schwartz inequality. Inequality (iv) follows from the assumed moment bound on $\mathbb{E}[W_j^{4m}]$ and $\mathbb{E}[X_j^{4m}]$. Now, for $m=1$, we have

$$\mathbb{E}[(T_{jk}^{(i)})^2] \lesssim \frac{2^4}{n^2}\left[3+\sigma_{W,jk}^2\right]$$

and hence

$$\begin{aligned}\left(\sum_{i=1}^{n}\mathbb{E}[(T_{jk}^{(i)})^2]\right)^m &\lesssim \left[\frac{2^4 n}{n^2}\left\{3+\sigma_{W,jk}^2\right\}\right]^m \\ &\overset{\text{(i)}}{\leq} \frac{2^m 2^{4m} n^m}{n^{2m}}\left[3^m+\sigma_{W,jk}^{2m}\right] \\ &= \frac{2^{5m} n^m}{n^{2m}}\left[3^m+\sigma_{W,jk}^{2m}\right]\end{aligned}$$

where inequality (i) uses the relationship $(a+b)^m \leq 2^m(a^m+b^m)$. Combined with the earlier bound (3.42)

$$\begin{aligned}&\mathbb{E}\left[\left(\sum_{i=1}^{n}T_{jk}^{(i)}\right)^{2m}\right] \\ &\leq C_m\left[\frac{2^{4m}n}{n^{2m}}\left\{2(1+K_{m,X})(1+K_{m,W})+K_{m,W}+\sigma_{W,jk}^{2m}\right\}\right. \\ &\qquad\qquad\left.+\frac{2^{5m}n^m}{n^{2m}}\left\{3^m+\sigma_{W,jk}^{2m}\right\}\right] \\ &\overset{\text{(i)}}{\leq} \frac{C_m 2^{4m} n^m}{n^{2m}}\left[2(1+K_{m,X})(1+K_{m,W})+K_{m,W}+\sigma_{W,jk}^{2m}+2^m 3^m+2^m\sigma_{W,jk}^{2m}\right] \\ &\leq \frac{C_m 2^{4m}}{n^m}\left[2^m 3^m+2(1+K_{m,X})(1+K_{m,W})+K_{m,W}+\sigma_{W,jk}^{2m}(1+2^m)\right]\end{aligned}$$

Inequality (i) holds since $n \leq n^m$. Substituting back to the Chebyshev's inequality in (3.41) yields the tail bound

$$\begin{aligned}&\Pr\left[\left|\sum_{i=1}^{n}T_{jk}^{(i)}\right|>\delta\right] \\ &\leq \frac{C_m 2^{4m}}{n^m\delta^{2m}}\left[2^m 3^m+2(1+K_{m,X})(1+K_{m,W})+K_{m,W}+\sigma_{W,jk}^{2m}(1+2^m)\right] \\ &\overset{\text{(i)}}{\leq} \frac{C_m 2^{4m}}{n^m\delta^{2m}}\left[6^m+2(1+K_{m,X})(1+K_{m,W})+K_{m,W}+(1+2^m)(\max_i \sigma_{W,ii}^{2m})\right]\end{aligned}$$

$$= \frac{C_m 2^{4m}}{n^m \delta^{2m}} \left[6^m + 2(1 + K_{m,X})(1 + K_{m,W}) + K_{m,W} + (1 + 2^m) \right]$$

where in inequality (i) the elementwise variance of W is replaced by the maximum elementwise variance $\max_i \sigma_{W,ii}$ of $\mathbf{\Sigma}_W$, which simplifies to 1 since the data are normalized. To upper bound for the elementwise max norm we can apply Lemma 7. Hence the claim is established. □

Proof of Lemma 3

Proof. Recall that, for a multiplicative measurement error model, we assume the observed matrix is $\mathbf{Z} = \mathbf{X} \odot \mathbf{W}$ where $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_n)^\top$ is a matrix of multiplicative errors where each row $\mathbf{w}_i \in \mathbb{R}^p$ of $\mathbf{W}$ is independent and identically distributed. Let $\mathbb{E}(W) = \mu_W \in \mathbb{R}^p$ be the known mean and $\mathbf{\Sigma}_W = (\sigma_{W,jk})_{p \times p}$ be the known population covariance matrix of the measurement error $\mathbf{W}$ for the multiplicative model. Given $\mathbf{S}$ as the sample covariance matrix for the data without any corruption, we have

$$\begin{aligned} \widehat{\mathbf{\Sigma}}_{\text{mult}} - \mathbf{S} &= \frac{1}{n} \mathbf{Z}^\top \mathbf{Z} \oslash (\mathbf{\Sigma}_W + \mu_W \mu_W^\top) - \frac{1}{n} \mathbf{X}^\top \mathbf{X} \\ &= \frac{1}{n} (\mathbf{X} \odot \mathbf{W})^\top (\mathbf{X} \odot \mathbf{W}) \oslash (\mathbf{\Sigma}_W + \mu_W \mu_W^\top) - \frac{1}{n} \mathbf{X}^\top \mathbf{X}. \end{aligned} \tag{3.43}$$

Let W_j and X_k be the jth and the kth column of $\mathbf{W}$ and $\mathbf{X}$ each with n elements, respectively, where $j, k = 1, \ldots, p$. We have that $W_j = (W_{1j}, W_{2j}, \ldots, W_{nj})^\top$ and $X_k = (X_{1k}, X_{2k}, \ldots, X_{nk})^\top$ are independent, sub-Gaussian random variables where each entry has a parameter at most σ_W^2 and σ_X^2, respectively, for all j and k. We also assumed that W_j and X_k are independent. Followed by (4.40), we have,

$$\widehat{\mathbf{\Sigma}}_{\text{mult},jk} - \mathbf{S}_{jk} = \frac{1}{n} \sum_{i=1}^{n} \frac{X_{ij} W_{ij} X_{ik} W_{ik}}{\mu_j \mu_k + \sigma_{W,jk}} - \frac{1}{n} \sum_{i=1}^{n} X_{ij} X_{ik}.$$

Since the errors are multiplicative, in order to have all the Z_{ij}'s to be close to the respective X_{ij}'s, we need to upper bound both X_{ij} and W_{ij}. To have a meaningful expression for $\widehat{\mathbf{\Sigma}}_{\text{mult}}$, we also need to impose a positive lower bound for the entries μ_W and $\mathbf{\Sigma}_W + \mu_W \mu_W^\top$. We impose the following regularity conditions for the multiplicative setup:

$$\begin{aligned} \Pr\Big(\max_{i,j} \mid X_{ij} \mid < \infty \Big) &= 1, \\ \min_{j,k} \mathbb{E}(W_j W_k^\top) &= m_{\min} > 0, \\ \min_j \mu_j &= \mu_{\min} > 0, \\ \max_j \mu_j &= \mu_{\max} < \infty \end{aligned} \tag{3.44}$$

where $m_{\min}$, $\mu_{\max}$ and $\mu_{\min}$ are constants. Under these regularity conditions, we have by triangle inequality

$$\begin{aligned} \Big| \widehat{\mathbf{\Sigma}}_{\text{mult},jk} - \mathbf{S}_{jk} \Big| &\leq \Big| \frac{1}{n} \sum_{i=1}^n \frac{X_{ij} W_{ij} X_{ik} W_{ik}}{\mu_j \mu_k + \sigma_{W,jk}} - \frac{1}{n} \sum_{i=1}^n X_{ij} X_{ik} \Big| \\ &\leq \frac{1}{\min_{j,k} \mathbb{E}(W_j W_k^\top)} \Big| \frac{1}{n} \sum_{i=1}^n X_{ij} W_{ij} X_{ik} W_{ik} \Big| + \frac{1}{n} \Big| \sum_{i=1}^n X_{ij} X_{ik} \Big| \\ &\leq \frac{1}{m_{\min}} \Big| \frac{1}{n} \sum_{i=1}^n X_{ij} W_{ij} X_{ik} W_{ik} \Big| + \frac{1}{n} \Big| \sum_{i=1}^n X_{ij} X_{ik} \Big|. \end{aligned} \tag{3.45}$$

In order to bound this, we would now bound each term in (3.45) separately. For brevity, we denote the two terms on the right hand side of (3.45) by T_1 and T_2, respectively. For the first term, we have a product of four sub-Gaussian random variables, namely, X_j, X_k, W_j and W_k. To find the distribution of this product we can apply the result introduced in Lemma A.1 in Götze et al. (2021) which states that for a random vector $X = (X_1, \ldots, X_k)$ with marginals having α-sub-exponential tails the following relationship holds

$$\Big\| \prod_{i=1}^k X_i \Big\|_{\psi_{\frac{\alpha}{k}}} \leq \prod_{i=1}^k \|X_i\|_{\psi_\alpha}.$$

Therefore, for $k = 4$ and $\alpha = 2$, we obtain that the product of four centered sub-Gaussians is a centered $\frac{1}{2}$-sub-exponential and the relationship $\|X_j X_k W_j W_k\|_{\psi_{1/2}} \leq \|X_j\|_{\psi_2}\|X_k\|_{\psi_2}\|W_j\|_{\psi_2}\|W_k\|_{\psi_2}$ holds. Next, we apply Corollary 1.4 from Götze et al. (2021) which gives the probabilistic tail bound for the sum of independent, centered $\frac{1}{2}$-sub-exponential random variables with $K = \max_i \|X_{ij}X_{ik}W_{ij}W_{ik}\|_{\psi_{1/2}}$. Therefore we get for $t > 0$

$$
\begin{aligned}
\Pr\left[\Big|\sum_{i=1}^{n} X_{ij}W_{ij}X_{ik}W_{ik}\Big| \geq t\right] &\leq \frac{1}{C}\exp\left[-c\min\left(\frac{t^2}{K^2}, \frac{\sqrt{t}}{\sqrt{K}}\right)\right] \\
&= \frac{1}{C}\exp\left[-c\min\left(\frac{t^2}{\left\{\max_i \|X_{ij}W_{ij}X_{ik}W_{ik}\|_{\psi_{1/2}}\right\}^2}, \right.\right. \\
&\qquad \left.\left. \frac{\sqrt{t}}{\sqrt{\max_i \|X_{ij}W_{ij}X_{ik}W_{ik}\|_{\psi_{1/2}}}}\right)\right] \\
&\leq \frac{1}{C}\exp\left[-c\min\left(\frac{t^2}{\left\{\max_i \left(\|X_{ij}\|_{\psi_2}\|W_{ij}\|_{\psi_2}\|X_{ik}\|_{\psi_2}\|W_{ik}\|_{\psi_2}\right)\right\}^2}, \right.\right. \\
&\qquad \left.\left. \frac{\sqrt{t}}{\sqrt{\max_i \left(\|X_{ij}\|_{\psi_2}\|W_{ij}\|_{\psi_2}\|X_{ik}\|_{\psi_2}\|W_{ik}\|_{\psi_2}\right)}}\right)\right] \\
&= \frac{1}{C}\exp\left[-c\min\left(\frac{t^2}{\max_i \|X_{ij}\|_{\psi_2}^2\|W_{ij}\|_{\psi_2}^2\|X_{ik}\|_{\psi_2}^2\|W_{ik}\|_{\psi_2}^2}, \right.\right. \\
&\qquad \left.\left. \frac{\sqrt{t}}{\sqrt{\max_i \|X_{ij}\|_{\psi_2}\|W_{ij}\|_{\psi_2}\|X_{ik}\|_{\psi_2}\|W_{ik}\|_{\psi_2}}}\right)\right] \\
&\leq \frac{1}{C}\exp\left[-c\min\left(\frac{t^2}{\sigma_X^4\sigma_W^4}, \frac{\sqrt{t}}{\sqrt{\sigma_X^2\sigma_W^2}}\right)\right].
\end{aligned}
$$

The last inequality is due to the implication from Lemma 5.5 of Vershynin (2010) that for two generic sub-Gaussian random variables X and W there exist universal constants m_1, m_2, M_1 and M_2 such that $m_1\|X\|_{\psi_2}^2 \leq \sigma_X^2 \leq M_1\|X\|_{\psi_2}^2$ and $m_2\|W\|_{\psi_2}^2 \leq \sigma_W^2 \leq M_2\|W\|_{\psi_2}^2$ hold. Now setting $t = \delta n m_{\min}$, we get,

$$
\Pr\left[\frac{1}{nm_{\min}}\Big|\sum_{i=1}^{n} X_{ij}W_{ij}X_{ik}W_{ik}\Big| \geq \delta\right] \leq \frac{1}{C}\exp\left[-c\min\left(\frac{\delta^2 n^2 m_{\min}^2}{\sigma_X^4\sigma_W^4}, \frac{\sqrt{\delta n m_{\min}}}{\sqrt{\sigma_X^2\sigma_W^2}}\right)\right]
$$

$$\leq \frac{1}{C}\exp\left[-c\sqrt{nm_{\min}}\min\left(\frac{\delta^2}{\sigma_X^4\sigma_W^4}, \frac{\sqrt{\delta}}{\sqrt{\sigma_X^2\sigma_W^2}}\right)\right].$$

Therefore, we obtain the probability bound for the first term in (3.45) where $\xi = \max(\sigma_W^4\sigma_X^4, \sigma_W\sigma_X)$. When $\delta \leq \sigma_X^2\sigma_W^2$, the bound can be simplified to

$$\Pr\Big(\frac{1}{nm_{\min}}\Big|\sum_{i=1}^{n} X_{ij}W_{ij}X_{ik}W_{ik}\Big| \geq \delta\Big) \leq \frac{1}{C}\exp\Big[-\frac{c\sqrt{nm_{\min}}\delta^2}{\sigma_X^4\sigma_W^4}\Big].$$

The term T_2 contains product of two sub-Gaussian random variables and therefore would follow sub-exponential by Lemma 2.7.7 from Vershynin (2018). Furthermore, following Corollary 5.17 from Vershynin (2010), the sum of independent, centered sub-exponential random variables have the following probabilistic bound:

$$\Pr\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} X_{ij}X_{ik}\Big| \geq \delta\Big) \leq \frac{1}{C}\exp\left[-cn\min\left(\frac{\delta^2}{\sigma_X^4}, \frac{\delta}{\sigma_X^2}\right)\right]$$

with $\xi = \max(\sigma_X^4, \sigma_X^2)$. When $\delta \leq \sigma_X^2$, the bound can be simplified to

$$\Pr\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} X_{ij}X_{ik}\Big| \geq \delta\Big) \leq \frac{1}{C}\exp[-\frac{cn\delta^2}{\sigma_X^4}].$$

To upper bound for the elementwise max norm we can apply Lemma 7. Putting these two pieces together, we see that $\widehat{\mathbf{\Sigma}}_{\text{mult}}$ and $\mathbf{S}$ satisfy the closeness condition in (3.10) with $\xi = \max(\sigma_W^4\sigma_X^4, \sigma_W\sigma_X, \sigma_X^4, \sigma_X^2)$. For a sufficiently small δ, specifically for $\delta \leq \min(\sigma_X^2\sigma_W^2, \sigma_X^2)$, $\widehat{\mathbf{\Sigma}}_{\text{mult}}$ and $\mathbf{S}$ satisfy the closeness condition in (3.10) with $\xi = \max(\sigma_X^4\sigma_W^4, \sigma_X^4)$. □

Proof of Lemma 4

Proof. For a multiplicative measurement error model, we assume the observed matrix is $\mathbf{Z} = \mathbf{X} \odot \mathbf{W}$ where $\mathbf{W}$ is a matrix of multiplicative error. Let $\mathbf{\Sigma}_{jk}^* = (\sigma_{X,jk})_{p\times p}$ and $\mathbf{\Sigma}_{W,jk} = (\sigma_{W,jk})_{p\times p}$ be the covariance matrices of $\mathbf{X}$ and the measurement error $\mathbf{W}$, for the

multiplicative model, respectively. Let X_j and W_k be the random variables representing the jth and the kth column, respectively, where $j, k = 1, \ldots, p$. Here, we assume that each column X_j and W_k are each independently and identically distributed with bounded moments

$$\mathbb{E}\left[X_j^{4m}\right] \leq K_{m,X} \qquad \text{and} \qquad \mathbb{E}\left[W_k^{4m}\right] \leq K_{m,W}.$$

Here m is a positive integer and $K_{m,X}, K_{m,W} \in \mathbb{R}^+$. We also assumed that X_j and W_k for any j and k are independent. Given $\mathbf{S}$ as the sample covariance matrix for the data without any corruption, we have

$$\widehat{\mathbf{\Sigma}}_{\text{mult}} - \mathbf{S} = \frac{1}{n}(\mathbf{X} \odot \mathbf{W})^\top (\mathbf{X} \odot \mathbf{W}) \oslash (\mathbf{\Sigma}_W + \mu_W \mu_W^\top) - \frac{1}{n}\mathbf{X}^\top \mathbf{X}.$$

The jkth element of this matrix can be written as

$$\begin{aligned}
\widehat{\mathbf{\Sigma}}_{\text{mult},jk} - \mathbf{S}_{jk} &= \frac{1}{n}\frac{\sum_{i=1}^n X_{ij}W_{ij}X_{ik}W_{ik}}{\mu_j\mu_k + \sigma_{W,jk}} - \frac{1}{n}\sum_{i=1}^n X_{ij}X_{ik} \\
&= \sum_{i=1}^n \left\{ \frac{1}{n}\frac{X_{ij}W_{ij}X_{ik}W_{ik}}{(\mu_j\mu_k + \sigma_{W,jk})} - \frac{1}{n}X_{ij}X_{ik} \right\}.
\end{aligned}$$

Let us define the random variable $T_{jk}^{(i)}$ as

$$T_{jk}^{(i)} = \frac{1}{n}\frac{X_{ij}W_{ij}X_{ik}W_{ik}}{(\mu_j\mu_k + \sigma_{W,jk})} - \frac{1}{n}X_{ij}X_{ik}$$

and note that they have mean zero. Applying the regularity conditions defined in (3.44), we have,

$$T_{jk}^{(i)} = \frac{X_{ij}W_{ij}X_{ik}W_{ik}}{nm_{\min}} - \frac{1}{n}X_{ij}X_{ik}$$

where $m_{\min}$ is a constant and it is defined in the proof of Lemma 3. By applying Chebyshev's inequality, we obtain,

$$\begin{aligned}\Pr\left[\left|\sum_{i=1}^{n} T_{jk}^{(i)}\right| > \delta\right] &\leq \Pr\left[\left(\sum_{i=1}^{n} T_{jk}^{(i)}\right)^{2m} > \delta^{2m}\right] \\ &\leq \frac{\mathbb{E}\left[\left(\sum_{i=1}^{n} T_{jk}^{(i)}\right)^{2m}\right]}{\delta^{2m}}. \end{aligned} \tag{3.46}$$

Now, applying Rosenthal's inequality (Rosenthal, 1970) to obtain that there exists a constant C_m, depending only on m, such that

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i=1}^{n} T_{jk}^{(i)}\right)^{2m}\right] &\leq C_m \max\left(\sum_{i=1}^{n}\mathbb{E}[(T_{jk}^{(i)})^{2m}], \left(\sum_{i=1}^{n}\mathbb{E}[(T_{jk}^{(i)})^{2}]\right)^{m}\right) \\ &\leq C_m\left(\sum_{i=1}^{n}\mathbb{E}[(T_{jk}^{(i)})^{2m}] + \left(\sum_{i=1}^{n}\mathbb{E}[(T_{jk}^{(i)})^{2}]\right)^{m}\right). \end{aligned} \tag{3.47}$$

Turning to each individual expectation, we have

$$\begin{aligned}\mathbb{E}[(T_{jk}^{(i)})^{2m}] &= \mathbb{E}\left[\left\{\frac{1}{n}\frac{X_{ij}W_{ij}X_{ik}W_{ik}}{m_{\min}} - \frac{1}{n}X_{ij}X_{ik}\right\}^{2m}\right] \\ &\overset{\text{(i)}}{\leq} 2^{2m-1}\mathbb{E}\left[\left\{\frac{1}{m_{\min}n}X_{ij}W_{ij}X_{ik}W_{ik}\right\}^{2m} + \left\{\frac{1}{n}X_{ij}X_{ik}\right\}^{2m}\right] \\ &\leq 2^{2m}\left[\frac{1}{m_{\min}^{2m}n^{2m}}\mathbb{E}[X_{ij}^{2m}W_{ij}^{2m}X_{ik}^{2m}W_{ik}^{2m}] + \frac{1}{n^{2m}}\mathbb{E}[X_{ij}^{2m}X_{ik}^{2m}]\right] \\ &\overset{\text{(ii)}}{\leq} \frac{2^{2m}}{n^{2m}}\left[\frac{1}{m_{\min}^{2m}}\mathbb{E}[X_{ij}^{2m}X_{ik}^{2m}]\mathbb{E}[W_{ij}^{2m}W_{ik}^{2m}] + \mathbb{E}[X_{ij}^{2m}X_{ik}^{2m}]\right] \\ &\overset{\text{(iii)}}{\leq} \frac{2^{2m}}{n^{2m}}\left[\frac{1}{m_{\min}^{2m}}\sqrt{\mathbb{E}[X_{ij}^{4m}]\mathbb{E}[X_{ik}^{4m}]}\sqrt{\mathbb{E}[W_{ij}^{4m}]\mathbb{E}[W_{ik}^{4m}]} + \sqrt{\mathbb{E}[X_{ij}^{4m}]\mathbb{E}[X_{ik}^{4m}]}\right] \\ &\overset{\text{(iv)}}{\leq} \frac{2^{2m}}{n^{2m}}\left[\frac{1}{m_{\min}^{2m}}K_{m,X}K_{m,W} + K_{m,X}\right] \end{aligned}$$

where inequality (i) follows because of the relationship $(\sum_{i=1}^{k} a_i)^n \leq k^{n-1}(\sum_{i=1}^{k} a_i^n)$. Specifically, since $(a+b)^{2m} \leq 2^{2m}(a^{2m}+b^{2m})$. The first term in inequality (ii) follows since X and W are independent. The terms in inequality (iii) follows from Cauchy Schwartz inequality. Inequality (iv) follows from the assumed moment bounds on $\mathbb{E}[X_j^{4m}]$ and $\mathbb{E}[W_j^{4m}]$. Now for $m = 1$, we have

$$\mathbb{E}[(T_{jk}^{(i)})^2] \lesssim \frac{2^2}{n^2}\left[1 + \frac{1}{m_{\min}^2}\right].$$

Hence,

$$\begin{aligned}\left(\sum_{i=1}^{n} \mathbb{E}[(T_{jk}^{(i)})^2]\right)^m &\lesssim \left[\frac{2^2 n}{n^2}\left\{1 + \frac{1}{m_{\min}^2}\right\}\right]^m \\ &\overset{\text{(i)}}{\leq} \frac{2^{3m} n^m}{n^{2m}}\left\{1 + \frac{1}{m_{\min}^{2m}}\right\}\end{aligned}$$

where inequality (i) uses the relationship $(a+b)^m \leq 2^m(a^m + b^m)$. Combined with the earlier bound (3.42), we have

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i=1}^{n} T_{jk}^{(i)}\right)^{2m}\right] &\leq C_m\left[\frac{2^{2m} n}{n^{2m}}\left\{\frac{1}{m_{\min}^{2m}} K_{m,X} K_{m,W} + K_{m,X}\right\} + \frac{2^{3m} n^m}{n^{2m}}\left\{1 + \frac{1}{m_{\min}^{2m}}\right\}\right] \\ &\overset{\text{(i)}}{\leq} \frac{C_m 2^{2m} n^m}{n^{2m}}\left[\frac{1}{m_{\min}^{2m}} K_{m,X} K_{m,W} + K_{m,X} + \frac{2}{m_{\min}^{2m}}\right] \\ &\leq \frac{C_m 2^{2m}}{n^m}\left[\frac{1}{m_{\min}^{2m}}\left(2 + K_{m,X} K_{m,W}\right) + K_{m,X}\right]\end{aligned}$$

Inequality (i) holds since $n \leq n^m$. Substituting back to the Chebyshev's inequality in (3.46) yields the tail bound

$$\Pr\left[\left|\sum_{i=1}^{n} T_{jk}^{(i)}\right| > \delta\right] \leq \frac{C_m 2^{2m}}{n^m \delta^{2m}}\left[\frac{1}{m_{\min}^{2m}}\left(2 + K_{m,X} K_{m,W}\right) + K_{m,X}\right]$$

To upper bound for the elementwise max norm we can apply Lemma 7. Hence the claim is established. □

Consistency of $\widehat{\Theta}$ and its error in elementwise ℓ_∞- norm:

For completeness, we now present the results required to show the consistency of $\widehat{\mathbf{\Theta}}$. The framework for the proofs are adopted from Ravikumar et al. (2011).

Lemma 8. *For any $\lambda_n > 0$ and the projected sample covariance $\widetilde{\mathbf{\Sigma}}$ with strictly positive diagonal elements, the ℓ_1-regularized log-determinant problem has a unique solution $\widehat{\mathbf{\Theta}}$ characterized by*

$$\widetilde{\mathbf{\Sigma}} - \widehat{\mathbf{\Theta}}^{-1} + \lambda_n \widehat{\mathbf{Z}} = 0 \tag{3.48}$$

where $\widehat{\mathbf{Z}}$ is an element of the sub-differential $\partial |\!|\!| \widehat{\mathbf{\Theta}} |\!|\!|_{1,\text{off}}$.

Proof. If $\lambda_n > 0$, then the CoGlasso objective function can be written in an equivalent constrained form using Lagrangian duality as follows:

$$\min_{\mathbf{\Theta} \in \mathcal{S}_{++}^p, |\!|\!| \mathbf{\Theta} |\!|\!|_{1,\text{off}} \leq C(\lambda_n)} \Big\{ \langle\!\langle \mathbf{\Theta}, \widetilde{\mathbf{\Sigma}} \rangle\!\rangle - \log\det(\mathbf{\Theta}) \Big\} \tag{3.49}$$

for some $C(\lambda_n) < \infty$. The behaviour of the objective function for sequences with possibly unbounded diagonal entries is the only possible concern, since the off-diagonal elements remain bounded within the ℓ_1-ball, since, $|\!|\!| \mathbf{\Theta} |\!|\!|_{1,\text{off}} \leq C(\lambda_n)$. The diagonal entries would be positive since any $\mathbf{\Theta}$ in the constraint set is positive-definite. We can show this by using the standard basis vector e_i defined by $e_i = 1$ at the i^{th} position and zero otherwise for $i = 1, 2, \ldots, p$. Since $\mathbf{\Theta}$ is positive definite, then $x^\top \mathbf{\Theta} x > 0$ for any non-zero vector $x \in \mathbb{R}^p$. Then $e_i^\top \mathbf{\Theta} e_i = \mathbf{\Theta}_{ii} > 0$, for all $i = 1, 2, \ldots, p$, showing that the diagonal elements are indeed positive for a positive definite symmetric matrix.

Next, we can upper bound the term $\log\det \mathbf{\Theta}$ using Hadamard's inequality (Horn and Johnson, 2012) for positive definite matrices which states that $\det \mathbf{\Theta} \leq \prod_{i=1}^p \mathbf{\Theta}_{ii}$. Therefore, we can write $\log\det \mathbf{\Theta} \leq \sum_{i=1}^p \log \mathbf{\Theta}_{ii}$. Since the off-diagonal elements are bounded within the ℓ_1-ball, so we only need to show that the following function involving the diagonal

elements is coercive, that is,

$$\sum_{i=1}^{p} \boldsymbol{\Theta}_{ii}\widetilde{\boldsymbol{\Sigma}}_{ii} - \log\det\boldsymbol{\Theta} \geq \sum_{i=1}^{p}\left\{\boldsymbol{\Theta}_{ii}\widetilde{\boldsymbol{\Sigma}}_{ii} - \log\boldsymbol{\Theta}_{ii}\right\}.$$

diverges to infinity for any sequence indexed by t, $\|\boldsymbol{\Theta}^t_{11}, \ldots, \boldsymbol{\Theta}^t_{pp}\|_2 \to +\infty$, as long as $\widetilde{\boldsymbol{\Sigma}}_{ii} > 0$ for each $i = 1, \ldots, p$. Therefore, the minimum is attained. Here, $-\sum_{i=1}^{p}\log\boldsymbol{\Theta}_{ii}$ is termed the logarithmic-determinant barrier function and it is strictly convex since $\boldsymbol{\Theta}_{ii} > 0$. By the strict convexity of the log-determinant barrier, the minimum would be unique.
For the regularized form, the matrix $\widehat{\boldsymbol{\Theta}} \in \mathcal{S}^p_{++}$ is optimal if and only if the zero matrix belongs to the sub-differential of the objective, or equivalently if and only if there exists a matrix $\widehat{\mathbf{Z}}$ in the sub-differential of the off-diagonal norm $|||\cdot|||_{1,\text{off}}$ evaluated at $\widehat{\boldsymbol{\Theta}}$ such that $\widetilde{\boldsymbol{\Sigma}} - \widehat{\boldsymbol{\Theta}}^{-1} + \lambda_n\widehat{\mathbf{Z}} = 0$, as claimed, by standard optimality conditions for convex programs. □

The following lemma provides sufficient condition to show that step (d) from the primal-dual witness condition, the strict dual feasibility holds, so that $|||\widetilde{\mathbf{Z}}|||_{\max} < 1$.

Lemma 9. *[Strict dual feasibility.] Suppose that*

$$\max\left\{|||\mathbf{W}|||_{\max}, |||R(\boldsymbol{\Delta})|||_{\max}\right\} \leq \frac{\alpha\lambda_n}{8} \tag{3.50}$$

Then the vector $\widetilde{\mathbf{Z}}_{S^c}$ constructed in step (c) of primal-dual witness condition satisfies $|||\widetilde{\mathbf{Z}}_{S^c}|||_{\max} < 1$, and therefore $\widetilde{\boldsymbol{\Theta}} = \widehat{\boldsymbol{\Theta}}$.

Proof. We can re-write the stationarity condition (3.48) using (3.19) and (3.20) as

$$\begin{aligned}
0 =& \widetilde{\boldsymbol{\Sigma}} - \widetilde{\boldsymbol{\Theta}}^{-1} + \lambda_n\widetilde{\mathbf{Z}} \\
=& \widetilde{\boldsymbol{\Sigma}} - \widetilde{\boldsymbol{\Theta}}^{-1} + \widetilde{\boldsymbol{\Theta}}^{-1} - (\boldsymbol{\Theta}^*)^{-1} + (\boldsymbol{\Theta}^*)^{-1}\boldsymbol{\Delta}(\boldsymbol{\Theta}^*)^{-1} - R(\boldsymbol{\Delta}) + \lambda_n\widetilde{\mathbf{Z}} \\
=& \widetilde{\boldsymbol{\Sigma}} - (\boldsymbol{\Theta}^*)^{-1} + (\boldsymbol{\Theta}^*)^{-1}\boldsymbol{\Delta}(\boldsymbol{\Theta}^*)^{-1} - R(\boldsymbol{\Delta}) + \lambda_n\widetilde{\mathbf{Z}} \\
=& (\boldsymbol{\Theta}^*)^{-1}\boldsymbol{\Delta}(\boldsymbol{\Theta}^*)^{-1} + \mathbf{W} - R(\boldsymbol{\Delta}) + \lambda_n\widetilde{\mathbf{Z}}
\end{aligned} \tag{3.51}$$

This can be re-written as a linear equation by "vectorizing" the matrices. Let us use the notation $\mathrm{vec}(\mathbf{A})$ or equivalently $\bar{\mathbf{A}}$ for the vector version of the set or matrix $\mathbf{A}$, obtained by stacking up the rows of $\mathbf{A}$ into a single column vector.

$$\mathrm{vec}((\boldsymbol{\Theta}^*)^{-1}\boldsymbol{\Delta}(\boldsymbol{\Theta}^*)^{-1}) = ((\boldsymbol{\Theta}^*)^{-1} \otimes (\boldsymbol{\Theta}^*)^{-1})\bar{\boldsymbol{\Delta}} = \boldsymbol{\Gamma}^*\bar{\boldsymbol{\Delta}}.$$

Equation (3.51) can be decomposed into two blocks of linear equations in terms of the disjoint decomposition S and S^c as follows:

$$\boldsymbol{\Gamma}^*_{SS}\bar{\boldsymbol{\Delta}}_S + \bar{\mathbf{W}}_S - \bar{\mathbf{R}}_S + \lambda_n\bar{\widetilde{\mathbf{Z}}}_S = 0 \tag{3.52}$$

$$\boldsymbol{\Gamma}^*_{S^cS}\bar{\boldsymbol{\Delta}}_S + \bar{\mathbf{W}}_{S^c} - \bar{\mathbf{R}}_{S^c} + \lambda_n\bar{\widetilde{\mathbf{Z}}}_{S^c} = 0, \tag{3.53}$$

since by construction $\boldsymbol{\Delta}_{S^c} = 0$. We can solve for $\bar{\boldsymbol{\Delta}}_S$ from (3.52) as follows, since $\boldsymbol{\Gamma}^*_{SS}$ is invertible:

$$\bar{\boldsymbol{\Delta}}_S = (\boldsymbol{\Gamma}^*_{SS})^{-1}\left\{-\bar{\mathbf{W}}_S + \bar{\mathbf{R}}_S - \lambda_n\bar{\widetilde{\mathbf{Z}}}_S\right\}.$$

Substituting this into (3.53), we can solve for $\widetilde{\mathbf{Z}}_{S^c}$ as follows:

$$\begin{aligned}
\bar{\widetilde{\mathbf{Z}}}_{S^c} &= \frac{1}{\lambda_n}\left\{-\boldsymbol{\Gamma}^*_{S^cS}\bar{\boldsymbol{\Delta}}_S - \bar{\mathbf{W}}_{S^c} + \bar{\mathbf{R}}_{S^c}\right\} \\
&= \frac{1}{\lambda_n}\left\{-\boldsymbol{\Gamma}^*_{S^cS}(\boldsymbol{\Gamma}^*_{SS})^{-1}\left\{-\bar{\mathbf{W}}_S + \bar{\mathbf{R}}_S - \lambda_n\bar{\widetilde{\mathbf{Z}}}_S\right\} - \bar{\mathbf{W}}_{S^c} + \bar{\mathbf{R}}_{S^c}\right\} \\
&= \frac{1}{\lambda_n}\boldsymbol{\Gamma}^*_{S^cS}(\boldsymbol{\Gamma}^*_{SS})^{-1}(\bar{\mathbf{W}}_S - \bar{\mathbf{R}}_S) + \boldsymbol{\Gamma}^*_{S^cS}(\boldsymbol{\Gamma}^*_{SS})^{-1}\bar{\widetilde{\mathbf{Z}}}_S - \frac{1}{\lambda_n}(\bar{\mathbf{W}}_{S^c} - \bar{\mathbf{R}}_{S^c}).
\end{aligned}$$

Next, we take ℓ_∞-operator norm on both sides and apply the triangular inequality

$$\begin{aligned}
\|\bar{\widetilde{\mathbf{Z}}}_{S^c}\|_\infty &\leq \frac{1}{\lambda_n}\|\boldsymbol{\Gamma}^*_{S^cS}(\boldsymbol{\Gamma}^*_{SS})^{-1}(\bar{\mathbf{W}}_S - \bar{\mathbf{R}}_S)\|_\infty + \|\boldsymbol{\Gamma}^*_{S^cS}(\boldsymbol{\Gamma}^*_{SS})^{-1}\bar{\widetilde{\mathbf{Z}}}_S\|_\infty \\
&\qquad + \frac{1}{\lambda_n}(\|\bar{\mathbf{W}}_{S^c}\|_\infty + \|\bar{\mathbf{R}}_{S^c}\|_\infty) \\
&\overset{\text{(i)}}{\leq} \frac{1}{\lambda_n}|||\boldsymbol{\Gamma}^*_{S^cS}(\boldsymbol{\Gamma}^*_{SS})^{-1}|||_\infty(\|\bar{\mathbf{W}}_S\|_\infty + \|\bar{\mathbf{R}}_S\|_\infty) + \|\boldsymbol{\Gamma}^*_{S^cS}(\boldsymbol{\Gamma}^*_{SS})^{-1}\|_\infty\|\bar{\widetilde{\mathbf{Z}}}_S\|_\infty
\end{aligned}$$

$$
\begin{aligned}
&\qquad\qquad + \frac{1}{\lambda_n}(\|\bar{\mathbf{W}}_{S^c}\|_\infty + \|\bar{\mathbf{R}}_{S^c}\|_\infty) \\
&\overset{\text{(ii)}}{\leq} \frac{(1-\alpha)}{\lambda_n}(\|\bar{\mathbf{W}}_S\|_\infty + \|\bar{\mathbf{R}}_S\|_\infty) + (1-\alpha) + \frac{1}{\lambda_n}(\|\bar{\mathbf{W}}_{S^c}\|_\infty + \|\bar{\mathbf{R}}_{S^c}\|_\infty) \\
&\leq (1-\alpha) + \frac{2}{\lambda_n}(\|\bar{\mathbf{W}}\|_\infty + \|\bar{\mathbf{R}}\|_\infty) - \frac{\alpha}{\lambda_n}(\|\bar{\mathbf{W}}\|_\infty + \|\bar{\mathbf{R}}\|_\infty) \\
&\overset{\text{(iii)}}{\leq} (1-\alpha) + \frac{\alpha}{2} - \frac{\alpha^2}{4} \\
&\leq 1 - \alpha + \frac{\alpha}{2} \\
&< 1,
\end{aligned}
$$

as claimed. Inequality (i) holds because of sub-multiplicative property of operator norms, $\|\mathbf{A}\mathbf{x}\|_\infty \leq |\!|\!|\mathbf{A}|\!|\!|_\infty \|\mathbf{x}\|_\infty$. Inequality (ii) is due to the mutual incoherence condition (3.18) and the fact that $\|\bar{\tilde{\mathbf{Z}}}_S\|_\infty \leq 1$, by construction. The third inequality utilizes the assumption made in the statement of Lemma 9. □

The following lemma is to relate the behaviour of the remainder term (3.20) to the deviation $\boldsymbol{\Delta} = \widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*$.

Lemma 10. *[Control of Remainder.] Suppose that the elementwise ℓ_∞-bound $|\!|\!|\boldsymbol{\Delta}|\!|\!|_{\max} \leq 1/(3\kappa_{\boldsymbol{\Sigma}^*} d)$ holds. Then the matrix $\mathbf{J} = \sum_{k=0}^{\infty}(-1)^k((\boldsymbol{\Theta}^*)^{-1}\boldsymbol{\Delta})^k$ satisfies the ℓ_∞- operator norm $|\!|\!|\mathbf{J}^T|\!|\!|_\infty \leq 3/2$, and moreover, the matrix*

$$
R(\boldsymbol{\Delta}) = (\boldsymbol{\Theta}^*)^{-1}\boldsymbol{\Delta}(\boldsymbol{\Theta}^*)^{-1}\boldsymbol{\Delta}\mathbf{J}(\boldsymbol{\Theta}^*)^{-1}, \tag{3.54}
$$

has elementwise ℓ_∞-norm bounded as

$$
|\!|\!|R(\boldsymbol{\Delta})|\!|\!|_{\max} \leq \frac{3}{2} d |\!|\!|\boldsymbol{\Delta}|\!|\!|^2_{\max} \kappa^3_{\boldsymbol{\Sigma}^*}. \tag{3.55}
$$

Proof. With $\boldsymbol{\Delta} = \widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*$, the remainder term can be rewritten as follows:

$$
R(\boldsymbol{\Delta}) = (\boldsymbol{\Theta}^* + \boldsymbol{\Delta})^{-1} - (\boldsymbol{\Theta}^*)^{-1} + (\boldsymbol{\Theta}^*)^{-1}\boldsymbol{\Delta}(\boldsymbol{\Theta}^*)^{-1}.
$$

Using matrix expansion of the first term in the expression of the remainder term, we get

$$\begin{aligned}
(\mathbf{\Theta}^* + \mathbf{\Delta})^{-1} &= (\mathbf{\Theta}^*(\mathbf{I} + (\mathbf{\Theta}^*)^{-1}\mathbf{\Delta}))^{-1} \\
&= (\mathbf{I} + (\mathbf{\Theta}^*)^{-1}\mathbf{\Delta})^{-1}(\mathbf{\Theta}^*)^{-1} \\
&= \sum_{k=0}^{\infty}(-1)^k((\mathbf{\Theta}^*)^{-1}\mathbf{\Delta})^k(\mathbf{\Theta}^*)^{-1} \\
&= (\mathbf{\Theta}^*)^{-1} - (\mathbf{\Theta}^*)^{-1}\mathbf{\Delta}(\mathbf{\Theta}^*)^{-1} + \sum_{k=2}^{\infty}(-1)^k((\mathbf{\Theta}^*)^{-1}\mathbf{\Delta})^k(\mathbf{\Theta}^*)^{-1} \\
&= (\mathbf{\Theta}^*)^{-1} - (\mathbf{\Theta}^*)^{-1}\mathbf{\Delta}(\mathbf{\Theta}^*)^{-1} + (\mathbf{\Theta}^*)^{-1}\mathbf{\Delta}(\mathbf{\Theta}^*)^{-1}\mathbf{\Delta}\mathbf{J}(\mathbf{\Theta}^*)^{-1},
\end{aligned}$$

where $\mathbf{J} = \sum_{k=0}^{\infty}(-1)^k((\mathbf{\Theta}^*)^{-1}\mathbf{\Delta})^k$.

To prove the bound for the remainder term, let e_i denote the unit vector with 1 in position i and zeros elsewhere. We have,

$$\begin{aligned}
|||R(\mathbf{\Delta})|||_{\max} &= \max_{i,j} \mid e_i^\top(\mathbf{\Theta}^*)^{-1}\mathbf{\Delta}(\mathbf{\Theta}^*)^{-1}\mathbf{\Delta}\mathbf{J}(\mathbf{\Theta}^*)^{-1}e_j \mid \\
&\overset{\text{(i)}}{\leq} \max_{i,j} \|e_i^\top(\mathbf{\Theta}^*)^{-1}\mathbf{\Delta}\|_\infty \|(\mathbf{\Theta}^*)^{-1}\mathbf{\Delta}\mathbf{J}(\mathbf{\Theta}^*)^{-1}e_j\|_1 \\
&= \max_{i} \|e_i^\top(\mathbf{\Theta}^*)^{-1}\mathbf{\Delta}\|_\infty \max_{j} \|(\mathbf{\Theta}^*)^{-1}\mathbf{\Delta}\mathbf{J}(\mathbf{\Theta}^*)^{-1}e_j\|_1 \\
&\overset{\text{(ii)}}{\leq} \max_{i} \|e_i^\top(\mathbf{\Theta}^*)^{-1}\|_1 |||\mathbf{\Delta}|||_{\max} \max_{j} \|(\mathbf{\Theta}^*)^{-1}\mathbf{\Delta}\mathbf{J}(\mathbf{\Theta}^*)^{-1}e_j\|_1 \\
&\overset{\text{(iii)}}{\leq} |||(\mathbf{\Theta}^*)^{-1}|||_\infty |||\mathbf{\Delta}|||_{\max} |||(\mathbf{\Theta}^*)^{-1}\mathbf{\Delta}\mathbf{J}(\mathbf{\Theta}^*)^{-1}|||_1 \\
&\overset{\text{(iv)}}{=} |||(\mathbf{\Theta}^*)^{-1}|||_\infty |||\mathbf{\Delta}|||_{\max} |||(\mathbf{\Theta}^*)^{-1}\mathbf{J}^T\mathbf{\Delta}(\mathbf{\Theta}^*)^{-1}|||_\infty \\
&\leq \kappa_{\mathbf{\Sigma}^*} |||\mathbf{\Delta}|||_{\max} |||(\mathbf{\Theta}^*)^{-1}|||_\infty^2 |||\mathbf{J}^\top|||_\infty |||\mathbf{\Delta}|||_\infty \\
&\leq \kappa_{\mathbf{\Sigma}^*}^3 |||\mathbf{\Delta}|||_{\max} |||\mathbf{J}^\top|||_\infty |||\mathbf{\Delta}|||_\infty
\end{aligned}$$

where the first inequality follows from Hölder's inequality. The second inequality follows since for a vector $\mathbf{u} \in \mathbb{R}^p$, $\|\mathbf{u}^\top\mathbf{\Delta}\|_\infty = \max_j \mid \sum_i u_i\Delta_{ij} \mid \leq |||\mathbf{\Delta}|||_{\max} \mid \sum_i u_i \mid \leq \|\mathbf{u}\|_1 |||\mathbf{\Delta}|||_{\max}$. The third inequality uses the fact that $|||\mathbf{A}|||_\infty = \max_i \sum_j \mid a_{ij} \mid$, i.e. the maximum absolute row sum of the matrix and $|||\mathbf{A}|||_1 = \max_j \sum_i \mid a_{ij} \mid$, i.e. the maximum absolute column

sum of the matrix. The fourth inequality follows from the relationship $|||\mathbf{A}|||_1 = |||\mathbf{A}^\top|||_\infty$. The last line following by using the definition of $\kappa_{\mathbf{\Sigma}^*} = |||(\mathbf{\Theta}^*)^{-1}|||_\infty$.

Since $\mathbf{J} = \sum_{k=0}^{\infty}(-1)^k((\mathbf{\Theta}^*)^{-1}\mathbf{\Delta})^k$ and using the submultiplicativity of $|||\cdot|||_\infty$ matrix norm, we have,

$$|||\mathbf{J}^\top|||_\infty \leq \sum_{k=0}^{\infty} |||\mathbf{\Delta}(\mathbf{\Theta}^*)^{-1}|||_\infty^k = \frac{1}{1-|||\mathbf{\Delta}(\mathbf{\Theta}^*)^{-1}|||_\infty} \leq \frac{1}{1-|||(\mathbf{\Theta}^*)^{-1}|||_\infty |||\mathbf{\Delta}|||_\infty} \leq \frac{1}{2/3} = \frac{3}{2},$$

because $|||(\mathbf{\Theta}^*)^{-1}|||_\infty |||\mathbf{\Delta}|||_\infty < 1/3$ from (3.56). We can verify that by applying submultiplicativity of the $|||\cdot|||_\infty$ matrix form to the term $(\mathbf{\Theta}^*)^{-1}\mathbf{\Delta}$. For any $p \times p$ matrices, we can write

$$|||(\mathbf{\Theta}^*)^{-1}\mathbf{\Delta}|||_\infty \leq |||(\mathbf{\Theta}^*)^{-1}|||_\infty |||\mathbf{\Delta}|||_\infty \leq \kappa_{\mathbf{\Sigma}^*} d |||\mathbf{\Delta}|||_{\max} \overset{\text{(i)}}{<} \frac{1}{3}, \tag{3.56}$$

where d is the maximum number of non-zeros in any row/column of $\mathbf{\Delta}$ and $\kappa_{\mathbf{\Sigma}^*} = |||\mathbf{\Sigma}^*|||_\infty$. We also used the fact $|||\mathbf{\Delta}|||_\infty \leq d|||\mathbf{\Delta}|||_{\max}$. This is true because

$$|||\mathbf{\Delta}|||_\infty = \max_i \sum_j |\ a_{ij}\ | \leq \max_i (d.\max_j |\ a_{ij}\ |) = d \max_{i,j} |\ a_{ij}\ | = d|||\mathbf{\Delta}|||_{\max}.$$

Inequality (i) follows from the assumption stated in the lemma that $|||\mathbf{\Delta}|||_{\max} \leq 1/3\kappa_{\mathbf{\Sigma}^*} d$.

Therefore, we have

$$\begin{aligned} |||R(\mathbf{\Delta})|||_{\max} &\leq \frac{3}{2}\kappa_{\mathbf{\Sigma}^*}^3 |||\mathbf{\Delta}|||_{\max} |||\mathbf{\Delta}|||_\infty \\ &\leq \frac{3}{2} d |||\mathbf{\Delta}|||_{\max}^2 \kappa_{\mathbf{\Sigma}^*}^3, \end{aligned}$$

since $|||\mathbf{\Delta}|||_\infty \leq d|||\mathbf{\Delta}|||_{\max}$, as $\mathbf{\Delta}$ has at most d non-zeros per row/column. Hence the proof is complete. □

To prove the sufficient condition for ℓ_∞-bounds, the next lemma provides control on the deviation $\mathbf{\Delta} = \widetilde{\mathbf{\Theta}} - \mathbf{\Theta}^*$, measured in elementwise ℓ_∞ norm.

Lemma 11. *[Control of the Error Deviation.] Suppose that*

$$r = 2\kappa_{\mathbf{\Gamma}^*}\left\{|||\mathbf{W}|||_{\max} + \lambda_n\right\} \leq \min\left\{\frac{1}{3\kappa_{\mathbf{\Sigma}^*}d}, \frac{1}{3\kappa^3_{\mathbf{\Sigma}^*}\kappa_{\mathbf{\Gamma}^*}d}\right\} \tag{3.57}$$

where $\kappa_{\mathbf{\Gamma}^*} = |||(\mathbf{\Gamma}^*_{SS})^{-1}|||_\infty$, $\kappa_{\mathbf{\Sigma}^*} = |||\mathbf{\Sigma}^*|||_\infty$ *and* $\mathbf{W} = \widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*$. *Then we have the elementwise* ℓ_∞ *bound*

$$|||\mathbf{\Delta}|||_{\max} = |||\widetilde{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_{\max} \leq r. \tag{3.58}$$

Proof. As proved in Lemma 8, we can conclude that the regularized problem (3.49) has a unique optimum $\widetilde{\mathbf{\Theta}}$. We proceed by noting that $\widetilde{\mathbf{\Theta}}_{S^c} = \mathbf{\Theta}^*_{S^c} = 0$, so that $|||\mathbf{\Delta}|||_{\max} = |||\mathbf{\Delta}_S|||_{\max}$. We get the zero-gradient condition by taking partial derivatives of the Lagrangian of the regularized problem with respect to the unconstrained elements $\mathbf{\Theta}_S$, since the partial derivatives are zero at the optimum

$$G(\mathbf{\Theta}_S) = -\mathbf{\Theta}_S^{-1} + \widetilde{\mathbf{\Sigma}}_S + \lambda_n \widetilde{\mathbf{Z}}_S = 0, \tag{3.59}$$

where $\mathbf{\Theta}$ is the $p \times p$ matrix with entries in S equal to $\mathbf{\Theta}_S$ and entries in S^c equal to zero. The zero-gradient condition is necessary and sufficient to achieve an optimum of the Lagrangian problem and therefore the problem has $\widetilde{\mathbf{\Theta}}_S$ as the unique solution.

To bound $\mathbf{\Delta} = \widetilde{\mathbf{\Theta}} - \mathbf{\Theta}^*$, we want to show that there exists a solution $\mathbf{\Delta}$ to the zero-gradient condition (3.59) that is contained within the ball

$$\mathbb{B}(r) = \left\{\mathbf{\Theta}_S \mid |||\mathbf{\Theta}_S|||_{\max} \leq r\right\}, \qquad \text{with } r = 2\kappa_{\mathbf{\Gamma}^*}\left\{|||\mathbf{W}|||_{\max} + \lambda_n\right\}. \tag{3.60}$$

Then, by the uniqueness of the optimal solution, we can conclude that $\widetilde{\mathbf{\Theta}} - \mathbf{\Theta}^*$ belongs to $\mathbb{B}(r)$. To do so, let us first define the error deviation in vectorized form, $\bar{\mathbf{\Delta}}_S = \bar{\widetilde{\mathbf{\Theta}}}_S - \bar{\mathbf{\Theta}}^*_S$. Next, let us define a continuous map $F : \mathbf{\Delta}_S \mapsto F(\mathbf{\Delta}_S)$ such that its fixed points are

equivalent to zeros of this gradient expression via

$$F(\bar{\mathbf{\Delta}}_S) = \bar{\mathbf{\Delta}}_S - (\mathbf{\Gamma}^*_{SS})^{-1}(\bar{G}(\mathbf{\Theta}^*_S + \mathbf{\Delta}_S)), \tag{3.61}$$

where $\bar{G}$ denotes the vectorized form of G. We have, by construction, $F(\bar{\mathbf{\Delta}}_S) = \bar{\mathbf{\Delta}}_S$ holds if and only if $G(\mathbf{\Theta}^*_S + \mathbf{\Delta}_S) = G(\widetilde{\mathbf{\Theta}}_S) = 0$. We can now apply Brouwer's fixed point theorem which states (Ortega and Rheinboldt, 2000) that for any continuous function f mapping a compact convex set to itself there is a point x_0 such that $f(x_0) = x_0$.

Since F is continuous and $\mathbb{B}(r)$ is convex and compact, by Brouwer's fixed point theorem, this inclusion implies that there exists some fixed point $\bar{\mathbf{\Delta}}_S \in \mathbb{B}(r)$. Therefore, we can claim that $F(\mathbb{B}(r)) \subseteq \mathbb{B}(r)$, that is, F indeed has a fixed point inside $\mathbb{B}(r)$. By uniqueness of the zero gradient condition and hence fixed points of F, it can be concluded that $|||\widetilde{\mathbf{\Theta}}_S - \mathbf{\Theta}^*_S|||_{\max} \leq r$.

Let $\mathbf{\Delta} \in \mathbb{R}^{p \times p}$ denote the zero-padded matrix which is equal to $\mathbf{\Delta}_S$ on S and zero on S^c. We can rewrite the zero-gradient expression by adding and subtracting $[(\mathbf{\Theta}^*)^{-1}]_S$

$$\begin{aligned} G(\mathbf{\Theta}^*_S + \mathbf{\Delta}_S) &= -[(\mathbf{\Theta}^* + \mathbf{\Delta})^{-1}]_S + \widetilde{\mathbf{\Sigma}}_S + \lambda_n \widetilde{\mathbf{Z}}_S \\ &= -[(\mathbf{\Theta}^* + \mathbf{\Delta})^{-1}]_S + (\widetilde{\mathbf{\Sigma}}_S - [(\mathbf{\Theta}^*)^{-1}]_S) + [(\mathbf{\Theta}^*)^{-1}]_S + \lambda_n \widetilde{\mathbf{Z}}_S \\ &= -[(\mathbf{\Theta}^* + \mathbf{\Delta})^{-1}]_S + [(\mathbf{\Theta}^*)^{-1}]_S + \mathbf{W}_S + \lambda_n \widetilde{\mathbf{Z}}_S, \end{aligned} \tag{3.62}$$

by using the definition of $\mathbf{W} = \widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*$.

Next, we can write the vectorized form of the remainder term (3.20), restricting the entries to S as

$$\begin{aligned} \bar{\mathbf{R}}_S &= \mathrm{vec}((\mathbf{\Theta}^* + \mathbf{\Delta})^{-1} - (\mathbf{\Theta}^*)^{-1}))_S + \mathrm{vec}((\mathbf{\Theta}^*)^{-1}\mathbf{\Delta}(\mathbf{\Theta}^*)^{-1})_S \\ &= \mathrm{vec}((\mathbf{\Theta}^* + \mathbf{\Delta})^{-1} - (\mathbf{\Theta}^*)^{-1}))_S + ((\mathbf{\Theta}^*)^{-1} \otimes (\mathbf{\Theta}^*)^{-1})_{SS}\bar{\mathbf{\Delta}}_S \\ &= \mathrm{vec}((\mathbf{\Theta}^* + \mathbf{\Delta})^{-1} - (\mathbf{\Theta}^*)^{-1}))_S + \mathbf{\Gamma}^*_{SS}\bar{\mathbf{\Delta}}_S. \end{aligned}$$

This expression is equal to the vectorized form of expansion (3.54) where the entries are restricted to S. Therefore,

$$\begin{aligned} &\text{vec}((\boldsymbol{\Theta}^* + \boldsymbol{\Delta})^{-1} - (\boldsymbol{\Theta}^*)^{-1}))_S + \boldsymbol{\Gamma}^*_{SS}\bar{\boldsymbol{\Delta}}_S = \text{vec}(((\boldsymbol{\Theta}^*)^{-1}\boldsymbol{\Delta})^2\mathbf{J}(\boldsymbol{\Theta}^*)^{-1})_S \qquad (3.63) \\ \implies &\text{vec}((\boldsymbol{\Theta}^* + \boldsymbol{\Delta})^{-1} - (\boldsymbol{\Theta}^*)^{-1}))_S = \text{vec}[((\boldsymbol{\Theta}^*)^{-1}\boldsymbol{\Delta})^2\mathbf{J}(\boldsymbol{\Theta}^*)^{-1}]_S - \boldsymbol{\Gamma}^*_{SS}\bar{\boldsymbol{\Delta}}_S \end{aligned}$$

We can rewrite (3.61), combined with (3.63) and (3.62),

$$\begin{aligned} F(\bar{\boldsymbol{\Delta}}_S) &= \bar{\boldsymbol{\Delta}}_S - (\boldsymbol{\Gamma}^*_{SS})^{-1}(\bar{G}(\boldsymbol{\Theta}^*_S + \boldsymbol{\Delta}_S)) \\ &= \bar{\boldsymbol{\Delta}}_S + (\boldsymbol{\Gamma}^*_{SS})^{-1}\text{vec}\left\{(\boldsymbol{\Theta}^* + \boldsymbol{\Delta})^{-1} - (\boldsymbol{\Theta}^*)^{-1}) - \mathbf{W}_S - \lambda_n\widetilde{\mathbf{Z}}_S\right\}_S \\ &= \bar{\boldsymbol{\Delta}}_S + (\boldsymbol{\Gamma}^*_{SS})^{-1}\left\{\text{vec}[((\boldsymbol{\Theta}^*)^{-1}\boldsymbol{\Delta})^2\mathbf{J}(\boldsymbol{\Theta}^*)^{-1}]_S - \boldsymbol{\Gamma}^*_{SS}\bar{\boldsymbol{\Delta}}_S\right\} - (\boldsymbol{\Gamma}^*_{SS})^{-1}\left\{\bar{\mathbf{W}}_S + \lambda_n\bar{\widetilde{\mathbf{Z}}}_S\right\} \\ &= \bar{\boldsymbol{\Delta}}_S - (\boldsymbol{\Gamma}^*_{SS})^{-1}\boldsymbol{\Gamma}^*_{SS}\bar{\boldsymbol{\Delta}}_S + (\boldsymbol{\Gamma}^*_{SS})^{-1}\left\{\text{vec}[((\boldsymbol{\Theta}^*)^{-1}\boldsymbol{\Delta})^2\mathbf{J}(\boldsymbol{\Theta}^*)^{-1}]_S\right\} \\ &\qquad\qquad - (\boldsymbol{\Gamma}^*_{SS})^{-1}\left\{\bar{\mathbf{W}}_S + \lambda_n\bar{\widetilde{\mathbf{Z}}}_S\right\} \\ &= (\boldsymbol{\Gamma}^*_{SS})^{-1}\left\{\text{vec}[((\boldsymbol{\Theta}^*)^{-1}\boldsymbol{\Delta})^2\mathbf{J}(\boldsymbol{\Theta}^*)^{-1}]_S\right\} - (\boldsymbol{\Gamma}^*_{SS})^{-1}\left\{\bar{\mathbf{W}}_S + \lambda_n\bar{\widetilde{\mathbf{Z}}}_S\right\}. \qquad (3.64) \end{aligned}$$

Notice, that for any $\boldsymbol{\Delta}_S \in \mathbb{B}(r)$, by sub-multiplicativity of the $|||\cdot|||_\infty$ we have

$$|||(\boldsymbol{\Theta}^*)^{-1}\boldsymbol{\Delta}|||_\infty \leq |||(\boldsymbol{\Theta}^*)^{-1}|||_\infty |||\boldsymbol{\Delta}|||_\infty \leq \kappa_{\boldsymbol{\Sigma}^*} d|||\boldsymbol{\Delta}|||_{\max},$$

where $|||\boldsymbol{\Delta}|||_{\max}$ and $|||\boldsymbol{\Delta}|||_\infty$ denote the elementwise ℓ_∞-norm and ℓ_∞-operator norm respectively with d being the maximum number of non-zero entries per row/column of $\boldsymbol{\Delta}$.

Now, we can apply the results of Lemma 10 to the error deviation. By definition of the radius r and the assumed upper bound defined in (3.57), we have

$$|||\boldsymbol{\Delta}|||_{\max} \leq r \leq \frac{1}{3\kappa_{\boldsymbol{\Sigma}^*} d}.$$

We can take ℓ_∞-norm on both sides of (3.64), use the triangular inequality and submultiplicativity property of matrix operator norms to calculate our desired bound. Beginning with the second term, we have

$$\begin{aligned}
\|(\mathbf{\Gamma}^*_{SS})^{-1}\left\{\bar{\mathbf{W}}_S + \lambda_n \bar{\tilde{\mathbf{Z}}}_S\right\}\|_\infty &\leq |||(\mathbf{\Gamma}^*_{SS})^{-1}|||_\infty (\|\bar{\mathbf{W}}_S\|_\infty + \lambda_n) \\
&\leq \kappa_{\mathbf{\Gamma}^*}(\|\bar{\mathbf{W}}\|_\infty + \lambda_n) \\
&= \frac{r}{2},
\end{aligned}$$

where the inequality follows from the assumed upper bound in (3.57).

To show the bound for the first term, we have

$$\begin{aligned}
\|(\mathbf{\Gamma}^*_{SS})^{-1}\left\{\mathrm{vec}[((\mathbf{\Theta}^*)^{-1}\mathbf{\Delta})^2\mathbf{J}(\mathbf{\Theta}^*)^{-1}]_S\right\}\|_\infty &\leq |||(\mathbf{\Gamma}^*_{SS})^{-1}|||_\infty \|\mathrm{vec}[((\mathbf{\Theta}^*)^{-1}\mathbf{\Delta})^2\mathbf{J}(\mathbf{\Theta}^*)^{-1}]_S\|_\infty \\
&\leq \kappa_{\mathbf{\Gamma}^*}\|\bar{\mathbf{R}}_S\|_\infty \\
&\leq \kappa_{\mathbf{\Gamma}^*}|||R(\mathbf{\Delta})|||_{\max} \\
&\overset{\text{(i)}}{\leq} \kappa_{\mathbf{\Gamma}^*}\frac{3}{2}d|||\mathbf{\Delta}|||^2_{\max}\kappa^3_{\mathbf{\Sigma}^*} \\
&\leq \frac{3}{2}d\kappa^3_{\mathbf{\Sigma}^*}\kappa_{\mathbf{\Gamma}^*}r^2,
\end{aligned}$$

where the first inequality follows by applying the bound from Lemma 10. Since $r \leq 1/3d\kappa^3_{\mathbf{\Sigma}^*}\kappa_{\mathbf{\Gamma}^*}$ by assumption (3.57), we conclude that

$$\|(\mathbf{\Gamma}^*_{SS})^{-1}\left\{\mathrm{vec}[((\mathbf{\Theta}^*)^{-1}\mathbf{\Delta})^2\mathbf{J}(\mathbf{\Theta}^*)^{-1}]_S\right\}\|_\infty \leq \frac{3}{2}d\kappa^3_{\mathbf{\Sigma}^*}\kappa_{\mathbf{\Gamma}^*}\frac{1}{3d\kappa^3_{\mathbf{\Sigma}^*}\kappa_{\mathbf{\Gamma}^*}}r \leq \frac{r}{2}.$$

Therefore, putting the pieces together, we can establish that

$$|||\mathbf{\Delta}|||_{\max} = |||\widetilde{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_{\max} \leq 2\kappa_{\mathbf{\Gamma}^*}\left\{|||\mathbf{W}|||_{\max} + \lambda_n\right\}.$$

□

A guarantee on the sign consistency of the primal witness matrix $\widetilde{\mathbf{\Theta}}$ can be achieved by lower bounding the minimum value $\mathbf{\Theta}^*_{\min}$ with a combination of Lemma 11.

Lemma 12. *[Sign Consistency of Oracle Estimator.] Suppose the conditions of Lemma 11 holds and further that the minimum absolute value $\mathbf{\Theta}^*_{\min}$ of non-zero entries in the true concentration matrix $\mathbf{\Theta}^*$ is lower bounded as*

$$| \mathbf{\Theta}^*_{\min} | \geq 4\kappa_{\mathbf{\Gamma}^*}(|||\mathbf{W}|||_{\max} + \lambda_n) \tag{3.65}$$

then $\operatorname{sgn}(\widetilde{\mathbf{\Theta}}_S) = \operatorname{sgn}(\mathbf{\Theta}^*_S)$ *holds.*

Proof. We have, from the bound (3.58), $| \widetilde{\mathbf{\Theta}}_{ij} - \mathbf{\Theta}^*_{ij} | \leq r, \forall (i,j) \in S$. Therefore, combining the definition of r, we can write

$$| \widetilde{\mathbf{\Theta}}_{ij} - \mathbf{\Theta}^*_{ij} | \leq 2\kappa_{\mathbf{\Gamma}^*}(|||\mathbf{W}|||_{\max} + \lambda_n).$$

This yields that for all $(i,j) \in S$, the estimate $\widetilde{\mathbf{\Theta}}_{ij}$ cannot differ enough from $\mathbf{\Theta}^*_{ij}$ to change sign. □

Proof of Theorem 2

Proof. Using the lower bound on sample size n (3.32) and the monotonicity condition (3.13), we can write

$$\begin{aligned}
&\delta = \frac{1}{2\kappa_{\mathbf{\Gamma}^*}\Big(1 + (8/\alpha)\Big)\theta^{-1}_{\min}} \geq \bar{\delta}_{f_*}(n, p^\gamma) \\
\implies &\frac{1}{2\kappa_{\mathbf{\Gamma}^*}\Big(1 + (8/\alpha)\Big)\theta^{-1}_{\min}\bar{\delta}_{f_*}(n, p^\gamma)} \geq 1 \\
\implies &\frac{1}{2\kappa_{\mathbf{\Gamma}^*}\Big(1 + (8/\alpha)\Big)\bar{\delta}_{f_*}(n, p^\gamma)} \geq \theta^{-1}_{\min} \\
\implies &\theta_{\min} \geq 2\kappa_{\mathbf{\Gamma}^*}\Big(1 + (8/\alpha)\Big)\bar{\delta}_{f_*}(n, p^\gamma) > 4\kappa_{\mathbf{\Gamma}^*}\Big(1 + (8/\alpha)\Big)\bar{\delta}_{f_*}(n, p^\gamma).
\end{aligned}$$

Therefore, from Theorem 1 we have the equality $\widetilde{\mathbf{\Theta}} = \widehat{\mathbf{\Theta}}$, and also that $|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_{\max} \leq \theta_{\min}/2$ with probability at least $1 - 1/p^{\gamma-2}$. We can apply Lemma 12 which guarantees that $\operatorname{sgn}(\widetilde{\mathbf{\Theta}}_{ij}) = \operatorname{sgn}(\mathbf{\Theta}^*_{ij})$ for all $(i,j) \in E$. To conclude, we establish that with probability at least $1 - (1/p^{\gamma-2})$, the sign consistency condition $\operatorname{sgn}(\widehat{\mathbf{\Theta}}_{ij}) = \operatorname{sgn}(\mathbf{\Theta}^*_{ij})$ holds for all $(i,j) \in E$. □

Proof of Corollary 1

Proof. Theorem 1 guarantees that with probability at least $1 - (1/p^{\gamma-2})$ that $|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_{\max} \leq 2\kappa_{\mathbf{\Gamma}^*}\Big(1 + 8/\alpha\Big)\bar{\delta}_{f_*}(n, p^\tau)$. Recall that $\mathbf{\Theta}^*$ has at most $s + p$ non-zero elements, where, $s = \mid E(\mathbf{\Theta}^*) \mid$ denotes the total number of off-diagonal non-zeros in $\mathbf{\Theta}^*$ and p is the number of diagonal elements. Since the edge set $\widehat{\mathbf{\Theta}}$ is a subset of the edge set of $\mathbf{\Theta}^*$, we can conclude that

$$\begin{aligned}
|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_F &= \left[\sum_{i=1}^{p}(\widehat{\mathbf{\Theta}}_{ii} - \mathbf{\Theta}^*_{ii})^2 + \sum_{(i,j)\in E}(\widehat{\mathbf{\Theta}}_{ij} - \mathbf{\Theta}^*_{ij})^2\right]^{1/2} \\
&\leq \sqrt{s+p}|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_{\max} \\
&= 2\kappa_{\mathbf{\Gamma}^*}\Big(1 + (8/\alpha)\Big)\bar{\delta}_{f_*}(n, p^\tau)\sqrt{s+p}.
\end{aligned}$$

The inequality follows from the fact that $|||\mathbf{A}|||_F \leq \sqrt{d}|||\mathbf{A}|||_{\max}$, if the matrix $\mathbf{A}$ has d non-zero elements.

Notice that for a symmetric matrix, we have

$$|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_2 \overset{(i)}{\leq} |||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_\infty \overset{(ii)}{\leq} d|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_{\max} \tag{3.66}$$

where the first inequality follows from the equivalence relationship between the l_2 and l_∞-operator norm and the second inequality follows from the fact that for a matrix $\mathbf{A}$, $|||\mathbf{A}|||_\infty \leq d|||\mathbf{A}|||_{\max}$, where d is the maximum number of non-zero elements per row/column.

Since the Frobenius norm upper bounds the operator norm, therefore, as claimed, we have

$$|||\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*|||_2 \leq \left\{2\kappa_{\boldsymbol{\Gamma}^*}\left(1 + (8/\alpha)\right)\right\} \min\left\{\sqrt{s+p}, d\right\} \bar{\delta}_{f_*}(n, p^\tau).$$

□

Chapter 4

Joint Estimation of Regression Coefficients and Precision Matrix in Noisy Data

In Chapter 3, we were only interested to inspect the graphical structure of the p-dimensional vector $X := (X_1, ..., X_p)^\top \in \mathbb{R}^p$. In this Chapter, we introduce a q-dimensional random vector as responses $Y := (Y_1, ..., Y_q)^\top \in \mathbb{R}^q$ and want to inspect the conditional relationship of Y given X in the presence of missing data in the responses.

4.1 Introduction

Consider a q-dimensional Gaussian random vector $Y := (Y_1, ..., Y_q)^\top \in \mathbb{R}^q$ and a p-dimensional deterministic covariate vector $X := (X_1, ..., X_p)^\top \in \mathbb{R}^p$. We assume that Y and X have been centered, thus the intercept term is omitted. Assuming that the underlying random variables Y come from a multivariate Gaussian distribution, we can form a linear relationship between Y and X as

$$Y = \mathbf{B}^{*\top} X + \varepsilon \tag{4.1}$$

where $\mathbf{B}^* \in \mathbb{R}^{p\times q}$ is the matrix of regression coefficients and $\varepsilon \in \mathbb{R}^q$ is the random error assumed to follow a multivariate Gaussian distribution with mean zero and covariance $\mathbf{\Sigma}^*_{\varepsilon\varepsilon} \in \mathbb{R}^{q\times q}$. The model in (4.1) implies that the regression function has the form $\mathbb{E}(Y|X) = \mathbf{B}^{*\top}X$ and also $\mathrm{Cov}(Y|X) = \mathbf{\Sigma}^*_{\varepsilon\varepsilon}$.

Suppose we have n independent and identically distributed observations from some joint distribution of Y and X denoted by $\mathcal{D}(Y, X)$. In matrix notation, we can rewrite (4.1) as a model of n stacked observations

$$\mathbf{Y} = \mathbf{X}\mathbf{B}^* + \boldsymbol{\varepsilon}, \tag{4.2}$$

where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top \in \mathbb{R}^{n\times p}$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^\top \in \mathbb{R}^{n\times q}$ denote the data matrices, and $\boldsymbol{\varepsilon} = [\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n]^\top \in \mathbb{R}^{n\times q}$ denotes the matrix of random noises. We assume that the design matrix $\mathbf{X}$ has normalized columns, that is, $(1/n)\sum_{i=1}^n x_{ij}^2 = 1$ for every $j = 1, \ldots, p$. Typically, the goal is to estimate the coefficient matrix $\mathbf{B}^*$ and the covariance $\mathbf{\Sigma}^*_{\varepsilon\varepsilon}$. Before stating the ways to solve this problem, let us define some notations that we will use throughout the chapter.

Notation and Conventions: In this chapter, we will denote a random variable by an uppercase letter and an data matrix with uppercase bold letter, for example, A is a random variable and $\mathbf{A}$ is a data matrix.

For a matrix $\mathbf{A}$, we denote by $\mathbf{A} \succeq 0$ when $\mathbf{A}$ is positive semi-definite. Let $|||\mathbf{A}|||_1$ be the operator norm induced by ℓ_1 norm for vectors, which can be computed by $|||\mathbf{A}|||_1 = \max_j \sum_i \mid a_{ij} \mid$, i.e. the maximum absolute column sum of the matrix. Denoted by $|||\mathbf{A}|||_2$ the operator norm that can be computed as the greatest singular value of $\mathbf{A}$, i.e. $|||\mathbf{A}|||_2 = \max_j \sigma_j(\mathbf{A})$. Let $|||\mathbf{A}|||_\infty$ be the operator norm induced by ℓ_∞ norm, which can be computed by $|||\mathbf{A}|||_\infty = \max_i \sum_j \mid a_{ij} \mid$, i.e. the maximum absolute row sum of the matrix. We also introduced the following elementwise matrix norm. Let $|||\mathbf{A}|||_{1,1} = \sum_{i,j} \mid a_{ij} \mid$ be the elementwise ℓ_1-norm, $|||\mathbf{A}|||_F = \sqrt{\sum_{i,j} \mid a_{ij} \mid^2}$ be the Frobenius norm, and $|||\mathbf{A}|||_{\max} =$

$\max_{i,j} \mid a_{ij} \mid$ be the elementwise maximum norm. Let $\Lambda_{\min}(\mathbf{A})$ and $\Lambda_{\max}(\mathbf{A})$ denote the smallest and largest eigenvalues of $\mathbf{A}$.

For two matrices $\mathbf{A} = (a_{ij})$ and $\mathbf{B} = (b_{ij})$, we define $\mathbf{A} \odot \mathbf{B} = (a_{ij}b_{ij})$ as their elementwise product, and $\mathbf{A} \oslash \mathbf{B} = (a_{ij}/b_{ij})$ as their elementwise division. For any $a \in \mathbb{N} = \{1, 2, \ldots\}$, we denote $[a] = \{1, \ldots, a\}$ as sample indices. For instance, if I is an index set, we write, $I \subset [q]$ or $I \subset [p] \times [q]$. If $\mathbf{a} \in \mathbb{R}^p$ and if $I \subset [p]$, we use $\mathbf{a}_I$ to denote the same vector as $\mathbf{a}$ but with elements $[p] \backslash I$ set to zero. We denote the (k, l)th element of a matrix $\mathbf{M}$ by $(\mathbf{M})_{kl}$, its kth row by $(\mathbf{M})_{k\bullet}$ and lth column by $(\mathbf{M})_{\bullet l}$.

Let $\mathbb{Q}$ be a generic Euclidean space. Let $\mathbf{M}$ and $\mathbf{N}$ be any conformable matrices or vectors in $\mathbb{Q}$. Let us define the inner product $\langle \cdot, \cdot \rangle$ as $\langle \mathbf{M}, \mathbf{N} \rangle = \mathrm{tr}(\mathbf{M}^\top \mathbf{N})$. For a norm $\mathcal{R}$ defined on $\mathbb{Q}$, the dual norm $\mathcal{R}^*$ can be defined by

$$\mathcal{R}^*(\mathbf{M}) \equiv \sup_{\mathbf{N} \in \mathbb{Q} \backslash \{0\}} \frac{\langle \mathbf{M}, \mathbf{N} \rangle}{\mathcal{R}(\mathbf{N})}.$$

4.1.1 Model Setup

To jointly estimate $\mathbf{B}^*$ and $\mathbf{\Sigma}^*_{\varepsilon\varepsilon}$, Rothman et al. (2010) proposed a method called multivariate regression with covariance estimation (MRCE) to estimate $\mathbf{B}^*$ and $\mathbf{\Theta}^*_{\varepsilon\varepsilon} := (\mathbf{\Sigma}^*_{\varepsilon\varepsilon})^{-1}$ by minimizing the negative log-likelihood with ℓ_1 penalization as follows:

$$(\widehat{\mathbf{\Theta}}, \widehat{\mathbf{B}}) = \underset{\mathbf{\Theta} \succeq 0, \mathbf{B}}{\arg\min}\, \mathrm{tr}\left[\frac{1}{2n}(\mathbf{Y} - \mathbf{XB})^\top (\mathbf{Y} - \mathbf{XB})\mathbf{\Theta}\right] - \frac{1}{2}\log\det(\mathbf{\Theta}) + \lambda_{\mathbf{\Theta}} |\!|\!| \mathbf{\Theta} |\!|\!|_{1,\text{off}} + \lambda_{\mathbf{B}} |\!|\!| \mathbf{B} |\!|\!|_{1,1} \tag{4.3}$$

where $|\!|\!| \mathbf{\Theta} |\!|\!|_{1,\text{off}} = \sum_{j' \neq j} |\Theta_{jj'}|$, $|\!|\!| \mathbf{B} |\!|\!|_{1,1} = \sum_{j,k} |B_{jk}|$, and $\lambda_{\mathbf{\Theta}}, \lambda_{\mathbf{B}} \geq 0$ are tuning parameters controlling the sparsity in $\widehat{\mathbf{\Theta}}$ and $\widehat{\mathbf{B}}$, respectively. In the case of fully observed data, Equation (4.3) can be efficiently solved when the error vector ε is uncorrelated. When $\mathbf{\Sigma}^*_{\varepsilon\varepsilon}$ is assumed to be diagonal, that is, all off-diagonal elements of $\mathbf{\Sigma}^*_{\varepsilon\varepsilon}$ are zeros, the objective function in (4.3) can be solved by a series of separate standard Lasso problems, equivalent to performing q separate penalized least square regressions. Rothman et al. (2010) showed

that the estimation accuracy can be improved by exploiting the additional information when the response variables are correlated.

When there are many predictors and responses, prediction with a multivariate regression model becomes challenging since it is required to estimate pq parameters. When the responses are correlated, the assumptions of sparsity on both $\mathbf{B}^*$ and the off-diagonal elements of $\mathbf{\Theta}^*_{\varepsilon\varepsilon}$ becomes necessary to be able to estimate the additional $\mathcal{O}(q^2)$ parameters in $\mathbf{\Theta}^*_{\varepsilon\varepsilon}$. When $n \ll p$ or $s \equiv |S| \ll pq$, where $S = \{(j,k)|B^*_{jk} \neq 0\}$ is the support of $\mathbf{B}^*$, assuming sparsity in $\mathbf{B}^*$ and $\mathbf{\Theta}^*_{\varepsilon\varepsilon}$ considerably reduces variability in the estimation. Under the multivariate normal assumption, the precision matrix has the interpretation of a conditional Gaussian graphical model (Lauritzen, 1996), since a zero off-diagonal element implies conditional independence among the covariates. Since $\mathbf{\Theta}^*_{\varepsilon\varepsilon}$ captures the conditional dependencies among the response variables, the resulting network structure becomes highly interpretable.

The standard methods proposed in the literature to solve the problem in (4.3) are established when the data are fully observed. A detailed review of those works are presented in Section 2.3 of Chapter 2. However, in practical applications, the data may be corrupted or missing such that the responses or the covariates are only partially observed. A naive way to handle missing data could be to delete all the cases that contain missing values either listwise or pairwise and work with the complete cases only. However, that would result into decreased statistical power, substantial information loss and may lead to biased estimates when the data are not missing completely at random (MCAR).

Other ad hoc imputation based methods are also available in literature where the missing observations are imputed by the corresponding mean along with more systematic approaches based on likelihoods (Little and Rubin, 2019; Schafer, 1997). Städler and Bühlmann (2012) developed an Expectation Maximization (EM)-based method for sparse inverse covariance matrix estimation, which can be used in missing data scenarios. The challenge with classical approaches is that they do not often scale to high-dimensional

problems and it becomes difficult to provide theoretical guarantees to their algorithmic counterparts. Errors-in-variables regressions (Hwang, 1986; Xu and You, 2007) had been extensively studied in literature. Loh and Wainwright (2012) studied the case when the observed covariates are corrupted in univariate regression ($q = 1$) under high-dimensional settings when $p \gg n$. They proposed an unbiased surrogate estimate of the sample covariance matrix based on Xu and You (2007) and provided a non-convex solution proving that under some restricted eigenvalue conditions and deviation bounds, the projected gradient descent method converges to a near-global minimizer. One criticism of such a non-convex solution is that it depends on an additional side constraints that requires knowledge on some unknown constants a priori that we actually want to estimate. The convergence rates and the computational results depend on this assumption and therefore, it makes this approach difficult to use in practice.

Datta and Zou (2017) proposed the convex conditioned Lasso (CoCoLasso) by defining a nearest positive semi-definite matrix projection operator for square matrix that makes the underlying optimization problem to be convex. It is well known that there are theoretical and computational benefits of convexity, which makes this approach more lucrative. Unlike the non-convex approach by Loh and Wainwright (2012), the projected surrogate estimator of the sample covariance matrix is guaranteed to be positive semi-definite, thereby ensuring the convexity of the problem.

4.1.2 Unbiased Surrogate Estimators with Corrupted Responses

In this work, we extend the results of Loh and Wainwright (2012) to the case of multiple responses but using the projected surrogate estimates of the sample covariance matrix as proposed by Datta and Zou (2017) instead of a non-convex approach. We propose to solve (4.3) when the data are not fully observed and contains missing data. We propose a three step estimation to the problem defined in (4.3) when the data are corrupted and may be missing completely at random (MCAR) in a high-dimensional setting.

Multiplicative noise

In this section we establish the general formulation of the model for missing data, since missing data are a special case of multiplicative error. Let $\mathbf{Y} \in \mathbb{R}^{n\times q}$ be the unobserved response variables and $\mathbf{W} \in \mathbb{R}^{n\times q}$ is a noise matrix with (i,j)th element $w_{ij} \geq 0$, so that we have an observed response data matrix $\mathbf{Z} = \mathbf{Y} \odot \mathbf{W}$. We assume that the rows $w_{i\bullet}$ of $\mathbf{W}$ are drawn independently and identically from some multivariate distribution having strictly positive entries in the first and the second-order expectations $\boldsymbol{\mu}_W = \mathbb{E}[W] \in \mathbb{R}^q$ and $\mathbb{E}[WW^\top] \in \mathbb{R}^{q\times q}$. For simplicity, we assume that the mean and covariance of W are known or can be estimated from the data.

We can expand the quadratic terms and rewrite (4.3) as

$$(\widehat{\boldsymbol{\Theta}}, \widehat{\mathbf{B}}) = \underset{\boldsymbol{\Theta}\succeq 0, \mathbf{B}}{\arg\min}\ \operatorname{tr}\left[\frac{1}{2}(\mathbf{S}_{yy} - 2\widehat{\mathbf{S}}_{xy}^\top \mathbf{B} + \mathbf{B}^\top \mathbf{S}_{xx}\mathbf{B})\boldsymbol{\Theta}\right] - \frac{1}{2}\log\det(\boldsymbol{\Theta}) + \lambda_{\boldsymbol{\Theta}}|||\boldsymbol{\Theta}|||_{1,\text{off}} + \lambda_B|||\mathbf{B}|||_{1,1}, \tag{4.4}$$

where $\mathbf{S}_{yy} = \frac{1}{n}\mathbf{Y}^\top\mathbf{Y}$, $\mathbf{S}_{xy} = \frac{1}{n}\mathbf{X}^\top\mathbf{Y}$ and $\mathbf{S}_{xx} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$ are empirical covariance matrices, which are unbiased estimators of $\boldsymbol{\Sigma}^*_{yy} = \mathbb{E}[YY^\top]$, $\boldsymbol{\Sigma}^*_{xy} = \mathbb{E}[XY^\top]$ and $\boldsymbol{\Sigma}^*_{xx} = \mathbb{E}[XX^\top]$, respectively. Since we do not observe the clean data matrix $\mathbf{Y}$ and only observe the corrupted version $\mathbf{Z}$ of the response matrix, we cannot directly estimate $\boldsymbol{\Theta}^*_{\varepsilon\varepsilon}$ and $\mathbf{B}^*$ using (4.4) because $\mathbf{S}_{yy}$ and $\mathbf{S}_{xy}$ are biased estimates. Following Loh and Wainwright (2012) we can calculate the alternative unbiased surrogate estimator $\widehat{\mathbf{S}}_{yy}$ and $\widehat{\mathbf{S}}_{xy}$ by calculating the sufficient statistics:

$$\widehat{\mathbf{S}}_{yy} := \frac{1}{n}\mathbf{Z}^\top\mathbf{Z} \oslash \mathbb{E}[WW^\top],\ \ \widehat{\mathbf{S}}_{xy} := \frac{1}{n}\mathbf{X}^\top\mathbf{Z} \oslash [\mathbb{E}[W], ..., \mathbb{E}[W]]^\top. \tag{4.5}$$

Then the estimates of $\mathbf{\Theta}^*_{\varepsilon\varepsilon}$ and $\mathbf{B}^*$ can be obtained by solving the following optimization problem:

$$(\widehat{\mathbf{\Theta}}, \widehat{\mathbf{B}}) = \underset{\mathbf{\Theta}\succeq 0, \mathbf{B}}{\arg\min}\ \mathrm{tr}[\frac{1}{2}(\widehat{\mathbf{S}}_{yy} - 2\widehat{\mathbf{S}}_{xy}^{\top}\mathbf{B} + \mathbf{B}^{\top}\mathbf{S}_{xx}\mathbf{B})\mathbf{\Theta}] - \frac{1}{2}\log\det(\mathbf{\Theta}) + \lambda_{\mathbf{\Theta}}|||\mathbf{\Theta}|||_{1,\mathrm{off}} + \lambda_B|||\mathbf{B}|||_{1,1}. \tag{4.6}$$

To incorporate the missing data case in the objective function of the fully observed case (4.4), the empirical covariance matrices $\mathbf{S}_{yy}$ and $\mathbf{S}_{xy}$ have been replaced with their unbiased surrogates $\widehat{\mathbf{S}}_{yy}$ and $\widehat{\mathbf{S}}_{xy}$. Notice that $\mathbf{S}_{xx}$ remains unchanged since we are assuming that the covariate matrix $\mathbf{X}$ is fully observed.

Missing data

Since the missing data case is a special case of the multiplicative measurement errors, we construct a missing complete at random scenario where the entries w_{ij} of $\mathbf{W}$ are assumed to be independent Bernoulli $(1-\rho_j), \forall j = 1, ..., q$ random variables with values

$$w_{ij} = \begin{cases} 1 & \text{with probability } 1-\rho_j, \\ 0 & \text{otherwise.} \end{cases} \tag{4.7}$$

That is, each element of $\mathbf{Y}$ in the jth column has probability ρ_j of being missing. We observe $z_{ij} = y_{ij}$ with probability $1-\rho_j$ and zero otherwise. Under the missing completely at random assumption, $\mathbb{E}[W]$ and $\mathbb{E}[WW^{\top}]$ have the following specific forms:

$$\mathbb{E}[W] = [(1-\rho_1), \cdots, (1-\rho_q)]^{\top}, \ \ \mathbb{E}[WW^{\top}]_{ij} = \begin{cases} (1-\rho_i)(1-\rho_j) & \text{if } i \neq j, \\ (1-\rho_i) & \text{if } i = j. \end{cases} \tag{4.8}$$

The problem reduces to the standard MRCE model (Rothman et al., 2010) for fully observed data when $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_q)^{\top} = \mathbf{0}$. However, in practice, $\boldsymbol{\rho}$ may not be known and must be

estimated empirically from the data. Loh and Wainwright (2012) suggested estimating ρ_j using $\widehat{\rho}_j$, where $\widehat{\rho}_j$ is the empirical missing probability of the jth column.

4.2 Estimation

4.2.1 First Stage Estimation of the Coefficient Matrix B^*

In the first stage, we assume that there is no correlation among the response variables, that is, we assume the precision matrix to be an identity matrix ($\mathbf{\Theta} = \mathbf{I}$). This assumption lets us to perform a column-by-column estimation of $\mathbf{B}^*$. Assuming that $\mathbf{B}^*$ is elementwise sparse, we obtain a preliminary estimator $\widehat{\mathbf{B}}^{(1)}$ by regularizing the least squares problem with ℓ_1 penalty:

$$\widehat{\mathbf{B}}^{(1)} = \arg\min_{\mathbf{B}} \operatorname{tr}(\mathbf{B}^\top \mathbf{S}_{xx}\mathbf{B}/2 - \widehat{\mathbf{S}}_{xy}^\top \mathbf{B}) + \lambda_{\mathbf{B}} |\!|\!| \mathbf{B} |\!|\!|_{1,1}, \tag{4.9}$$

where $\widehat{\mathbf{S}}_{xy}$ is an unbiased estimator of $\mathbb{E}[XY^\top]$ defined in (4.5). This loss function is convex because $\mathbf{X}$ is fully observed and therefore $\mathbf{S}_{xx}$ is positive semi-definite. Define $\widehat{\mathbf{B}}^{(1)} = [\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_q]$ with lth column as $\widehat{\boldsymbol{\beta}}_l = (\widehat{\boldsymbol{\beta}}_{l1}, \ldots, \widehat{\boldsymbol{\beta}}_{lp})^\top \in \mathbb{R}^p$. We can compute each column of $\widehat{\mathbf{B}}^{(1)}$, by considering (4.9) as a column-by column solution of multiple penalized least squares problems with univariate response.

Let us define the least square Lasso loss function $\mathcal{L}_l : \mathbb{R}^p \to \mathbb{R}$ for each $l = 1, \ldots, q$, i.e. for the lth column of $\mathbf{B}^*$ by setting for arbitrary $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$\widehat{\boldsymbol{\beta}}_l = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \mathcal{L}_l(\boldsymbol{\beta}) + \lambda_l \mathcal{R}(\boldsymbol{\beta}), \tag{4.10}$$

where $\mathcal{L}_l(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \mathbf{S}_{xx}\boldsymbol{\beta}/2 - \left\{(\widehat{\mathbf{S}}_{xy})_{\bullet l}\right\}^\top \boldsymbol{\beta}$ where $(\widehat{\mathbf{S}}_{xy})_{\bullet l}$ denotes lth column of $\widehat{\mathbf{S}}_{xy}$. and $\mathcal{R} : \mathbb{R}^p \to \mathbb{R}$ with $\mathcal{R}(\cdot) = \|\cdot\|_1$ is the penalty function. Hence the problem can be solved by various efficient algorithms for the standard Lasso.

4.2.2 Estimation of the Precision Matrix $\boldsymbol{\Theta}^*_{\varepsilon\varepsilon}$

In the second stage of the estimation, we estimate $\boldsymbol{\Theta}^*_{\varepsilon\varepsilon}$ by the solution to the graphical Lasso problem

$$\underset{\boldsymbol{\Theta}\in\mathbb{R}^{q\times q}:\boldsymbol{\Theta}\succeq\mathbf{0}}{\arg\min}\left\{\langle\boldsymbol{\Theta},\widehat{\mathbf{S}}_{\varepsilon\varepsilon}\rangle-\log\det(\boldsymbol{\Theta})+\lambda_{\boldsymbol{\Theta}}|||\boldsymbol{\Theta}|||_{1,\text{off}}\right\}, \tag{4.11}$$

where $\widehat{\mathbf{S}}_{\varepsilon\varepsilon}\coloneqq\widehat{\mathbf{S}}_{yy}-\widehat{\mathbf{B}}^{(1)\top}\mathbf{S}_{xx}\widehat{\mathbf{B}}^{(1)}$ is a plug-in estimate of the covariance, in which $\widehat{\mathbf{S}}_{yy}$ is an unbiased surrogate estimator of $\mathbb{E}[YY^\top]$ and $\widehat{\mathbf{B}}^{(1)}$ is estimated in the first step. This should lead to an improved estimate of the regression coefficients in the third stage of the estimation since we take into account the dependence structure of the precision matrix.

Since we are interested to estimate the precision for the responses which is prone to having missing data, as discussed in Chapter 3, the objective function may no longer be convex since the input estimator of the covariance matrix $\widehat{\mathbf{S}}_{\varepsilon\varepsilon}$ may not be positive semi-definite. We can illustrate this phenomenon for a specific situation. Suppose that all columns of responses have the same probability $\boldsymbol{\rho}\in[0,1]$ of being missing, hence $\widehat{\mathbf{S}}_{yy}$ has a certain form so that $\widehat{\mathbf{S}}_{yy}=\widetilde{\mathbf{Z}}^\top\widetilde{\mathbf{Z}}/n-\rho_j\text{diag}(\widetilde{\mathbf{Z}}^\top\widetilde{\mathbf{Z}}/n)$, in which $\tilde{z}_{ij}=z_{ij}/(1-\rho_j)$. In fact, when $n\ll q$, $\widehat{\mathbf{S}}_{yy}$ is guaranteed to have a large number of negative eigenvalues even if under a moderate missingness, given the fact that $\widetilde{\mathbf{Z}}^\top\widetilde{\mathbf{Z}}/n$ has rank at most n, so the subtracted diagonal matrix may cause $\widehat{\mathbf{S}}_{yy}$ to have negative eigenvalues. As a result, a non-positive semi-definite estimate of the covariance matrix $\widehat{\mathbf{S}}_{\varepsilon\varepsilon}$ makes the the objective function in (4.11) non-convex and unbounded from below.

To avoid the non-convex objective function and to ensure the positive semi-definiteness of the covariance matrix, we can easily project $\widehat{\mathbf{S}}_{\varepsilon\varepsilon}$ onto the semi-definite cone (Datta and Zou, 2017) to produce an update $\widetilde{\mathbf{S}}_{\varepsilon\varepsilon}$ and substitute for $\widehat{\mathbf{S}}_{\varepsilon\varepsilon}$ in (4.11). We can define a nearest positive semi-definite matrix projection operator for any square matrix $\widehat{\mathbf{S}}_{\varepsilon\varepsilon}$ that

$$\widetilde{\mathbf{S}}_{\varepsilon\varepsilon}:=\underset{\mathbf{K}\succeq 0}{\text{argmin}}\,|||\widehat{\mathbf{S}}_{\varepsilon\varepsilon}-\mathbf{K}|||_{\max}, \tag{4.12}$$

where $\mathbf{K}$ is a positive semi-definite matrix and $|||\cdot|||_{\max}$ is the elementwise maximum norm. As established in Chapter 3, this is nothing but the CoGlasso estimate and can be obtained as a solution to the objective function

$$\widehat{\boldsymbol{\Theta}}_{\varepsilon\varepsilon} = \underset{\boldsymbol{\Theta}\in\mathbb{R}^{q\times q}:\boldsymbol{\Theta}\succeq\mathbf{0}}{\arg\min} \left\{ \langle \boldsymbol{\Theta}, \widetilde{\mathbf{S}}_{\varepsilon\varepsilon} \rangle - \log\det(\boldsymbol{\Theta}) + \lambda_{\boldsymbol{\Theta}} |||\boldsymbol{\Theta}|||_{1,\text{off}} \right\} \tag{4.13}$$

where $\widetilde{\mathbf{S}}_{\varepsilon\varepsilon} := \widehat{\mathbf{S}}_{yy} - \widehat{\mathbf{B}}^{(1)\top}\mathbf{S}_{xx}\widehat{\mathbf{B}}^{(1)}$. Since $\mathbf{S}_{xx}$ is positive semi-definite by construction, therefore $\widetilde{\mathbf{S}}_{\varepsilon\varepsilon}$ is guaranteed to be positive semi-definite by definition. The problem in (4.13) can be solved using our CoGlasso algorithm.

4.2.3 Second Stage Estimation of the Coefficient Matrix $\mathbf{B}^*$

In the last stage of estimation, we utilize the estimated precision matrix $\widehat{\boldsymbol{\Theta}}_{\varepsilon\varepsilon}$ in the previous stage to get a refined estimate of $\mathbf{B}^*$. Solving for $\widehat{\mathbf{B}}^{(2)}$ with a fixed plug-in estimate $\widehat{\boldsymbol{\Theta}}_{\varepsilon\varepsilon}$ is equivalent to finding the solution to the objective function

$$\widehat{\mathbf{B}}^{(2)} = \underset{\mathbf{B}}{\arg\min}\ \text{tr}[(\mathbf{B}^\top\mathbf{S}_{xx}\mathbf{B}/2 - \widehat{\mathbf{S}}_{xy}^\top\mathbf{B})\widehat{\boldsymbol{\Theta}}_{\varepsilon\varepsilon}] + \lambda_{\mathbf{B}}|||\mathbf{B}|||_{1,1}. \tag{4.14}$$

The problem in (4.14) can be solved using a proximal gradient descent algorithm, specifically an iterative soft-thresholding algorithm (ISTA) in this particular case of an ℓ_1-penalty.

4.3 Theoretical Properties

4.3.1 Recovery Rate for $\widehat{\mathbf{B}}^{(1)}$

The following assumption is imposed on the population covariance matrix to mildly control the error of the lasso solution. Under unfavorable settings, where the loss function is flat around its minimizer, it is not necessarily true that a small loss difference implies a small error. Especially, in high-dimensional settings, we can only hope to obtain some

form of restricted curvature of the loss function in certain directions, specifically, along the cone set $\mathbb{C}(S_l)$ defined below, under a sufficiently large sample size.

Assumption 1 (Restricted eigenvalue condition.)**.** *Define the cone set*

$$\mathbb{C}(S_l) = \left\{ \boldsymbol{\delta} \in \mathbb{R}^p : \|(\boldsymbol{\delta})_{S_l^c}\|_1 \leq 3\|(\boldsymbol{\delta})_{S_l}\|_1 \right\}.$$

We assume the following restricted eigenvalue condition (Page 208, Wainwright, 2019) for $\mathbf{S}_{xx}$ *over* $\mathbb{C}(S_l)$*,*

$$0 < \kappa_l = \min_{\boldsymbol{\delta} \neq 0,\ \boldsymbol{\delta} \in \mathbb{C}(S_l)} \frac{\boldsymbol{\delta}^\top \mathbf{S}_{xx} \boldsymbol{\delta}}{\|\boldsymbol{\delta}\|_2^2}.$$

We would require the following Lemmas to prove Proposition 1. Without loss of generality, we can write the lth column of the true regression coefficient matrix as $\boldsymbol{\beta}_l^* = (\boldsymbol{\beta}_{S_l}^{*\top}, \mathbf{0}^\top)^\top$ and the corresponding $\mathbf{X} = ((\mathbf{X})_{\bullet S_l}, (\mathbf{X})_{\bullet S_l^c})$. Hence, the true model for the lth column of $\mathbf{Y}$ can be written as $(\mathbf{Y})_{\bullet l} = (\mathbf{X})_{\bullet S_l} \boldsymbol{\beta}_{S_l}^* + (\boldsymbol{\varepsilon})_{\bullet l}$.

Lemma 13. *Assume that each row of the error matrix* $\boldsymbol{\varepsilon} \in \mathbb{R}^{n \times q}$ *and each row of multiplicative error matrix* $\mathbf{W} \in \mathbb{R}^{n \times q}$ *are two sub-Gaussian random vectors where each elements of the vectors follow the sub-Gaussian distribution with parameters* σ_ε^2 *and* σ_W^2 *respectively. Under this assumption, the elementwise max norm of the deviation between* $\widehat{\mathbf{S}}_{xy}$ *and* $\mathbf{S}_{xy}$ *satisfies the following probability bound for any* $t \leq t_0^{(1)}$ *with* $t_0^{(1)} = \sigma_\varepsilon \sigma_W X_{\max} / \mu_{\min}$*,*

$$\Pr[|||\widehat{\mathbf{S}}_{xy} - \mathbf{S}_{xy}|||_{\max} \geq t] \leq pqC \exp\left(-\frac{cn\mu_{\min}^2 t^2}{\sigma_W^2 X_{\max}^2 \max(s_{\max}^2 X_{\max}^2 B_{\max}^2, \sigma_\varepsilon^2)} \right),$$

where $\mu_{\min} = \min_j(1-\rho_j) > 0$*,* $X_{\max} = \max_{i,k} |X_{ik}| < \infty$*,* $B_{\max} = \max_{k,j} |\beta_{kj}^*|$*,* $s_{\max} = \max_j s_j$*, for* $j \in \{1, \ldots, q\}$*. Let* $(\widehat{\mathbf{S}}_{xy})_{\bullet l}$ *and* $(\mathbf{S}_{xy})_{\bullet l}$ *be the* l*th columns of* $\widehat{\mathbf{S}}_{xy}$ *and* $\mathbf{S}_{xy}$*, respectively, then we have*

$$\Pr[\|(\widehat{\mathbf{S}}_{xy})_{\bullet l} - (\mathbf{S}_{xy})_{\bullet l}\|_\infty \geq t] \leq pC \exp\left(-\frac{cn\mu_{\min}^2 t^2}{\sigma_W^2 X_{\max}^2 \max(s_{\max}^2 X_{\max}^2 B_{\max}^2, \sigma_\varepsilon^2)} \right),$$

Lemma 14. *Assume that each row of the error matrix $\boldsymbol{\varepsilon} \in \mathbb{R}^{n\times q}$ is a sub-Gaussian random vector where each element of the vector follows the sub-Gaussian distribution with parameter σ_ε^2. Under this assumption, the elementwise max norm of the deviation between $\mathbf{S}_{xy}$ and $\mathbf{S}_{xx}\mathbf{B}^*$ satisfies the following probability bound*

$$\Pr(|||\mathbf{S}_{xy} - \mathbf{S}_{xx}\mathbf{B}^*|||_{\max} \geq t) \leq pqC\exp\left[-\frac{cnt^2}{\sigma_\varepsilon^2 X_{\max}^2}\right]$$

and

$$\Pr(\|(\mathbf{S}_{xy})_{\bullet l} - \mathbf{S}_{xx}\boldsymbol{\beta}_l^*\|_\infty \geq t) \leq pC\exp\left[-\frac{cnt^2}{\sigma_\varepsilon^2 X_{\max}^2}\right]$$

where $(\mathbf{S}_{xy})_{\bullet l}$ and $\boldsymbol{\beta}_l^*$ are the lth columns of $\mathbf{S}_{xy}$ and $\mathbf{B}^*$, respectively, and $X_{\max} = \max_{i,k}|X_{ik}| < \infty$.

Proposition 1. *We assume that the lth column of the true coefficient matrix from* (4.1)*, $\boldsymbol{\beta}_l^*$ has support $S_l \subseteq \{1,\ldots,p\}$ with cardinality $s_l \coloneqq |S_l|$, meaning that $\boldsymbol{\beta}_{lj}^* = 0$ for all $j \in S_l^c$, where S_l^c denotes the complement of S_l. Let us consider that Assumption 1 on the parameter $\kappa_l > 0$ hold. We assume that the tuning parameter λ_l in* (4.10) *satisfies*

$$\lambda_l \geq 2X_{\max}\max\left[\sigma_W s_{\max} X_{\max} B_{\max}/\mu_{\min}, \sigma_W\sigma_\varepsilon/\mu_{\min}, \sigma_\varepsilon\right]\sqrt{\frac{\log p}{n}}. \tag{4.15}$$

Then any estimate $\widehat{\boldsymbol{\beta}}_l$ from (4.10) *satisfy the following bounds*

$$\|\widehat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l^*\|_2 \leq 3\sqrt{s_l}\lambda_l/\kappa_l, \qquad \textit{and} \qquad \|\widehat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l^*\|_1 \leq 4\sqrt{s_l}\|\widehat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l^*\|_2 = 12s_l\lambda_l/\kappa_l$$

with a probability at least $1 - C\exp(-c\log p)$.

Proof. From the definition of (4.10), we have

$$\begin{aligned}
&\mathcal{L}(\widehat{\boldsymbol{\beta}}_l) - \mathcal{L}(\boldsymbol{\beta}_l^*) \\
&= \frac{\widehat{\boldsymbol{\beta}}_l^\top \mathbf{S}_{xx}\widehat{\boldsymbol{\beta}}_l}{2} - \frac{\boldsymbol{\beta}_l^{*\top}\mathbf{S}_{xx}\boldsymbol{\beta}_l^*}{2} - \left\{(\widehat{\mathbf{S}}_{xy})_{\bullet l}\right\}^\top(\widehat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l^*) - (\widehat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l^*)^\top\mathbf{S}_{xx}\boldsymbol{\beta}_l^* + (\widehat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l^*)^\top\mathbf{S}_{xx}\boldsymbol{\beta}_l^*
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\widehat{\boldsymbol{\beta}}_l^\top \mathbf{S}_{xx} \widehat{\boldsymbol{\beta}}_l}{2} + \frac{\boldsymbol{\beta}_l^{*\top} \mathbf{S}_{xx} \boldsymbol{\beta}_l^*}{2} - \widehat{\boldsymbol{\beta}}_l^\top \mathbf{S}_{xx} \boldsymbol{\beta}_l^* - \left\{(\widehat{\mathbf{S}}_{xy})_{\bullet l}\right\}^\top (\widehat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l^*) + (\widehat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l^*)^\top \mathbf{S}_{xx} \boldsymbol{\beta}_l^* \\
&= \frac{1}{2}(\widehat{\boldsymbol{\beta}}_l^\top \mathbf{S}_{xx} \widehat{\boldsymbol{\beta}}_l + \boldsymbol{\beta}_l^{*\top} \mathbf{S}_{xx} \boldsymbol{\beta}_l^* - 2\widehat{\boldsymbol{\beta}}_l^\top \mathbf{S}_{xx} \boldsymbol{\beta}_l^*) - ((\widehat{\mathbf{S}}_{xy})_{\bullet l} - \mathbf{S}_{xx}\boldsymbol{\beta}_l^*)^\top (\widehat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l^*) \\
&\overset{\text{(i)}}{=} \frac{1}{2}[(\widehat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l^*)^\top \mathbf{S}_{xx} (\widehat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l^*)] - ((\widehat{\mathbf{S}}_{xy})_{\bullet l} - \mathbf{S}_{xx}\boldsymbol{\beta}_l^*)^\top (\widehat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l^*) \\
&= \frac{1}{2}[\boldsymbol{\delta}^\top \mathbf{S}_{xx} \boldsymbol{\delta}] - ((\widehat{\mathbf{S}}_{xy})_{\bullet l} - \mathbf{S}_{xx}\boldsymbol{\beta}_l^*)^\top \boldsymbol{\delta}
\end{aligned}
$$

where $\boldsymbol{\delta} = \widehat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l^*$ and equality (i) follows by completing the square. Since $\widehat{\boldsymbol{\beta}}_l$ is the solution of (4.10)

$$
\mathcal{L}(\widehat{\boldsymbol{\beta}}_l) + \lambda_l \|\widehat{\boldsymbol{\beta}}_l\|_1 \leq \mathcal{L}(\boldsymbol{\beta}_l^*) + \lambda_l \|\boldsymbol{\beta}_l^*\|_1.
$$

Plugging in $\mathcal{L}(\widehat{\boldsymbol{\beta}}_l) - \mathcal{L}(\boldsymbol{\beta}_l^*)$ yields

$$
\begin{aligned}
\frac{1}{2}[\boldsymbol{\delta}^\top \mathbf{S}_{xx} \boldsymbol{\delta}] + \lambda_l \|\widehat{\boldsymbol{\beta}}_l\|_1 &\leq ((\widehat{\mathbf{S}}_{xy})_{\bullet l} - \mathbf{S}_{xx}\boldsymbol{\beta}_l^*)^\top \boldsymbol{\delta} + \lambda_l \|\boldsymbol{\beta}_l^*\|_1 \\
&\leq \|\boldsymbol{\delta}\|_1 \|(\widehat{\mathbf{S}}_{xy})_{\bullet l} - \mathbf{S}_{xx}\boldsymbol{\beta}_l^*\|_\infty + \lambda_l \|\boldsymbol{\beta}_l^*\|_1 \qquad (4.16)
\end{aligned}
$$

The second inequality follows from Hölder's inequality. In order to obtain an upper bound for the left-hand side, we first bound the quantity $\|(\widehat{\mathbf{S}}_{xy})_{\bullet l} - \mathbf{S}_{xx}\boldsymbol{\beta}_l^*\|_\infty$. Using the triangular inequality, we get

$$
\|(\widehat{\mathbf{S}}_{xy})_{\bullet l} - \mathbf{S}_{xx}\boldsymbol{\beta}_l^*\|_\infty \leq \|(\widehat{\mathbf{S}}_{xy})_{\bullet l} - (\mathbf{S}_{xy})_{\bullet l}\|_\infty + \|(\mathbf{S}_{xy})_{\bullet l} - \mathbf{S}_{xx}\boldsymbol{\beta}_l^*\|_\infty.
$$

The first term can be bounded by applying Lemma 13 by setting $t = \lambda_l/4$. We see that for $\lambda_l \leq 4t_0^{(1)}$, we have

$$
\Pr[\|(\widehat{\mathbf{S}}_{xy})_{\bullet l} - (\mathbf{S}_{xy})_{\bullet l}\|_\infty \geq \lambda_l/4] \leq pC \exp\left(-\frac{cn\mu_{\min}^2 \lambda_l^2}{\sigma_W^2 X_{\max}^2 \max(s_{\max}^2 X_{\max}^2 B_{\max}^2, \sigma_\varepsilon^2)}\right).
$$

The second term can be bounded by applying Lemma 14 by setting $t = \lambda_l/4$

$$
\Pr(\|(\mathbf{S}_{xy})_{\bullet l} - \mathbf{S}_{xx}\boldsymbol{\beta}_l^*\|_\infty \geq \lambda_l/4) \leq pC \exp\left[-\frac{cn\lambda_l^2}{\sigma_\varepsilon^2 X_{\max}^2}\right].
$$

Define the event $\mathcal{E}_1 = \{\|(\widehat{\mathbf{S}}_{xy})_{\bullet l} - (\mathbf{S}_{xy})_{\bullet l}\|_\infty \geq \lambda_l/4\}$ and $\mathcal{E}_2 = \{\|(\mathbf{S}_{xy})_{\bullet l} - \mathbf{S}_{xx}\boldsymbol{\beta}_l^*\|_\infty \geq \lambda_l/4\}$, and the "good" event $\mathcal{G}(\lambda_l) = \{\lambda_l/2 \geq \|(\widehat{\mathbf{S}}_{xy})_{\bullet l} - \mathbf{S}_{xx}\boldsymbol{\beta}_l^*\|_\infty\}$. Then by Boole's inequality

$$
\begin{aligned}
\Pr(\mathcal{G}(\lambda_l)) &= \Pr(\|(\widehat{\mathbf{S}}_{xy})_{\bullet l} - \mathbf{S}_{xx}\boldsymbol{\beta}_l^*\|_\infty < \lambda_l/2) \\
&\geq \Pr(\mathcal{E}_1^c \cap \mathcal{E}_2^c) = \Pr\left[(\mathcal{E}_1 \cup \mathcal{E}_2)^c\right] = 1 - \Pr\left[(\mathcal{E}_1 \cup \mathcal{E}_2)\right] \\
&\geq 1 - \Pr(\mathcal{E}_1) - \Pr(\mathcal{E}_2) \\
&\geq 1 - pC\exp\left(-\frac{cn\mu_{\min}^2\lambda_l^2}{\sigma_W^2 X_{\max}^2 \max(s_{\max}^2 X_{\max}^2 B_{\max}^2, \sigma_\varepsilon^2)}\right) - pC\exp\left[-\frac{cn\lambda_l^2}{\sigma_\varepsilon^2 X_{\max}^2}\right] \\
&\geq 1 - 2pC\max\left\{\exp\left(-\frac{cn\lambda_l^2}{\sigma_W^2 X_{\max}^2 \max(s_{\max}^2 X_{\max}^2 B_{\max}^2, \sigma_\varepsilon^2)/\mu_{\min}^2}\right), \exp\left[-\frac{cn\lambda_l^2}{\sigma_\varepsilon^2 X_{\max}^2}\right]\right\} \\
&= 1 - 2pC\exp\left(-\frac{cn\lambda_l^2}{\max\left[\sigma_W^2 X_{\max}^2 \max(s_{\max}^2 X_{\max}^2 B_{\max}^2, \sigma_\varepsilon^2)/\mu_{\min}^2, \sigma_\varepsilon^2 X_{\max}^2\right]}\right) \\
&= 1 - pC\exp\left(-\frac{cn\lambda_l^2}{X_{\max}^2 \max\left[\sigma_W^2 s_{\max}^2 X_{\max}^2 B_{\max}^2/\mu_{\min}^2, \sigma_W^2\sigma_\varepsilon^2/\mu_{\min}^2, \sigma_\varepsilon^2\right]}\right) \qquad (4.17)
\end{aligned}
$$

Returning to (4.16), when the "good" event $\mathcal{G}(\lambda_l)$ holds, we have

$$
\begin{aligned}
\frac{1}{2}[\boldsymbol{\delta}^\top \mathbf{S}_{xx}\boldsymbol{\delta}] &\leq \|\boldsymbol{\delta}\|_1 \|(\widehat{\mathbf{S}}_{xy})_{\bullet l} - \mathbf{S}_{xx}\boldsymbol{\beta}_l^*\|_\infty + \lambda_l(\|\boldsymbol{\beta}_l^*\|_1 - \|\widehat{\boldsymbol{\beta}}_l\|_1) \\
&\leq \frac{\lambda_l}{2}\|\boldsymbol{\delta}\|_1 + \lambda_l(\|\boldsymbol{\beta}_l^*\|_1 - \|\boldsymbol{\beta}_l^* + \boldsymbol{\delta}\|_1). \qquad (4.18)
\end{aligned}
$$

Now since $(\boldsymbol{\beta}_l^*)_{S_l^c} = \mathbf{0}$, we have $\|\boldsymbol{\beta}_l^*\|_1 = \|(\boldsymbol{\beta}_l^*)_{S_l}\|_1$, and using the reverse triangle inequality

$$
\|\boldsymbol{\beta}_l^* + \boldsymbol{\delta}\|_1 = \|(\boldsymbol{\beta}_l^*)_{S_l} + \boldsymbol{\delta}_{S_l}\|_1 + \|\boldsymbol{\delta}_{S_l^c}\|_1 \geq \|(\boldsymbol{\beta}_l^*)_{S_l}\|_1 - \|\boldsymbol{\delta}_{S_l}\|_1 + \|\boldsymbol{\delta}_{S_l^c}\|_1.
$$

Substituting these relations into inequality (4.18) yields

$$
\begin{aligned}
\frac{1}{2}[\boldsymbol{\delta}^\top \mathbf{S}_{xx}\boldsymbol{\delta}] &\leq \frac{\lambda_l}{2}\|\boldsymbol{\delta}\|_1 + \lambda_l(\|\boldsymbol{\beta}_l^*\|_1 - \|\boldsymbol{\beta}_l^* + \boldsymbol{\delta}\|_1). \\
&\leq \frac{\lambda_l}{2}\|\boldsymbol{\delta}\|_1 + \lambda_l(\|\boldsymbol{\beta}_l^*\|_1 - \|(\boldsymbol{\beta}_l^*)_{S_l}\|_1 + \|\boldsymbol{\delta}_{S_l}\|_1 - \|\boldsymbol{\delta}_{S_l^c}\|_1) \\
&= \frac{\lambda_l}{2}(\|\boldsymbol{\delta}_{S_l}\|_1 + \|\boldsymbol{\delta}_{S_l^c}\|_1) + \lambda_l(\|\boldsymbol{\delta}_{S_l}\|_1 - \|\boldsymbol{\delta}_{S_l^c}\|_1) \\
&\overset{\text{(i)}}{=} \frac{3\lambda_l}{2}\|\boldsymbol{\delta}_{S_l}\|_1 - \frac{\lambda_l}{2}\|\boldsymbol{\delta}_{S_l^c}\|_1 \qquad (4.19)
\end{aligned}
$$

$$\leq \frac{3\lambda_l}{2}\|\boldsymbol{\delta}_{S_l}\|_1 \leq \frac{3}{2}\sqrt{s_l}\lambda_l\|\boldsymbol{\delta}_{S_l}\|_2$$

This allows us to apply the restricted eigenvalue condititon to $\boldsymbol{\delta}$, which ensures that $\kappa_l\|\boldsymbol{\delta}\|_2^2 \leq \boldsymbol{\delta}^\top \mathbf{S}_{xx}\boldsymbol{\delta}$. Combining this lower bound with our earlier inequality yields

$$\frac{\kappa_l}{2}\|\boldsymbol{\delta}\|_2^2 \leq \frac{1}{2}[\boldsymbol{\delta}^\top \mathbf{S}_{xx}\boldsymbol{\delta}] \leq \frac{3}{2}\sqrt{s_l}\lambda_l\|\boldsymbol{\delta}_{S_l}\|_2 \leq \frac{3}{2}\sqrt{s_l}\lambda_l\|\boldsymbol{\delta}\|_2$$

and rearranging yields the bound

$$\|\widehat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l^*\|_2 \leq 3\sqrt{s_l}\lambda_l/\kappa_l.$$

Returning to equality (i) in (4.19), we can now show that the condition for $\mathbb{C}(S_l)$ holds

$$0 \leq \frac{1}{2}[\boldsymbol{\delta}^\top \mathbf{S}_{xx}\boldsymbol{\delta}] \leq \frac{3\lambda_l}{2}\|\boldsymbol{\delta}_{S_l}\|_1 - \frac{\lambda_l}{2}\|\boldsymbol{\delta}_{S_l^c}\|_1.$$

Hence $\|\boldsymbol{\delta}_{S_l^c}\|_1 \leq 3\|\boldsymbol{\delta}_{S_l}\|_1$. Now, we can also derive the ℓ_1-norm bound for the estimation error.

$$\begin{aligned}
\|\widehat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l^*\|_1 &= \|\boldsymbol{\delta}\|_1 \\
&= \|\boldsymbol{\delta}_{S_l}\|_1 + \|\boldsymbol{\delta}_{S_l^c}\|_1 \\
&\leq \|\boldsymbol{\delta}_{S_l}\|_1 + 3\|\boldsymbol{\delta}_{S_l}\|_1 \qquad \forall \boldsymbol{\delta} \in \mathbb{C}(S_l) \\
&\leq 4\|\boldsymbol{\delta}_{S_l}\|_1 \\
&\leq 4\sqrt{s_l}\|\boldsymbol{\delta}_{S_l}\|_2 \\
&\leq 12 s_l\lambda_l/\kappa_l.
\end{aligned}$$

When we set

$$\lambda_l/2 \geq \lambda_0/2 \coloneqq X_{\max}\max\left[\sigma_W s_{\max} X_{\max} B_{\max}/\mu_{\min}, \sigma_W\sigma_\varepsilon/\mu_{\min}, \sigma_\varepsilon\right]\sqrt{\frac{\log p}{n}} \tag{4.20}$$

in inequality (4.17), then

$$
\begin{aligned}
\Pr(\mathcal{G}_l) &= \Pr(\|(\widehat{\mathbf{S}}_{xy})_{\bullet l} - \mathbf{S}_{xx}\boldsymbol{\beta}_l^*\|_\infty \leq \lambda_l/2) \\
&\geq \Pr(\|(\widehat{\mathbf{S}}_{xy})_{\bullet l} - \mathbf{S}_{xx}\boldsymbol{\beta}_l^*\|_\infty \leq \lambda_0/2) \\
&\geq 1 - pC\exp\left(-\frac{cn(X_{\max}\max\left[\sigma_W s_{\max}X_{\max}B_{\max}/\mu_{\min}, \sigma_W\sigma_\varepsilon/\mu_{\min}, \sigma_\varepsilon\right])^2(\log p/n)}{X_{\max}^2\max\left[\sigma_W^2 s_{\max}^2 B_{\max}^2/\mu_{\min}^2, \sigma_W^2\sigma_\varepsilon^2/\mu_{\min}^2, \sigma_\varepsilon^2\right]}\right) \\
&= 1 - pC\exp\left(-c\log p\right) \\
&= 1 - C\exp\left(\log p - c\log p\right) \\
&= 1 - C\exp((1-c)\log p) \\
&= 1 - C\exp(-c\log p) \qquad (4.21)
\end{aligned}
$$

With slight abuse of the notation, we define a new constant $-c$ that is equal to $1-c$. Notice that the lower bound λ_0 of λ_l in (4.20) must satisfy the condition assumed in (4.41)

$$
\lambda_0/2 \coloneqq X_{\max}\max\left[\sigma_W s_{\max}X_{\max}B_{\max}/\mu_{\min}, \sigma_W\sigma_\varepsilon/\mu_{\min}, \sigma_\varepsilon\right]\sqrt{\frac{\log p}{n}} \leq t_0^{(1)} \coloneqq X_{\max}\sigma_W\sigma_\varepsilon/\mu_{\min}
$$

which implies that sample size n must be sufficiently large such that

$$
\begin{aligned}
\sqrt{\frac{\log p}{n}} &\leq \frac{X_{\max}\sigma_W\sigma_\varepsilon/\mu_{\min}}{X_{\max}\max\left[\sigma_W s_{\max}X_{\max}B_{\max}/\mu_{\min}, \sigma_W\sigma_\varepsilon/\mu_{\min}, \sigma_\varepsilon\right]} \\
&= \frac{1}{\max(s_{\max}B_{\max}X_{\max}/\sigma_\varepsilon, 1, \mu_{\min}/\sigma_W)} \\
&= \min(\sigma_\varepsilon/(s_{\max}B_{\max}X_{\max}), \sigma_W/\mu_{\min}, 1)
\end{aligned}
$$

Therefore, Lemma (13) can be applied as in (4.17) of this proof. □

4.3.2 Recovery Rate for the Estimator $\widehat{\boldsymbol{\Theta}}_{\varepsilon\varepsilon}$

We can express the deviation between the projected estimate $\widetilde{\mathbf{S}}_{\varepsilon\varepsilon}$ and the truth $\boldsymbol{\Sigma}_{\varepsilon\varepsilon}^*$ in terms of the deviation between the surrogate estimate and the truth using the following

inequality. Note that by the definition of $\widetilde{\mathbf{S}}_{\varepsilon\varepsilon}$, $|||\widetilde{\mathbf{S}}_{\varepsilon\varepsilon} - \widehat{\mathbf{S}}_{\varepsilon\varepsilon}|||_{\max} \leq |||\widehat{\mathbf{S}}_{\varepsilon\varepsilon} - \mathbf{\Sigma}^*_{\varepsilon\varepsilon}|||_{\max}$, since $\mathbf{\Sigma}^*_{\varepsilon\varepsilon}$ is positive semi-definite as well. Combining this with the triangular inequality, we have

$$|||\widetilde{\mathbf{S}}_{\varepsilon\varepsilon} - \mathbf{\Sigma}^*_{\varepsilon\varepsilon}|||_{\max} \leq |||\widetilde{\mathbf{S}}_{\varepsilon\varepsilon} - \widehat{\mathbf{S}}_{\varepsilon\varepsilon}|||_{\max} + |||\widehat{\mathbf{S}}_{\varepsilon\varepsilon} - \mathbf{\Sigma}^*_{\varepsilon\varepsilon}|||_{\max} \leq 2|||\widehat{\mathbf{S}}_{\varepsilon\varepsilon} - \mathbf{\Sigma}^*_{\varepsilon\varepsilon}|||_{\max}. \tag{4.22}$$

Following Ravikumar et al. (2011), we define the maximum degree or row cardinality of $\mathbf{\Theta}^*_{\varepsilon\varepsilon}$ as

$$d_p = \max_{l \in [q]} \operatorname{card}\left[\{l' \in [q] \backslash \{l\} : (\mathbf{\Theta}_{\varepsilon\varepsilon})_{ll'} \neq 0\}\right],$$

and let $\kappa_{\mathbf{\Sigma}^*_{\varepsilon\varepsilon}} = |||\mathbf{\Sigma}^*_{\varepsilon\varepsilon}|||_\infty$. We further let $S = \{(l, l') \in [q] \times [q] : (\mathbf{\Theta}_{\varepsilon\varepsilon})_{ll'} \neq 0\}$, $S^c = [q] \times [q] \backslash S$, and $\mathbf{\Gamma} = \mathbf{\Sigma}^*_{\varepsilon\varepsilon} \otimes \mathbf{\Sigma}^*_{\varepsilon\varepsilon} \in \mathbb{R}^{q^2} \times \mathbb{R}^{q^2}$. For any two subsets T and T' of $[q^2]$, let $(\mathbf{\Gamma})_{TT'}$ denote the $\operatorname{card}(T) \times \operatorname{card}(T')$ matrix with rows and columns of $\mathbf{\Gamma}$ indexed by T and T', respectively. Then we set $\kappa_{\mathbf{\Gamma}} = |||(\mathbf{\Gamma})^{-1}_{SS}|||_\infty$. Finally, set

$$\begin{aligned}\Delta = {} & \sigma_W^2 \max\left\{X_{\max}^2 B_{\max}^2 s_{\max}^2/m_{\min}, X_{\max} B_{\max} s_{\max} \sigma_\varepsilon^2/m_{\min}, \sigma_\varepsilon^2/m_{\min}\right\} \sqrt{\frac{\log(q^2)}{n}} \\ & + X_{\max}^2 \left(\max_{l \in [q]} 12 s_l \lambda_l/\kappa_l\right)^2 + X_{\max}^2 s_{\max} B_{\max} \max_{l \in [q]} 12 s_l \lambda_l/\kappa_l + \sigma_\varepsilon X_{\max} s_{\max} B_{\max} \sqrt{\frac{\log(q^2)}{n}} \\ & + \sigma_\varepsilon^2 \sqrt{\frac{\log(q^2)}{n}}.\end{aligned}$$

The details of deriving Δ is shown in Lemma 16. To derive the recovery rate for the estimator $\widehat{\mathbf{\Theta}}_{\varepsilon\varepsilon}$, let us introduce the irrepresentability condition introduced in Assumption 1 in Ravikumar et al. (2011) for graphical Lasso without any corrupted data.

We would require the following assumption and the Lemmas to prove Proposition 2.

Assumption 2 (Irrepresentability condition)**.** *There exists $\alpha \in (0, 1]$ such that*

$$\max_{e \in S^c} |||(\mathbf{\Gamma})_{\{e\}S} (\mathbf{\Gamma})^{-1}_{SS}|||_1 \leq 1 - \alpha.$$

Lemma 15. *Assume that each row of random error matrix $\boldsymbol{\varepsilon} \in \mathbb{R}^{n \times q}$ and each row of multiplicative error matrix $\mathbf{W} \in \mathbb{R}^{n \times q}$ follow the sub-Gaussian distribution with parameters σ_ε^2 and σ_W^2*

respectively, then the elementwise max norm of the deviation between $\widehat{\mathbf{S}}_{yy}$ *and* $\mathbf{S}_{yy}$ *satisfies the probability bound for any* $t \leq t_0^{(2)}$ *with*

$$t_0^{(2)} \coloneqq \min(X_{\max}^2 B_{\max}^2 s_{\max}^2 \sigma_W^2 / m_{\min}, X_{\max} B_{\max} s_{\max} \sigma_W^2 \sigma_\varepsilon^2 / m_{\min} n^{1/3}, \sigma_W^2 \sigma_\varepsilon^2 / m_{\min} n^{1/3})$$

$$\Pr(\|\widehat{\mathbf{S}}_{yy} - \mathbf{S}_{yy}\|_{\max} \geq t) \leq 4q^2 C \left\{ \exp\left(- \frac{cnt^2 m_{\min}^2}{\sigma_W^4 \max\{X_{\max}^4 B_{\max}^4 s_{\max}^4, X_{\max}^2 B_{\max}^2 s_{\max}^2 \sigma_\varepsilon^4, \sigma_\varepsilon^4\}} \right) \right\}.$$

where $m_{\min} = \min_{j,k} \mathbb{E}(W_j W_k^\top) > 0$, $X_{\max} = \max_{i,k} |X_{ik}| < \infty$, $B_{\max} = \max_{k,j} |\beta_{kj}^*|$, $s_{\max} = \max_j s_j$, *for* $j \in \{1, \ldots, q\}$.

Lemma 16. *Assume that each row of the error matrix* $\boldsymbol{\varepsilon} \in \mathbb{R}^{n \times q}$ *and each row of multiplicative error matrix* $\mathbf{W} \in \mathbb{R}^{n \times q}$ *are two sub-Gaussian random vectors where each elements of the vectors follow the sub-Gaussian distribution with parameters* σ_ε^2 *and* σ_W^2 *respectively. Under this assumption, the elementwise max norm of the deviation between* $\widehat{\mathbf{S}}_{\varepsilon\varepsilon}$ *and* $\boldsymbol{\Sigma}_{\varepsilon\varepsilon}^*$ *satisfies the following condition*

$$\Pr(\|\widehat{\mathbf{S}}_{\varepsilon\varepsilon} - \boldsymbol{\Sigma}_{\varepsilon\varepsilon}^*\|_{\max} \leq \Delta) \geq 1 - C\exp(-c\log q^2) - C\exp(-c\log(pq)),$$

where Δ *is defined as follows*

$$\begin{aligned} \Delta = {} & \sigma_W^2 \max\left\{ X_{\max}^2 B_{\max}^2 s_{\max}^2 / m_{\min}, X_{\max} B_{\max} s_{\max} \sigma_\varepsilon^2 / m_{\min}, \sigma_\varepsilon^2 / m_{\min} \right\} \sqrt{\frac{\log(q^2)}{n}} \\ & + X_{\max}^2 \left(\max_{l \in [q]} 12 s_l \lambda_l / \kappa_l \right)^2 + X_{\max}^2 s_{\max} B_{\max} \max_{l \in [q]} 12 s_l \lambda_l / \kappa_l + \sigma_\varepsilon X_{\max} s_{\max} B_{\max} \sqrt{\frac{\log(q^2)}{n}} \\ & + \sigma_\varepsilon^2 \sqrt{\frac{\log(q^2)}{n}}. \end{aligned} \tag{4.23}$$

Proposition 2. *Suppose that, for all* $l \in [q]$, $\kappa_l > 0$, λ_l *in* (4.10) *satisfies* (4.15) *and* n *is sufficiently large to ensure that Proposition* (1) *applies. Further, assume that Assumption* (2) *is satisfied and that* n *is sufficiently large to ensure that*

$$6 \left(1 + \frac{8}{\alpha} \right)^2 \max\left(\kappa_{\boldsymbol{\Sigma}_{\varepsilon\varepsilon}^*}, \kappa_{\boldsymbol{\Gamma}}, \kappa_{\boldsymbol{\Sigma}_{\varepsilon\varepsilon}^*}^3, \kappa_{\boldsymbol{\Gamma}}^2 \right) d_p \times \Delta \leq 1. \tag{4.24}$$

Finally, suppose that the tuning parameter $\lambda_{\mathbf{\Theta}}$ in (4.11) *satisfies*

$$\lambda_{\mathbf{\Theta}} = \frac{8\Delta}{\alpha}. \tag{4.25}$$

Then with probability at least

$$1 - C\exp(-c\log q^2) - C\exp(-c\log(pq)) \tag{4.26}$$

the estimator $\widehat{\mathbf{\Theta}}_{\varepsilon\varepsilon}$ satisfies

$$|||\widehat{\mathbf{\Theta}}_{\varepsilon\varepsilon} - \mathbf{\Theta}^*_{\varepsilon\varepsilon}|||_{\max} \leq \left\{2\kappa_{\mathbf{\Gamma}}(1 + \frac{8}{\alpha})\right\}\Delta \coloneqq \Delta_\infty(\mathbf{\Theta}^*_{\varepsilon\varepsilon}) \tag{4.27}$$

and

$$|||\widehat{\mathbf{\Theta}}_{\varepsilon\varepsilon} - \mathbf{\Theta}^*_{\varepsilon\varepsilon}|||_2 \leq d_p\Delta_\infty(\mathbf{\Theta}^*_{\varepsilon\varepsilon}) \coloneqq \Delta_1(\mathbf{\Theta}^*_{\varepsilon\varepsilon}). \tag{4.28}$$

Proof. Since $\widehat{\mathbf{S}}_{\varepsilon\varepsilon}$ is not necessarily positive semi-definite, we can produce an update $\widetilde{\mathbf{S}}_{\varepsilon\varepsilon}$ as described in Datta and Zou (2017) by projecting it onto the nearest semi-definite cone and substituting $\widehat{\mathbf{S}}_{\varepsilon\varepsilon}$ by $\widetilde{\mathbf{S}}_{\varepsilon\varepsilon}$. Then, $\widehat{\mathbf{S}}_{\varepsilon\varepsilon}$ would satisfy an inequality analogous to (4.22). Specifically,

$$|||\widetilde{\mathbf{S}}_{\varepsilon\varepsilon} - \mathbf{\Sigma}^*_{\varepsilon\varepsilon}|||_{\max} \leq 2|||\widehat{\mathbf{S}}_{\varepsilon\varepsilon} - \mathbf{\Sigma}^*_{\varepsilon\varepsilon}|||_{\max} \leq 2\Delta.$$

Applying Lemma 16, we obtain $|||\widetilde{\mathbf{S}}_{\varepsilon\varepsilon} - \mathbf{\Sigma}^*_{\varepsilon\varepsilon}|||_{\max} \leq 2\Delta$ occurs with probability $1 - C\exp(-c\log q^2) - C\exp(-c\log(pq))$. The conclusion of this Proposition follows from a slight variation of Theorem 1 from Ravikumar et al. (2011) where the observed sample was uncontaminated and one could calculate the sample covariance matrix $\widehat{\mathbf{\Sigma}}_{\varepsilon\varepsilon}$ with $|||\widehat{\mathbf{\Sigma}}_{\varepsilon\varepsilon} - \mathbf{\Sigma}^*_{\varepsilon\varepsilon}|||_{\max} \leq \bar{\delta}_f(n, p^\tau)$. In our case, $\widehat{\mathbf{\Sigma}}_{\varepsilon\varepsilon}$ is replaced by $\widetilde{\mathbf{S}}_{\varepsilon\varepsilon}$ and $\bar{\delta}_f(n, p^\tau) = 2\Delta$. Suppose $\widetilde{\mathbf{S}}_{\varepsilon\varepsilon}$ satisfies the error bound $|||\widetilde{\mathbf{S}}_{\varepsilon\varepsilon} - \mathbf{\Sigma}^*_{\varepsilon\varepsilon}|||_{\max} \leq 2\Delta$ on the intersection of events, that is, $\mathcal{B}_1 \cap \mathcal{B}_2 \cap \mathcal{B}_3 \cap \mathcal{B}_4 \cap \mathcal{B}_5$ defined in Lemma 16 and if the tuning parameter $\lambda_{\mathbf{\Theta}}$ satisfies (4.25),

then by Theorem 1 from Ravikumar et al. (2011), we have

$$|||\widehat{\boldsymbol{\Theta}}_{\varepsilon\varepsilon} - \boldsymbol{\Theta}^*_{\varepsilon\varepsilon}|||_{\max} \leq \left\{2\kappa_{\boldsymbol{\Gamma}}(1 + \frac{8}{\alpha})\right\} \Delta := \Delta_\infty(\boldsymbol{\Theta}^*_{\varepsilon\varepsilon}).$$

We can also show that (4.28) holds as follows

$$|||\widehat{\boldsymbol{\Theta}}_{\varepsilon\varepsilon} - \boldsymbol{\Theta}^*_{\varepsilon\varepsilon}|||_2 \leq \sqrt{d_p d_p}|||\widehat{\boldsymbol{\Theta}}_{\varepsilon\varepsilon} - \boldsymbol{\Theta}^*_{\varepsilon\varepsilon}|||_{\max} = d_p|||\widehat{\boldsymbol{\Theta}}_{\varepsilon\varepsilon} - \boldsymbol{\Theta}^*_{\varepsilon\varepsilon}|||_{\max} \leq d_p\Delta_\infty(\boldsymbol{\Theta}^*_{\varepsilon\varepsilon}) := \Delta_2(\boldsymbol{\Theta}^*_{\varepsilon\varepsilon})$$

and

$$|||\widehat{\boldsymbol{\Theta}}_{\varepsilon\varepsilon} - \boldsymbol{\Theta}^*_{\varepsilon\varepsilon}|||_1 \leq \sqrt{d_p}|||\widehat{\boldsymbol{\Theta}}_{\varepsilon\varepsilon} - \boldsymbol{\Theta}^*_{\varepsilon\varepsilon}|||_2 = \sqrt{d_p}\Delta_2(\boldsymbol{\Theta}^*_{\varepsilon\varepsilon}) := \Delta_1(\boldsymbol{\Theta}^*_{\varepsilon\varepsilon}).$$

□

4.3.3 Recovery Rate for $\widehat{\mathbf{B}}^{(2)}$

Let $\mathcal{L}(\cdot; \mathbf{S}_{xx}, \mathbf{S}_{xy}, \boldsymbol{\Theta}) : \mathbb{R}^{p\times q} \to \mathbb{R}$ be the loss functions to estimate $\mathbf{B}^*$ which depend on matrices $\mathbf{S}_{xx} \in \mathbb{R}^{p\times p}$, $\mathbf{S}_{xy} \in \mathbb{R}^{p\times q}$ and $\boldsymbol{\Theta} \in \mathbb{R}^{q\times q}$. Let $\mathbf{B} \in \mathbb{R}^{p\times q}$ be any arbitrary matrix and set

$$\begin{aligned}\mathcal{L}(\mathbf{B}; \mathbf{S}_{xx}, \mathbf{S}_{xy}, \boldsymbol{\Theta}) &= \langle \mathbf{B}^\top \mathbf{S}_{xx}\mathbf{B}/2 - \mathbf{S}_{xy}^\top \mathbf{B}, \boldsymbol{\Theta}\rangle \\ &= \operatorname{vec}(\mathbf{B})^\top(\boldsymbol{\Theta} \otimes \mathbf{S}_{xx}) \operatorname{vec}(\mathbf{B})/2 - \operatorname{tr}(\boldsymbol{\Theta}\mathbf{S}_{xy}^\top \mathbf{B}). \end{aligned} \tag{4.29}$$

The quantity $\mathbf{S}_{xx}$ in (4.29) will be replaced by the estimate of $\boldsymbol{\Sigma}^*_{xx}$ from the data since there is no missingness in $\mathbf{X}$. The quantities $\mathbf{S}_{xy}$ and $\boldsymbol{\Theta}$ will be replaced by the surrogate and the estimates of $\boldsymbol{\Sigma}^*_{xy}$ and $\boldsymbol{\Theta}^*_{\varepsilon\varepsilon}$, respectively. For brevity, we write, $\mathcal{L}(\cdot; \mathbf{S}_{xx}, \mathbf{S}_{xy}, \boldsymbol{\Theta})$ as $\mathcal{L}$ in the following derivations.

We consider the element-wise sparsity of $\mathbf{B}^*$. A matrix $\mathbf{B}^*$ is called element-wise sparse if its support set $S \subset \mathbb{R}^{p\times q}$ is such that $s = \operatorname{card}(S) \ll pq$. In order to obtain an element-wise sparse estimator of $\mathbf{B}^*$, it is natural to regularize the least squares program with the $|||\cdot|||_{1,1}$

penalty of $\mathbf{B}$,

$$\widehat{\mathbf{B}}^{(2)} = \underset{\mathbf{B}\in\mathbb{R}^{p\times q}}{\arg\min} \left\{ \mathcal{L}(\mathbf{B}; \mathbf{S}_{xx}, \widehat{\mathbf{S}}_{xy}, \widehat{\boldsymbol{\Theta}}_{\varepsilon\varepsilon}) + \lambda_{\mathbf{B}} |\!|\!| \mathbf{B} |\!|\!|_{1,1} \right\}, \tag{4.30}$$

where $\lambda_{\mathbf{B}} > 0$ is a tuning parameter.

Next, we establish the restricted eigenvalue (RE) condition for the loss function. Following Negahban et al. (2012), let us define, for all $\boldsymbol{\Delta} = \mathbf{B} - \mathbf{B}^* \in \mathbb{Q}$,

$$\begin{aligned}
\mathcal{EL}(\boldsymbol{\Delta}, \mathbf{B}^*) &= \mathcal{EL}(\boldsymbol{\Delta}, \mathbf{B}^*; \mathbf{S}_{xx}, \mathbf{S}_{xy}, \boldsymbol{\Theta}^*_{\varepsilon\varepsilon}) \\
&= \mathcal{L}(\mathbf{B}^* + \boldsymbol{\Delta}; \mathbf{S}_{xx}, \mathbf{S}_{xy}, \boldsymbol{\Theta}^*_{\varepsilon\varepsilon}) - \mathcal{L}(\mathbf{B}^*; \mathbf{S}_{xx}, \mathbf{S}_{xy}, \boldsymbol{\Theta}^*_{\varepsilon\varepsilon}) - \langle \nabla_{\mathbf{B}} \mathcal{L}(\mathbf{B}^*; \mathbf{S}_{xx}, \mathbf{S}_{xy}, \boldsymbol{\Theta}^*_{\varepsilon\varepsilon}), \boldsymbol{\Delta} \rangle \\
&= \langle \boldsymbol{\Delta}^\top \mathbf{S}_{xx} \boldsymbol{\Delta}, \boldsymbol{\Theta}^*_{\varepsilon\varepsilon} \rangle / 2 \\
&= \operatorname{vec}(\boldsymbol{\Delta})^\top (\boldsymbol{\Theta}^*_{\varepsilon\varepsilon} \otimes \mathbf{S}_{xx}) \operatorname{vec}(\boldsymbol{\Delta}) / 2,
\end{aligned}$$

where

$$\nabla_{\mathbf{B}} \mathcal{L}(\mathbf{B}; \mathbf{S}_{xx}, \mathbf{S}_{xy}, \boldsymbol{\Theta}^*_{\varepsilon\varepsilon}) = \mathbf{S}_{xx} \mathbf{B} \boldsymbol{\Theta}^*_{\varepsilon\varepsilon} - \mathbf{S}_{xy} \boldsymbol{\Theta}^*_{\varepsilon\varepsilon}$$

To derive the recovery rate for the estimator $\widehat{\mathbf{B}}^{(2)}$ we need the following assumption on the RE condition for the loss function.

Assumption 3. *The loss function $\mathcal{L}(\cdot; \mathbf{S}_{xx}, \mathbf{S}_{xy}, \boldsymbol{\Theta}^*_{\varepsilon\varepsilon})$ satisfies the RE condition*

$$\mathcal{EL}(\boldsymbol{\Delta}, \mathbf{B}^*) \geq \kappa |\!|\!| \boldsymbol{\Delta} |\!|\!|_2^2 \qquad \forall \boldsymbol{\Delta} \in \mathbb{C}(S). \tag{4.31}$$

with constant $\kappa > 0$ over the cone set $\mathbb{C}(S) = \{\mathbf{M} \in \mathbb{R}^{p\times q} : |\!|\!| (\mathbf{M})_{S^c} |\!|\!|_{1,1} \leq 3 |\!|\!| (\mathbf{M})_S |\!|\!|_{1,1}\}$.

Theorem 3. *Suppose that Assumption 3 and the assumptions of Proposition 2 hold. Further suppose that for $s = \operatorname{card}(S)$, where $S \subset \mathbb{R}^{p\times q}$ is the support set*

a) Assume $\kappa' \geq \kappa > 0$, where κ' is defined as

$$\kappa' = \kappa - |\!|\!| \mathbf{S}_{xx} |\!|\!|_2 \Delta_1(\boldsymbol{\Theta}^*_{\varepsilon\varepsilon}) \tag{4.32}$$

b) The tuning parameter $\lambda_{\mathbf{B}}$ in (4.30) satisfies

$$\lambda_{\mathbf{B}}/2 \geq (\lambda_0/2)(|||\widehat{\mathbf{\Theta}}_{\varepsilon\varepsilon} - \mathbf{\Theta}^*_{\varepsilon\varepsilon}|||_1 + |||\mathbf{\Theta}^*_{\varepsilon\varepsilon}|||_1), \tag{4.33}$$

then with probability $1 - qC\exp(-c\log p)$, the estimator $\widehat{\mathbf{B}}^{(2)}$ satisfies

$$|||\widehat{\mathbf{B}}^{(2)} - \mathbf{B}^*|||_F \leq 3\sqrt{s}\lambda_{\mathbf{B}}/\kappa', \qquad |||\widehat{\mathbf{B}}^{(2)} - \mathbf{B}^*|||_{1,1} \leq 12s\lambda_{\mathbf{B}}/\kappa'. \tag{4.34}$$

Proof. The proof relies partly on Proposition 3 stated below. The proposition verifies that the empirical loss function at this stage satisfies the RE condition..

Proposition 3. *Suppose that Assumption 3 and the assumptions of Proposition 2 hold. Then the empirical loss $\mathcal{L}(\cdot; \mathbf{S}_{xx}, \widehat{\mathbf{S}}_{xy}, \widehat{\mathbf{\Theta}}_{\varepsilon\varepsilon})$ satisfies RE condition with curvature κ' introduced in* (4.32) *and tolerance function equal to zero over the cone set $\mathbb{C}(S)$.*

Proof. We fix arbitrary $\mathbf{\Delta} \in \mathbb{C}(S)$. we have

$$\begin{aligned}\mathcal{EL}(\mathbf{\Delta}, \mathbf{B}^*; \mathbf{S}_{xx}, \widehat{\mathbf{S}}_{xy}, \widehat{\mathbf{\Theta}}_{\varepsilon\varepsilon}) = \operatorname{vec}(\mathbf{\Delta})^\top \mathbf{\Theta}^*_{\varepsilon\varepsilon} \otimes \mathbf{S}_{xx} \operatorname{vec}(\mathbf{\Delta})/2 \\ + \langle \mathbf{\Delta}^\top \mathbf{S}_{xx} \mathbf{\Delta}, \widehat{\mathbf{\Theta}}_{\varepsilon\varepsilon} - \mathbf{\Theta}^*_{\varepsilon\varepsilon} \rangle/2\end{aligned} \tag{4.35}$$

For the first term, by Assumption 3,

$$\operatorname{vec}(\mathbf{\Delta})^\top \mathbf{\Theta}^*_{\varepsilon\varepsilon} \otimes \mathbf{S}_{xx} \operatorname{vec}(\mathbf{\Delta})/2 \geq \kappa |||\mathbf{\Delta}|||_F^2$$

For the second term, we have

$$\begin{aligned}\left|\langle \mathbf{\Delta}^\top \mathbf{S}_{xx} \mathbf{\Delta}, \widehat{\mathbf{\Theta}}_{\varepsilon\varepsilon} - \mathbf{\Theta}^*_{\varepsilon\varepsilon} \rangle\right| &= \left|\langle \mathbf{S}_{xx}\mathbf{\Delta}, \mathbf{\Delta}(\widehat{\mathbf{\Theta}}_{\varepsilon\varepsilon} - \mathbf{\Theta}^*_{\varepsilon\varepsilon}) \rangle\right| \\ &\overset{\text{(i)}}{\leq} |||\mathbf{S}_{xx}\mathbf{\Delta}|||_F \times |||\mathbf{\Delta}(\widehat{\mathbf{\Theta}}_{\varepsilon\varepsilon} - \mathbf{\Theta}^*_{\varepsilon\varepsilon})|||_F \\ &\overset{\text{(ii)}}{\leq} |||\mathbf{S}_{xx}|||_F |||\mathbf{\Delta}|||_F \times |||\mathbf{\Delta}|||_F |||\widehat{\mathbf{\Theta}}_{\varepsilon\varepsilon} - \mathbf{\Theta}^*_{\varepsilon\varepsilon}|||_F\end{aligned}$$

$$\overset{\text{(iii)}}{\leq} |||\mathbf{S}_{xx}|||_2 |||\widehat{\boldsymbol{\Theta}}_{\varepsilon\varepsilon} - \boldsymbol{\Theta}^*_{\varepsilon\varepsilon}|||_2 |||\boldsymbol{\Delta}|||_F^2, \tag{4.36}$$

where inequality (i) follows from Hölder's inequality, inequality (ii) follows from the submultiplicative property of the Frobenius norm and inequality (iii) follows from the fact $|||\mathbf{A}|||_F = |||\mathbf{A}|||_2$ since the trace of a matrix is equal to the sum of its eigenvalues.

Therefore, (4.35) can be bounded as

$$\mathcal{EL}(\boldsymbol{\Delta}, \mathbf{B}^*; \mathbf{S}_{xx}, \widehat{\mathbf{S}}_{xy}, \widehat{\boldsymbol{\Theta}}_{\varepsilon\varepsilon}) \geq \kappa |||\boldsymbol{\Delta}|||_F^2 - (|||\mathbf{S}_{xx}|||_2 |||\widehat{\boldsymbol{\Theta}}_{\varepsilon\varepsilon} - \boldsymbol{\Theta}^*_{\varepsilon\varepsilon}|||_2 |||\boldsymbol{\Delta}|||_F^2)/2 \tag{4.37}$$

For the regularized Lasso problem, since $\boldsymbol{\Delta} \in \mathbb{C}(S)$, and therefore, we can write

$$|||\boldsymbol{\Delta}|||_{1,1} = |||\boldsymbol{\Delta}_S|||_{1,1} + |||\boldsymbol{\Delta}_{S^c}|||_{1,1} \leq 4|||\boldsymbol{\Delta}_S|||_{1,1} \leq 4\sqrt{s}|||\boldsymbol{\Delta}|||_F. \tag{4.38}$$

Combining (4.37) and (4.38), we conclude that

$$\begin{aligned}\mathcal{EL}(\boldsymbol{\Delta}, \mathbf{B}^*; \mathbf{S}_{xx}, \widehat{\mathbf{S}}_{xy}, \widehat{\boldsymbol{\Theta}}_{\varepsilon\varepsilon}) &\geq \left\{\kappa - (|||\mathbf{S}_{xx}|||_2 |||\widehat{\boldsymbol{\Theta}}_{\varepsilon\varepsilon} - \boldsymbol{\Theta}^*_{\varepsilon\varepsilon}|||_2)/2\right\} |||\boldsymbol{\Delta}|||_F^2 \\ &\geq (\kappa - |||\mathbf{S}_{xx}|||_2 \Delta_1(\boldsymbol{\Theta}^*_{\varepsilon\varepsilon})/2) |||\boldsymbol{\Delta}|||_F^2 \end{aligned} \tag{4.39}$$

If the assumptions of Proposition 2 are satisfied, then (4.27) and (4.28) also hold. Then we can further bound the right-hand side of (4.39) from below. This concludes the proof of Proposition 3. □

Next, we want to apply Theorem 7.13 from Wainwright (2019). To do so, first, we check if the conditions mentioned in Proposition 1 holds. We set,

$$\mathcal{M} \coloneqq \left\{\boldsymbol{\Delta} \in \mathbb{R}^{p\times q} : (\boldsymbol{\Delta})_{kl} = 0 \qquad \forall (k,l) \in S^c\right\}.$$

Then $\mathbf{B}^* \in \mathcal{M}$, and the penalty $\mathcal{R}$ is decomposable with respect to $(\mathcal{M}, \mathcal{M}^\perp)$. Next, by Proposition 3 and the assumption on κ', over the cone set $\mathbb{C}(S)$, the loss $\mathcal{L}(\cdot; \mathbf{S}_{xx}, \widehat{\mathbf{S}}_{xy}, \widehat{\boldsymbol{\Theta}}_{\varepsilon\varepsilon})$

satisfies RE condition with tolerance function equal to zero and curvature $\kappa' = \kappa - |||\mathbf{S}_{xx}|||_2 \Delta_1(\mathbf{\Theta}^*_{\varepsilon\varepsilon}) > 0$. Finally, the dual norm of $\mathcal{R}$ is $\mathcal{R}^*(\cdot) = |||\cdot|||_{\max}$, and

$$
\begin{aligned}
\mathcal{R}^*\left\{\nabla_{\mathbf{B}}\mathcal{L}(\mathbf{B}^*; \mathbf{S}_{xx}, \widehat{\mathbf{S}}_{xy}, \widehat{\mathbf{\Theta}}_{\varepsilon\varepsilon})\right\} &= |||\mathbf{S}_{xx}\mathbf{B}^*\widehat{\mathbf{\Theta}}_{\varepsilon\varepsilon} - \widehat{\mathbf{S}}_{xy}\widehat{\mathbf{\Theta}}_{\varepsilon\varepsilon}|||_{\max} \\
&= |||(\widehat{\mathbf{S}}_{xy} - \mathbf{S}_{xx}\mathbf{B}^*)(\widehat{\mathbf{\Theta}}_{\varepsilon\varepsilon} - \mathbf{\Theta}^*_{\varepsilon\varepsilon}) + (\widehat{\mathbf{S}}_{xy} - \mathbf{S}_{xx}\mathbf{B}^*)\mathbf{\Theta}^*_{\varepsilon\varepsilon}|||_{\max} \\
&\leq |||(\widehat{\mathbf{S}}_{xy} - \mathbf{S}_{xx}\mathbf{B}^*)(\widehat{\mathbf{\Theta}}_{\varepsilon\varepsilon} - \mathbf{\Theta}^*_{\varepsilon\varepsilon})|||_{\max} + |||(\widehat{\mathbf{S}}_{xy} - \mathbf{S}_{xx}\mathbf{B}^*)\mathbf{\Theta}^*_{\varepsilon\varepsilon}|||_{\max} \\
&\overset{\text{(i)}}{\leq} |||\widehat{\mathbf{S}}_{xy} - \mathbf{S}_{xx}\mathbf{B}^*|||_{\max}|||\widehat{\mathbf{\Theta}}_{\varepsilon\varepsilon} - \mathbf{\Theta}^*_{\varepsilon\varepsilon}|||_{\infty} + |||\widehat{\mathbf{S}}_{xy} - \mathbf{S}_{xx}\mathbf{B}^*|||_{\max}|||\mathbf{\Theta}^*_{\varepsilon\varepsilon}|||_{\infty} \\
&= |||\widehat{\mathbf{S}}_{xy} - \mathbf{S}_{xx}\mathbf{B}^*|||_{\max}(|||\widehat{\mathbf{\Theta}}_{\varepsilon\varepsilon} - \mathbf{\Theta}^*_{\varepsilon\varepsilon}|||_{\infty} + |||\mathbf{\Theta}^*_{\varepsilon\varepsilon}|||_{\infty}) \\
&= |||\widehat{\mathbf{S}}_{xy} - \mathbf{S}_{xx}\mathbf{B}^*|||_{\max}(|||\widehat{\mathbf{\Theta}}_{\varepsilon\varepsilon} - \mathbf{\Theta}^*_{\varepsilon\varepsilon}|||_{1} + |||\mathbf{\Theta}^*_{\varepsilon\varepsilon}|||_{1}) \\
&\leq (\lambda_0/2)(|||\widehat{\mathbf{\Theta}}_{\varepsilon\varepsilon} - \mathbf{\Theta}^*_{\varepsilon\varepsilon}|||_{1} + |||\mathbf{\Theta}^*_{\varepsilon\varepsilon}|||_{1})
\end{aligned}
$$

Inequality (i) follows from the fact $|||\mathbf{AB}|||_{\max} \leq |||\mathbf{A}|||_{\infty}|||\mathbf{B}|||_{\max}$ and equality (ii) follows from the relationship $|||\mathbf{A}|||_1 = |||\mathbf{A}^\top|||_{\infty}$. Denote $C_\Theta = |||\widehat{\mathbf{\Theta}}_{\varepsilon\varepsilon} - \mathbf{\Theta}^*_{\varepsilon\varepsilon}|||_1 + |||\mathbf{\Theta}^*_{\varepsilon\varepsilon}|||_1$, and choose

$$
\lambda_0/2 \coloneqq X_{\max} \max\left[\sigma_W s_{\max} X_{\max} B_{\max}/\mu_{\min}, \sigma_W \sigma_\varepsilon/\mu_{\min}, \sigma_\varepsilon\right] \sqrt{\frac{\log p}{n}}
$$

then we have

$$
\begin{aligned}
&\Pr(\mathcal{R}^*\{\nabla_{\mathbf{B}}\mathcal{L}(\mathbf{B}^*; \mathbf{S}_{xx}, \widehat{\mathbf{S}}_{xy}, \widehat{\mathbf{\Theta}}_{\varepsilon\varepsilon})\} \leq (\lambda_0/2)C_\Theta) \\
&\geq \Pr(|||\widehat{\mathbf{S}}_{xy} - \mathbf{S}_{xx}\mathbf{B}^*|||_{\max} C_\Theta \leq (\lambda_0/2)C_\Theta) \\
&= \Pr(|||\widehat{\mathbf{S}}_{xy} - \mathbf{S}_{xx}\mathbf{B}^*|||_{\max} \leq \lambda_0/2) \\
&= \Pr(\max_l \|(\widehat{\mathbf{S}}_{xy})_{\bullet l} - \mathbf{S}_{xx}\boldsymbol{\beta}^*_l\|_\infty \leq \lambda_0/2) \\
&= \Pr(\cap_l\{\|(\widehat{\mathbf{S}}_{xy})_{\bullet l} - \mathbf{S}_{xx}\boldsymbol{\beta}^*_l\|_\infty \leq \lambda_0/2\}) \\
&= 1 - \Pr(\cup_l\{\|(\widehat{\mathbf{S}}_{xy})_{\bullet l} - \mathbf{S}_{xx}\boldsymbol{\beta}^*_l\|_\infty \geq \lambda_0/2\}) \\
&\overset{\text{(i)}}{\geq} 1 - \sum_{l=1}^{q} \Pr(\|(\widehat{\mathbf{S}}_{xy})_{\bullet l} - \mathbf{S}_{xx}\boldsymbol{\beta}^*_l\|_\infty \geq \lambda_0/2)
\end{aligned}
$$

$$= 1 - qC\exp(-c\log p)$$

Inequality (i) follows from the tail probability bound in (4.21). The conclusion from (4.34) follows from Theorem 7.13 in Wainwright (2019) with probability $1 - qC\exp(-c\log p)$. □

4.4 Simulation

In this study, we focus on estimating the precision matrix and the regression coefficients in three stages as explained in Section 4.2. We can summarize the algorithm in the following pseudo code.

Algorithm 1: Pseudocode for the estimation of $\mathbf{B}^*$ and $\mathbf{\Theta}^*$

Stage I:

Step 1: Keeping $\mathbf{\Theta}$ fixed, estimate $\widehat{\mathbf{B}}^{(1)}$ from (4.9) by performing a column by column estimation of $\mathbf{B}^*$.

Stage II:

Step 2: Calculate $\widetilde{\mathbf{S}}_{\varepsilon\varepsilon}$, a positive semi-definite estimate of $\widehat{\mathbf{S}}_{\varepsilon\varepsilon}$ using (4.12).

Step 3: Estimate $\widehat{\mathbf{\Theta}}_{\varepsilon\varepsilon}$ from (4.13) using plug-in estimates of $\widehat{\mathbf{B}}^{(1)}$ and $\widetilde{\mathbf{S}}_{\varepsilon\varepsilon}$ calculated in **Step 1** and **2**.

Stage III:

Step 4: Calculate the refined estimate, $\widehat{\mathbf{B}}^{(2)}$, the second stage estimation of $\mathbf{B}^*$ using (4.14) by plugging in $\widehat{\mathbf{\Theta}}_{\varepsilon\varepsilon}$ from **Step 3**.

In Stage I, we performed a projected gradient descent to get the initial estimate $\widehat{\mathbf{B}}^{(1)}$ of $\mathbf{B}^*$ for each column of $\mathbf{Y}$. Since $\mathbf{Y}$ is corrupted due to missingness in the data, we used the surrogate estimates defined in (4.5) to perform column by column Lasso regression using projected gradient descent method. In Stage II, we used the CoGlasso method introduced in Chapter 3 to estimate $\widehat{\mathbf{\Theta}}_{\varepsilon\varepsilon}$ after plugging in $\widehat{\mathbf{B}}^{(1)}$. In the last stage, we performed an

iterative soft-thresholding algorithm (ISTA) after plugging in $\widehat{\mathbf{\Theta}}_{\varepsilon\varepsilon}$ to estimate the updated version of $\widehat{\mathbf{B}}^{(2)}$.

We used a simple simulation setting to demonstrate our method. We choose three simulation settings with $n = 300$ and $q = 20$ with enlarging $p = (50, 100, 300)$. We generated the covariance structure of $\mathbf{X}$, $\mathbf{\Sigma}_{xx}$ as an AR(1) process with $\rho_X = 0.7$. We generated the true precision matrix $\mathbf{\Theta}^*_{\varepsilon\varepsilon}$ as a chain graph with partial correlation, $\rho_E = (-0.1, -0.5, -0.7, -0.9)$. To obtain an elementwise sparse model on $\mathbf{B}^*$, we first generated a $p \times q$ matrix $\tilde{\mathbf{B}}^*$ so that in each of its columns, 80% of its p elements (chosen at random within each column) are equal to zero; the remaining 20% of the entries of $\tilde{\mathbf{B}}^*$ were drawn from the uniform distribution on [-1, 1]. We generated the samples of $(\mathbf{X}, \mathbf{Y})$ from a zero mean multivariate Gaussian distribution with covariance matrices $\mathbf{\Sigma}_{xx}$ and $\mathbf{\Sigma}_{\varepsilon\varepsilon} = (\mathbf{\Theta}^*_{\varepsilon\varepsilon})^{-1}$, respectively. Then we introduced 10% missingness completely at random in each column of $\mathbf{Y}$.

The metrics that we chose to compare are false positive rate and false negative rate for the Stage II estimator of the precision matrix defined in Chapter 3 in Section 3.6.2. To measure the performance of the updated estimator for $\mathbf{B}^*$, we calculated the prediction error, PE, as

$$\mathrm{PE}(\widehat{\mathbf{B}}^{(2)}, \mathbf{B}^*) = \mathrm{tr}\{(\widehat{\mathbf{B}}^{(2)} - \mathbf{B}^*)^\top \mathbf{\Sigma}_{xx} (\widehat{\mathbf{B}}^{(2)} - \mathbf{B}^*)\}$$

and squared error, SE, as

$$\mathrm{SE}(\widehat{\mathbf{B}}^{(2)}, \mathbf{B}^*) = |||\widehat{\mathbf{B}}^{(2)} - \mathbf{B}^*|||_F^2.$$

We compared the performances of the estimators where the Stage II precision matrix was calculated using CoGlasso method and the non-convex ADMM algorithm. We are referring to the two methods as Convex and NC respectively in the following tables. Note that, Stage I and Stage II of the algorithm is the same and we only varied how we calculate the precision matrix in Stage II for this comparison. We tuned the regularization parameters in

each stage from a equally spaced range of λs in a logarithmic scale between [-2,2]. All the results are averaged across 100 replications.

Table 4.1: *Scenario S1:* $n = 300, p = 50, q = 20, P_{\text{miss}} = 0.1$

ρ_E	Method	FPR	FNR	PE	SE
-0.1	Convex	0.00	1.00	2.45	5.46
	NC	0.02	0.95	2.44	5.45
-0.5	Convex	0.08	0.06	2.83	6.16
	NC	0.23	0.02	2.81	6.11
-0.7	Convex	0.19	0.00	3.19	6.63
	NC	0.42	0.00	3.19	6.63
-0.9	Convex	0.30	0.01	4.34	8.49
	NC	0.64	0.00	4.44	8.69

Table 4.2: *Scenario S2:* $n = 300, p = 100, q = 20, P_{\text{miss}} = 0.1$

ρ_E	Method	FPR	FNR	PE	SE
-0.10	Convex	0.00	1.00	5.97	14.52
	NC	0.02	0.98	5.90	14.36
-0.50	Convex	0.01	0.69	6.90	16.63
	NC	0.11	0.27	6.78	16.27
-0.70	Convex	0.17	0.03	7.73	17.81
	NC	0.46	0.01	7.63	17.56
-0.90	Convex	0.91	0.01	10.12	21.75
	NC	0.91	0.01	10.12	21.75

Table 4.3: *Scenario S3:* $n = 300, p = 300, q = 20, P_{\text{miss}} = 0.1$

ρ_E	Method	FPR	FNR	PE	SE
-0.10	Convex	0.00	1.00	29.48	101.54
	NC	0.16	0.84	29.25	100.75
-0.50	Convex	0.00	1.00	32.17	109.62
	NC	0.34	0.59	31.80	107.88
-0.70	Convex	0.01	0.92	37.12	123.23
	NC	0.76	0.14	36.12	118.74
-0.90	Convex	0.38	0.07	46.44	135.84
	NC	0.99	0.00	45.68	134.19

Both the methods tend to perform comparably in all the scenarios. We can clearly see that the prediction error and squared error are increasing as the number of covariates increase.

4.5 Discussion and Conclusion

In this chapter, we have studied the theoretical properties of estimating the regression coefficients and the precision matrix in the presence of missing data from a multivariate regression setup. We proposed a three step estimation procedure to efficiently estimate the parameters of the model. We also performed some simulations to illustrate the method. Note that, we have not compared our methods with the state-of-the-art methods in all stages. The results shown only vary in the second stage for the precision matrix matrix estimation where our method was compared with the non-convex approach of estimating precision matrix proposed by Fan et al. (2019) using an ADMM algorithm. Both the methods tend to perform similarly in this case. We have only considered a chain graph structure of the precision matrix corresponding to the error. Therefore, there are further scope of testing out our model numerically for more complicated graph structures with varying sparsity.

4.6 Technical Details

Proof of Lemma 13

Proof. Recall that, for a multiplicative measurement error model, we assume the observed matrix is $\mathbf{Z} = \mathbf{Y} \odot \mathbf{W}$ where $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_n)^\top$ is a matrix of multiplicative error. Given $\mathbf{S}_{xy} = \frac{1}{n}\mathbf{X}^\top\mathbf{Y}$ as the matrix that represents the covariance between $\mathbf{X}$ and uncontaminated

$\mathbf{Y}$, we have

$$\begin{aligned}\widehat{\mathbf{S}}_{xy} - \mathbf{S}_{xy} &= \frac{1}{n}\mathbf{X}^\top \mathbf{Z} \oslash [\mathbb{E}[W], ..., \mathbb{E}[W]]^\top - \frac{1}{n}\mathbf{X}^\top \mathbf{Y} \\ &= \frac{1}{n}\mathbf{X}^\top (\mathbf{Y} \odot \mathbf{W}) \oslash [\mathbf{1}_q - \boldsymbol{\rho}, \ldots, \mathbf{1}_q - \boldsymbol{\rho}]^\top - \frac{1}{n}\mathbf{X}^\top \mathbf{Y}. \end{aligned} \tag{4.40}$$

Let Y_{ij} be the ith row and jth column of $\mathbf{Y}$ for $j = 1, \ldots, q$ and $k = 1, \ldots, p$, also W_{ij} and X_{ik} are defined similarly. Followed by (4.40), we have,

$$\begin{aligned}(\widehat{\mathbf{S}}_{xy})_{kj} - (\mathbf{S}_{xy})_{kj} &= \frac{1}{n(1-\rho_j)} \sum_{i=1}^{n} Y_{ij} W_{ij} X_{ik} - \frac{1}{n} \sum_{i=1}^{n} Y_{ij} X_{ik} \\ &= \frac{1}{n(1-\rho_j)} \sum_{i=1}^{n} Y_{ij} W_{ij} X_{ik} - \frac{1}{n(1-\rho_j)} \sum_{i=1}^{n} Y_{ij} X_{ik} \mathbb{E} W_{ij} \\ &= \frac{1}{n(1-\rho_j)} \sum_{i=1}^{n} Y_{ij} X_{ik} (W_{ij} - \mathbb{E} W_{ij}),\end{aligned}$$

with $\mathbb{E}W_{ij} = 1 - \rho_j$. Now we plug in the true model $Y_{ij} = \sum_{k' \in S_j} X_{ik'} \beta^*_{k'j} + \varepsilon_{ij}$ and get

$$\begin{aligned}&\left|(\widehat{\mathbf{S}}_{xy})_{kj} - (\mathbf{S}_{xy})_{kj}\right| \\ &= \left| \frac{1}{n(1-\rho_j)} \sum_{i=1}^{n} (\sum_{k'=1}^{s_j} X_{ik'} \beta^*_{k'j} + \varepsilon_{ij}) X_{ik} (W_{ij} - \mathbb{E} W_{ij}) \right| \\ &= \frac{1}{(1-\rho_j)} \left[\left| \frac{1}{n} \sum_{i=1}^{n} \sum_{k'=1}^{s_j} \beta^*_{k'j} X_{ik'} X_{ik} (W_{ij} - \mathbb{E} W_{ij}) \right| + \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{ij} X_{ik} (W_{ij} - \mathbb{E} W_{ij}) \right| \right] \\ &\leq \frac{1}{\mu_{\min}} \left[\left| \frac{1}{n} \sum_{i=1}^{n} \sum_{k'=1}^{s_j} \beta^*_{k'j} X_{ik'} X_{ik} (W_{ij} - \mathbb{E} W_{ij}) \right| + \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{ij} X_{ik} (W_{ij} - \mathbb{E} W_{ij}) \right| \right] \\ &\leq \frac{1}{\mu_{\min}} \left[s_{\max} B_{\max} X^2_{\max} \underbrace{\left| \frac{1}{n} \sum_{i=1}^{n} (W_{ij} - \mathbb{E} W_{ij}) \right|}_{T_1} + X_{\max} \underbrace{\left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{ij} (W_{ij} - \mathbb{E} W_{ij}) \right|}_{T_2} \right]\end{aligned}$$

where $X_{\max} = \max_{i,k} |X_{ik}| < \infty$, $B_{\max} = \max_{k,j} |\beta^*_{kj}|$, $\mu_{\min} = \min_j \mu_j = \min_j 1 - \rho_j$. We find bounds for T_1 and T_2 separately. For T_1, applying Theorem 2.6.3 general Hoeffding

inequality on page 27 of Vershynin (2018), we get

$$
\begin{aligned}
&\Pr(|T_1| \geq t) \\
&= \Pr\left[\left|\sum_{i=1}^{n} \frac{1}{n}(W_{ij} - \mathbb{E}W_{ij})\right| \geq t\right] \\
&\leq 2\exp\left\{-\frac{ct^2}{K_W^2 \sum_{i=1}^{n}(1/n)^2}\right\} \\
&\overset{\text{(i)}}{\leq} 2\exp\left\{-\frac{cm_1 t^2}{\sigma_W^2 \sum_{i=1}^{n}(1/n)^2}\right\} \\
&\overset{\text{(ii)}}{\leq} 2\exp\left\{-\frac{cnt^2}{\sigma_W^2}\right\}
\end{aligned}
$$

where $K_W = \max_{i,j} \|W_{ij} - \mathbb{E}W_{ij}\|_{\psi_2}$. Inequality (i) is due to the implication of Lemma 5.5 of Vershynin (2010) that for sub-Gaussian random variable $W_{ij} - \mathbb{E}W_{ij}$ there exist universal constants m_1 and M_1 such that $m_1\|W_{ij} - \mathbb{E}W_{ij}\|_{\psi_2}^2 \leq \sigma_W^2 \leq M_1\|W_{ij} - \mathbb{E}W_{ij}\|_{\psi_2}^2$ hold. It can be simplified to inequality (ii) since m_1 is a constant and consequently gets absorbed into the universal constant c.

Now we will find the bound for T_2. Follow by Lemma 2.7.7 of Vershynin (2018), we notice that $\varepsilon_{ij}(W_{ij} - \mathbb{E}W_{ij})$ as the product of two independent, centered sub-Gaussian random variables follows sub-exponential distribution since

$$\|\varepsilon_{ij}(W_{ij} - \mathbb{E}W_{ij})\|_{\psi_1} \leq \|\varepsilon_{ij}\|_{\psi_2}\|W_{ij} - \mathbb{E}W_{ij}\|_{\psi_2}.$$

Also we can see that $\varepsilon_{ij}(W_{ij} - \mathbb{E}W_{ij})$ is centered since

$$\mathbb{E}[\varepsilon_{ij}(W_{ij} - \mathbb{E}W_{ij})] = \mathbb{E}\varepsilon_{ij}\mathbb{E}(W_{ij} - \mathbb{E}W_{ij}) = 0$$

with Now apply Theorem 2.8.2 of Vershynin (2018), we get

$$\Pr(|T_2| \geq t)$$

$$
\begin{aligned}
&= \Pr\left[\left|\sum_{i=1}^{n}\frac{1}{n}\varepsilon_{ij}(W_{ij}-\mathbb{E}W_{ij})\right| > t\right] \\
&\leq 2\exp\left\{-c\min\left(\frac{t^2}{\max_i\|\varepsilon_{ij}(W_{ij}-\mathbb{E}W_{ij})\|_{\psi_1}^2\sum_{i=1}^{n}(1/n)^2}, \frac{t}{\max_i\|\varepsilon_{ij}(W_{ij}-\mathbb{E}W_{ij})\|_{\psi_1}\max_i(1/n)}\right)\right\} \\
&\overset{\text{(i)}}{\leq} 2\exp\left\{-c\min\left(\frac{nt^2}{\max_i\|\varepsilon_{ij}\|_{\psi_2}^2\|W_{ij}-\mathbb{E}W_{ij}\|_{\psi_2}^2}, \frac{nt}{\max_i\|\varepsilon_{ij}\|_{\psi_2}\|W_{ij}-\mathbb{E}W_{ij}\|_{\psi_2}}\right)\right\} \\
&\leq 2\exp\left\{-cn\min\left(\frac{t^2}{\sigma_\varepsilon^2\sigma_W^2}, \frac{t}{\sigma_\varepsilon\sigma_W}\right)\right\}
\end{aligned}
$$

where inequality (i) is again due to Lemma 5.5 of Vershynin (2010) that there exist universal constants m_1, M_1, m_2 and M_2 such that $m_1\|W_{ij}-\mathbb{E}W_{ij}\|_{\psi_2}^2 \leq \sigma_W^2 \leq M_1\|W_{ij}-\mathbb{E}W_{ij}\|_{\psi_2}^2$ and $m_2\|\varepsilon_{ij}\|_{\psi_2}^2 \leq \sigma_\varepsilon^2 \leq M_2\|\varepsilon_{ij}\|_{\psi_2}^2$ hold.

We can combine the pieces together as follows:

$$
\begin{aligned}
&\Pr\left[\left|(\widehat{\mathbf{S}}_{xy})_{kj}-(\mathbf{S}_{xy})_{kj}\right| \geq t\right] \\
&\leq \Pr\left[\frac{s_{\max}B_{\max}X_{\max}^2}{\mu_{\min}}|T_1| + \frac{X_{\max}}{\mu_{\min}}|T_2| \geq t\right] \\
&= \Pr\left[s_{\max}B_{\max}X_{\max}^2|T_1| + X_{\max}|T_2| \geq t\mu_{\min}\right] \\
&\overset{\text{(i)}}{\leq} \Pr\left[2\max\left\{s_{\max}B_{\max}X_{\max}^2|T_1|, X_{\max}|T_2|\right\} \geq t\mu_{\min}\right] \\
&= \Pr\left[\max\left\{s_{\max}B_{\max}X_{\max}^2|T_1|, X_{\max}|T_2|\right\} \geq t\mu_{\min}/2\right] \\
&= \Pr\left[\left(s_{\max}B_{\max}X_{\max}^2|T_1| \geq t\mu_{\min}/2\right) \cup \left(X_{\max}|T_2| \geq t\mu_{\min}/2\right)\right] \\
&= \Pr\left[\left(|T_1| \geq t\mu_{\min}/(2s_{\max}B_{\max}X_{\max}^2)\right) \cup \left(|T_2| \geq t\mu_{\min}/(2X_{\max})\right)\right] \\
&\leq \Pr\left[|T_1| \geq t\mu_{\min}/(2s_{\max}B_{\max}X_{\max}^2)\right] + \Pr\left[|T_2| \geq t\mu_{\min}/(2X_{\max})\right] \\
&= C\exp\left\{-cn\frac{\mu_{\min}^2t^2}{s_{\max}^2\sigma_W^2X_{\max}^4B_{\max}^2}\right\} \\
&\qquad + C\exp\left\{-cn\min\left(\frac{\mu_{\min}^2t^2}{\sigma_\varepsilon^2\sigma_W^2X_{\max}^2}, \frac{\mu_{\min}t}{\sigma_\varepsilon\sigma_WX_{\max}}\right)\right\} \\
&\overset{\text{(ii)}}{=} C\exp\left\{-cn\frac{\mu_{\min}^2t^2}{s_{\max}^2\sigma_W^2X_{\max}^4B_{\max}^2}\right\} + C\exp\left\{-cn\frac{\mu_{\min}^2t^2}{\sigma_\varepsilon^2\sigma_W^2X_{\max}^2}\right\} \\
&\leq 2C\max\left\{\exp\left(-cn\frac{\mu_{\min}^2t^2}{s_{\max}^2\sigma_W^2X_{\max}^4B_{\max}^2}\right), \exp\left(-cn\frac{\mu_{\min}^2t^2}{\sigma_\varepsilon^2\sigma_W^2X_{\max}^2}\right)\right\}
\end{aligned}
$$

$$\leq C \exp\left(-cn \frac{\mu_{\min}^2 t^2}{\sigma_W^2 X_{\max}^2 \max(s_{\max}^2 X_{\max}^2 B_{\max}^2, \sigma_\varepsilon^2)}\right)$$

Inequality (i) is due to the relationship $A + B + |A - B| = 2\max(A, B)$, which implies $A + B \leq 2\max(A, B)$. Equality (ii) follows by assuming that $t \leq t_0^{(1)}$ with

$$t_0^{(1)} = \sigma_\varepsilon \sigma_W X_{\max} / \mu_{\min}. \tag{4.41}$$

Applying the union bound to find an upper bound for the elementwise max norm, we get,

$$\begin{aligned}
\Pr[|||\widehat{\mathbf{S}}_{xy} - \mathbf{S}_{xy}|||_{\max} \geq t] &\leq \Pr\left[\max_{j,k} \left|(\widehat{\mathbf{S}}_{xy})_{kj} - (\mathbf{S}_{xy})_{kj}\right| \geq t\right] \\
&\leq \sum_{j,k} \Pr\left[\left|(\widehat{\mathbf{S}}_{xy})_{kj} - (\mathbf{S}_{xy})_{kj}\right| \geq t\right] \\
&\leq pqC \exp\left(-cn \frac{\mu_{\min}^2 t^2}{\sigma_W^2 X_{\max}^2 \max(s_{\max}^2 X_{\max}^2 B_{\max}^2, \sigma_\varepsilon^2)}\right).
\end{aligned}$$

□

Proof of Lemma 14

Proof. Given the definition of $\mathbf{S}_{xy}$

$$\mathbf{S}_{xy} - \mathbf{S}_{xx}\mathbf{B}^* = \frac{1}{n}\mathbf{X}^\top \mathbf{Y} - \frac{1}{n}\mathbf{X}^\top \mathbf{X}\mathbf{B}^* = \frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{B}^* + \boldsymbol{\varepsilon}) - \frac{1}{n}\mathbf{X}^\top \mathbf{X}\mathbf{B}^* = \frac{1}{n}\mathbf{X}^\top \boldsymbol{\varepsilon} \tag{4.42}$$

Hence if we consider the kth row and jth column of this difference for $j = 1, \ldots, q$ and $k = 1, \ldots, p$, it is

$$(\mathbf{S}_{xy})_{kj} - (\mathbf{S}_{xx}\mathbf{B}^*)_{kj} = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_{ij} X_{ik}$$

We can bound it by applying Theorem 2.6.3 general Hoeffding inequality on page 27 of Vershynin (2018)

$$\begin{aligned}\Pr\left[\Big|\sum_{i=1}^{n}(X_{ik}/n)\varepsilon_{ij}\Big| \geq t\right] &< 2\exp\left\{-\frac{ct^2}{(\max_i \|\varepsilon_{ij}\|_{\psi_2})^2\sum_{i=1}^{n}(X_{ik}/n)^2}\right\} \\ &< 2\exp\left\{-\frac{cnt^2}{\sigma_\varepsilon^2 X_{\max}^2}\right\}.\end{aligned}$$

Applying the union bound, we get

$$\Pr(|||\mathbf{S}_{xy} - \mathbf{S}_{xx}\mathbf{B}^*|||_{\max} \geq t) \leq pqC\exp\left[-\frac{cnt^2}{\sigma_\varepsilon^2 X_{\max}^2}\right]$$

and

$$\Pr(\|(\mathbf{S}_{xy})_{\bullet l} - \mathbf{S}_{xx}\boldsymbol{\beta}_l^*\|_\infty \geq t) \leq pC\exp\left[-\frac{cnt^2}{\sigma_\varepsilon^2 X_{\max}^2}\right].$$

□

Proof of Lemma 15

Proof. Recall that, for a multiplicative measurement error model, we assume the observed matrix is $\mathbf{Z} = \mathbf{Y} \odot \mathbf{W}$ where $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_n)^\top$ is a matrix of multiplicative error. Let $\boldsymbol{\Sigma}_W$ be the known population covariance matrix of the measurement errors $\mathbf{W}$ for the multiplicative model. Given $\mathbf{S}_{yy}$ as the sample covariance matrix for the data without any corruption, we have

$$\begin{aligned}\widehat{\mathbf{S}}_{yy} - \mathbf{S}_{yy} &= \frac{1}{n}\mathbf{Z}^\top\mathbf{Z} \oslash (\boldsymbol{\Sigma}_W + \mu_W\mu_W^\top) - \frac{1}{n}\mathbf{Y}^\top\mathbf{Y} \\ &= \frac{1}{n}(\mathbf{Y}\odot\mathbf{W})^\top(\mathbf{Y}\odot\mathbf{W}) \oslash (\boldsymbol{\Sigma}_W + \mu_W\mu_W^\top) - \frac{1}{n}\mathbf{Y}^\top\mathbf{Y} \\ &= \frac{1}{n}(\mathbf{Y}\odot\mathbf{W})^\top(\mathbf{Y}\odot\mathbf{W}) \oslash \mathbb{E}[WW^\top] - \frac{1}{n}\mathbf{Y}^\top\mathbf{Y}, \qquad (4.43)\end{aligned}$$

where the last line is due to the fact that $\mathbb{E}[WW^\top] = \mathrm{Cov}(W) + \mathbb{E}[W]\mathbb{E}[W^\top] = \boldsymbol{\Sigma}_W + \mu_W\mu_W^\top$.

Let Y_{ij} and Y_{ik} be the ith row and jth and kth column of $\mathbf{Y}$ for $i = 1, \ldots, n$ and $j, k = 1, \ldots, q$, and W_{ij} and W_{ik} can be defined similarly. Let $X_{ik'}$ and $X_{ik''}$ be the ith row and k'th and k''th column of $\mathbf{X}$ for $k', k'' = 1, \ldots, p$. Then we can express the (k, j) element of $(\widehat{\mathbf{S}}_{yy} - \mathbf{S}_{yy})$ as

$$(\widehat{\mathbf{S}}_{yy})_{kj} - (\mathbf{S}_{yy})_{kj} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_{ij}W_{ij}Y_{ik}W_{ik}}{\mathbb{E}(W_jW_k)} - \frac{1}{n}\sum_{i=1}^{n} Y_{ij}Y_{ik}$$

where $\mathbb{E}(W_jW_k) = (\mathbb{E}[WW^\top])_{jk}$. Now by plugging in the true model

$$Y_{ij} = \sum_{k' \in S_j} X_{ik'}\beta^*_{k'j} + \varepsilon_{ij},$$
$$Y_{ik} = \sum_{k'' \in S_k} X_{ik''}\beta^*_{k''k} + \varepsilon_{ik}$$

we get

$$\begin{aligned}
&(\widehat{\mathbf{S}}_{yy})_{kj} - (\mathbf{S}_{yy})_{kj} \\
=&\frac{1}{\mathbb{E}(W_jW_k)}\left[\frac{1}{n}\sum_{i=1}^{n}(\sum_{k' \in S_j} X_{ik'}\beta^*_{k'j} + \varepsilon_{ij})(\sum_{k'' \in S_k} X_{ik''}\beta^*_{k''k} + \varepsilon_{ik})W_{ij}W_{ik}\right. \\
&\left. - \frac{1}{n}\sum_{i=1}^{n}(\sum_{k' \in S_j} X_{ik'}\beta^*_{k'j} + \varepsilon_{ij})(\sum_{k'' \in S_k} X_{ik''}\beta^*_{k''k} + \varepsilon_{ik})\mathbb{E}(W_jW_k)\right] \\
=&\frac{1}{\mathbb{E}(W_jW_k)}\left[\frac{1}{n}\sum_{i=1}^{n}(\sum_{k' \in S_j} X_{ik'}\beta^*_{k'j} + \varepsilon_{ij})(\sum_{k'' \in S_k} X_{ik''}\beta^*_{k''k} + \varepsilon_{ik})(W_{ij}W_{ik} - \mathbb{E}(W_jW_k))\right].
\end{aligned}$$

Since both W_{ij} and W_{ik} are sub-Gaussian with parameter σ_W^2, their product $W_{ij}W_{ik}$ is sub-exponential with

$$\|W_{ij}W_{ik}\|_{\psi_1} \leq \|W_{ij}\|_{\psi_2}\|W_{ik}\|_{\psi_2}.$$

Therefore,

$$U_{ikj} \coloneqq W_{ij}W_{ik} - \mathbb{E}(W_jW_k)$$

is a centered sub-exponential random variable, which has

$$\|U_{ikj}\|_{\psi_1} = \|W_{ij}W_{ik} - \mathbb{E}(W_jW_k)\|_{\psi_1} = C\|W_{ij}W_{ik}\|_{\psi_1} \leq C\|W_{ij}\|_{\psi_2}\|W_{ik}\|_{\psi_2}.$$

Hence

$$\begin{aligned}
&\left|(\widehat{\mathbf{S}}_{yy})_{kj} - (\mathbf{S}_{yy})_{kj}\right| \\
&\leq \frac{1}{\mathbb{E}(W_jW_k)}\left|\frac{1}{n}\sum_{i=1}^{n}\left\{U_{ikj}\sum_{k'\in S_j}\sum_{k''\in S_k}X_{ik'}\beta^*_{k'j}X_{ik''}\beta^*_{k''k}\right.\right. \\
&\qquad \left.\left. + U_{ikj}\sum_{k'\in S_j}X_{ik'}\beta^*_{k'j}\varepsilon_{ik} + U_{ikj}\sum_{k''\in S_k}X_{ik''}\beta^*_{k''k}\varepsilon_{ij} + U_{ikj}\varepsilon_{ij}\varepsilon_{ik}\right)\right\}\right| \\
&\leq \frac{1}{m_{\min}}\left\{X^2_{\max}B^2_{\max}s^2_{\max}\left|\underbrace{\sum_{i=1}^{n}(1/n)U_{ikj}}_{T_1}\right| + X_{\max}B_{\max}s_{\max}\left|\underbrace{\sum_{i=1}^{n}(1/n)U_{ikj}\varepsilon_{ik}}_{T_2}\right|\right. \\
&\qquad \left. + X_{\max}B_{\max}s_{\max}\left|\underbrace{\sum_{i=1}^{n}(1/n)U_{ikj}\varepsilon_{ij}}_{T_3}\right| + \left|\underbrace{\sum_{i=1}^{n}(1/n)U_{ikj}\varepsilon_{ij}\varepsilon_{ik}}_{T_4}\right|\right\}
\end{aligned}$$

where $m_{\min} = \min_{j,k}|\mathbb{E}(W_jW_k)| > 0$, $s_{\max} = \max_j s_j$, $X_{\max} = \max_{i,k}|X_{ik}| < \infty$, and $B_{\max} = \max_{k',j}|\beta^*_{k'j}|$. Notice that in the above formula, within each term, we have multiple products of sub-Gaussian random variables. Now we bound terms T_1, T_2, T_3 and T_4, separately.

Term T_1 is the average of n independent, mean zero, sub-exponential random variables. Therefore

$$\begin{aligned}
\Pr\left(|T_1| \geq t\right) &= \Pr\left(\left|\sum_{i=1}^{n}(1/n)U_{ikj}\right| \geq t\right) \\
&\leq 2\exp\left[-cn\min\left(\frac{t^2}{\max_i\|U_{ikj}\|^2_{\psi_1}}, \frac{t}{\max_i\|U_{ikj}\|_{\psi_1}}\right)\right] \\
&\leq C\exp\left[-cn\min\left(\frac{t^2}{\max_i\|W_{ij}\|^2_{\psi_2}\|W_{ik}\|^2_{\psi_2}}, \frac{t}{\max_i\|W_{ij}\|_{\psi_2}\|W_{ik}\|_{\psi_2}}\right)\right]
\end{aligned}$$

$$\leq C \exp\left[-cn \min\left(\frac{t^2}{\sigma_W^4}, \frac{t}{\sigma_W^2}\right)\right]$$

Now we look at term T_2 and T_3. We need to bound the product $U_{ikj}\varepsilon_{ik}$. We know that sub-Gaussian variable ε_{ik} is sub-exponential, since

$$\Pr\left(|\varepsilon_{ik}| \geq t\right) \leq 2\exp(-t^2/\sigma_\varepsilon^2) \leq 2\exp(-t/\sigma_\varepsilon^2).$$

We can use Lemma A.1 of Götze et al. (2021), which states that the product of D sub-exponential random variables has a α/D-sub-exponential tail with the Orlicz norm

$$\Big\| \prod_{i=1}^{D} X_i \Big\|_{\psi_{\frac{\alpha}{D}}} \leq \prod_{i=1}^{D} \|X_i\|_{\psi_\alpha}. \tag{4.44}$$

Specifically, in our case $D = 2$ and $\alpha = 1$ in (4.44), we obtain that the product of two sub-exponential is a $1/2$-sub-exponential with

$$\|U_{ikj}\varepsilon_{ik}\|_{\psi_{1/2}} \leq \|U_{ikj}\|_{\psi_1}\|\varepsilon_{ik}\|_{\psi_1} = \|W_{ij}\|_{\psi_2}\|W_{ij}\|_{\psi_2}\|\varepsilon_{ik}\|_{\psi_2}^2.$$

Meanwhile, we know $U_{ikj}\varepsilon_{ik}$ is centered due to $\mathbb{E}(U_{ikj}\varepsilon_{ik}) = \mathbb{E}(U_{ikj})\mathbb{E}(\varepsilon_{ik}) = 0$ and independence of U_{ikj} and ε_{ik}. We apply Corollary 1.4 of Götze et al. (2021)

$$\begin{aligned}
\Pr\left(|T_2| \geq t\right) &= \Pr\left(\left|\sum_{i=1}^{n}(1/n)U_{ikj}\varepsilon_{ik}\right| \geq t\right) \\
&\leq 2\exp\left(-c\min\left(\frac{nt^2}{\|U_{ikj}\varepsilon_{ik}\|_{\psi_{1/2}}^2}, \frac{n^{1/2}t^{1/2}}{\|U_{ikj}\varepsilon_{ik}\|_{\psi_{1/2}}^{1/2}}\right)\right) \\
&\leq 2\exp\left(-c\min\left(\frac{nt^2}{\|W_{ij}\|_{\psi_2}^2\|W_{ij}\|_{\psi_2}^2\|\varepsilon_{ik}\|_{\psi_2}^4}, \frac{n^{1/2}t^{1/2}}{\|W_{ij}\|_{\psi_2}^{1/2}\|W_{ij}\|_{\psi_2}^{1/2}\|\varepsilon_{ik}\|_{\psi_2}}\right)\right) \\
&\leq 2\exp\left(-c\min\left(\frac{nt^2}{\sigma_W^4\sigma_\varepsilon^4}, \frac{n^{1/2}t^{1/2}}{\sigma_W\sigma_\varepsilon}\right)\right)
\end{aligned}$$

The bound for term T_3 is the same

$$\Pr\left(|T_3| \geq t\right) = 2\exp\left(-c\min\left(\frac{nt^2}{\sigma_W^4\sigma_\varepsilon^4}, \frac{n^{1/2}t^{1/2}}{\sigma_W\sigma_\varepsilon}\right)\right).$$

Now we look at term T_4, in which $\varepsilon_{ij}\varepsilon_{ik}$ as the product of two sub-Gaussian is sub-exponential with

$$\|\varepsilon_{ij}\varepsilon_{ik}\|_{\psi_1} \leq \|\varepsilon_{ij}\|_{\psi_2}\|\varepsilon_{ik}\|_{\psi_2}$$

Therefore, we can view $U_{ikj}\varepsilon_{ij}\varepsilon_{ik}$ as the product of two sub-exponential random variables with

$$\|U_{ikj}\varepsilon_{ij}\varepsilon_{ik}\|_{\psi_{1/2}} \leq \|U_{ikj}\|_{\psi_1}\|\varepsilon_{ij}\varepsilon_{ik}\|_{\psi_1} \leq \|W_{ij}\|_{\psi_2}\|W_{ik}\|_{\psi_2}\|\varepsilon_{ij}\|_{\psi_2}\|\varepsilon_{ik}\|_{\psi_2}.$$

$$\begin{aligned}
\Pr\left(|T_4| \geq t\right) &= \Pr\left(\left|\sum_{i=1}^{n}(1/n)U_{ikj}\varepsilon_{ij}\varepsilon_{ik}\right| \geq t\right) \\
&\leq 2\exp\left(-c\min\left(\frac{nt^2}{\|U_{ikj}\varepsilon_{ij}\varepsilon_{ik}\|_{\psi_{1/2}}^2}, \frac{n^{1/2}t^{1/2}}{\|U_{ikj}\varepsilon_{ij}\varepsilon_{ik}\|_{\psi_{1/2}}^{1/2}}\right)\right) \\
&\leq 2\exp\left(-c\min\left(\frac{nt^2}{\|W_{ij}\|_{\psi_2}^2\|W_{ik}\|_{\psi_2}^2\|\varepsilon_{ij}\|_{\psi_2}^2\|\varepsilon_{ik}\|_{\psi_2}^2}, \frac{n^{1/2}t^{1/2}}{\|W_{ij}\|_{\psi_2}^{1/2}\|W_{ik}\|_{\psi_2}^{1/2}\|\varepsilon_{ij}\|_{\psi_2}^{1/2}\|\varepsilon_{ik}\|_{\psi_2}^{1/2}}\right)\right) \\
&\leq 2\exp\left(-c\min\left(\frac{nt^2}{\sigma_W^4\sigma_\varepsilon^4}, \frac{n^{1/2}t^{1/2}}{\sigma_W\sigma_\varepsilon}\right)\right).
\end{aligned}$$

Define the event

$$\begin{aligned}
\mathcal{A}_1 &= \left\{\frac{X_{\max}^2 B_{\max}^2 s_{\max}^2}{m_{\min}}\,|T_1| \geq t/4\right\} \\
\mathcal{A}_2 &= \left\{\frac{X_{\max} B_{\max} s_{\max}}{m_{\min}}\,|T_2| \geq t/4\right\} \\
\mathcal{A}_3 &= \left\{\frac{X_{\max} B_{\max} s_{\max}}{m_{\min}}\,|T_3| \geq t/4\right\} \\
\mathcal{A}_4 &= \left\{\frac{1}{m_{\min}}\,|T_4| \geq t/4\right\}
\end{aligned}$$

and the event $\mathcal{G} = \{|(\widehat{\mathbf{S}}_{yy})_{kj} - (\mathbf{S}_{yy})_{kj}| < t\}$. Then by Boole's inequality

$$
\begin{aligned}
\Pr(\mathcal{G}) &= \Pr(|(\widehat{\mathbf{S}}_{yy})_{kj} - (\mathbf{S}_{yy})_{kj}| < t) \\
&\geq \Pr(\mathcal{A}_1^c \cap \mathcal{A}_2^c \cap \mathcal{A}_3^c \cap \mathcal{A}_4^c) \\
&= \Pr[(\mathcal{A}_1 \cup \mathcal{A}_2 \cup \mathcal{A}_3 \cup \mathcal{A}_4)^c] \\
&\geq 1 - \Pr[(\mathcal{A}_1 \cup \mathcal{A}_2 \cup \mathcal{A}_3 \cup \mathcal{A}_4)] \\
&\geq 1 - \Pr(\mathcal{A}_1) - \Pr(\mathcal{A}_2) - \Pr(\mathcal{A}_3) - \Pr(\mathcal{A}_4) \\
&\geq 1 - \Pr\left\{|T_1| \geq \frac{tm_{\min}}{4X_{\max}^2 B_{\max}^2 s_{\max}^2}\right\} - \Pr\left\{|T_2| \geq \frac{tm_{\min}}{4X_{\max} B_{\max} s_{\max}}\right\} \\
&\qquad - \Pr\left\{|T_3| \geq \frac{tm_{\min}}{4X_{\max} B_{\max} s_{\max}}\right\} - \Pr\left\{|T_4| \geq tm_{\min}/4\right\} \\
&\geq 1 - C\exp\left(-c\min\left(\frac{nt^2 m_{\min}^2}{X_{\max}^4 B_{\max}^4 s_{\max}^4 \sigma_W^4}, \frac{ntm_{\min}}{X_{\max}^2 B_{\max}^2 s_{\max}^2 \sigma_W^2}\right)\right) \\
&\qquad - C\exp\left(-c\min\left(\frac{nt^2 m_{\min}^2}{X_{\max}^2 B_{\max}^2 s_{\max}^2 \sigma_W^4 \sigma_\varepsilon^4}, \frac{n^{1/2}t^{1/2}m_{\min}^{1/2}}{X_{\max}^{1/2} B_{\max}^{1/2} s_{\max}^{1/2} \sigma_W \sigma_\varepsilon}\right)\right) \\
&\qquad - C\exp\left(-c\min\left(\frac{nt^2 m_{\min}^2}{X_{\max}^2 B_{\max}^2 s_{\max}^2 \sigma_W^4 \sigma_\varepsilon^4}, \frac{n^{1/2}t^{1/2}m_{\min}^{1/2}}{X_{\max}^{1/2} B_{\max}^{1/2} s_{\max}^{1/2} \sigma_W \sigma_\varepsilon}\right)\right) \\
&\qquad - C\exp\left(-c\min\left(\frac{nt^2 m_{\min}^2}{\sigma_W^4 \sigma_\varepsilon^4}, \frac{n^{1/2}t^{1/2}m_{\min}^{1/2}}{\sigma_W \sigma_\varepsilon}\right)\right)
\end{aligned}
$$

Now, we can further simplify the terms involved inside the exponentiation of the right hand side of the above inequality if we assume t satisfies some additional conditions. Specifically, if $t \leq t_0^{(a)} \coloneqq X_{\max}^2 B_{\max}^2 s_{\max}^2 \sigma_W^2 / m_{\min}$, then we have

$$
\frac{nt^2 m_{\min}^2}{X_{\max}^4 B_{\max}^4 s_{\max}^4 \sigma_W^4} \leq \frac{ntm_{\min}}{X_{\max}^2 B_{\max}^2 s_{\max}^2 \sigma_W^2}.
$$

If $t \leq t_0^{(b)} \coloneqq X_{\max} B_{\max} s_{\max} \sigma_W^2 \sigma_\varepsilon^2 / (n^{1/3} m_{\min})$, then

$$
\frac{nt^2 m_{\min}^2}{X_{\max}^2 B_{\max}^2 s_{\max}^2 \sigma_W^4 \sigma_\varepsilon^4} \leq \frac{n^{1/2}t^{1/2}m_{\min}^{1/2}}{X_{\max}^{1/2} B_{\max}^{1/2} s_{\max}^{1/2} \sigma_W \sigma_\varepsilon}
$$

and if $t \leq t_0^{(c)} \coloneqq \sigma_W^2 \sigma_\varepsilon^2 / (n^{1/3} m_{\min})$, then

$$\frac{nt^2 m_{\min}^2}{\sigma_W^4 \sigma_\varepsilon^4} \leq \frac{n^{1/2} t^{1/2} m_{\min}^{1/2}}{\sigma_W \sigma_\varepsilon}$$

Therefore, if we assume that

$$\begin{aligned} t \leq & t_0^{(2)} \coloneqq \min(t_0^{(a)}, t_0^{(b)}, t_0^{(c)}) \\ & = \min(X_{\max}^2 B_{\max}^2 s_{\max}^2 \sigma_W^2 / m_{\min}, X_{\max} B_{\max} s_{\max} \sigma_W^2 \sigma_\varepsilon^2 / m_{\min} n^{1/3}, \sigma_W^2 \sigma_\varepsilon^2 / m_{\min} n^{1/3}) \end{aligned}$$

then the lower bound for $\Pr(\mathcal{G})$ can be simplified as

$$\begin{aligned} & \Pr\left[\mathcal{G}\right] \\ \geq & 1 - C \exp\left(- \frac{cnt^2 m_{\min}^2}{X_{\max}^4 B_{\max}^4 s_{\max}^4 \sigma_W^4} \right) - C \exp\left(- \frac{cnt^2 m_{\min}^2}{X_{\max}^2 B_{\max}^2 s_{\max}^2 \sigma_W^4 \sigma_\varepsilon^4} \right) \\ & - C \exp\left(- \frac{cnt^2 m_{\min}^2}{X_{\max}^2 B_{\max}^2 s_{\max}^2 \sigma_W^4 \sigma_\varepsilon^4} \right) - C \exp\left(- \frac{cnt^2 m_{\min}^2}{\sigma_W^4 \sigma_\varepsilon^4} \right) \\ \geq & 1 - 4C \max\left\{ \exp\left(- \frac{cnt^2 m_{\min}^2}{X_{\max}^4 B_{\max}^4 s_{\max}^4 \sigma_W^4} \right), \exp\left(- \frac{cnt^2 m_{\min}^2}{X_{\max}^2 B_{\max}^2 s_{\max}^2 \sigma_W^4 \sigma_\varepsilon^4} \right), \exp\left(- \frac{cnt^2 m_{\min}^2}{\sigma_W^4 \sigma_\varepsilon^4} \right) \right\} \\ \geq & 1 - 4C \left\{ \exp\left(- \frac{cnt^2 m_{\min}^2}{\max\left\{ X_{\max}^4 B_{\max}^4 s_{\max}^4 \sigma_W^4, X_{\max}^2 B_{\max}^2 s_{\max}^2 \sigma_W^4 \sigma_\varepsilon^4, \sigma_W^4 \sigma_\varepsilon^4 \right\}} \right) \right\} \\ \geq & 1 - 4C \left\{ \exp\left(- \frac{cnt^2 m_{\min}^2}{\sigma_W^4 \max\left\{ X_{\max}^4 B_{\max}^4 s_{\max}^4, X_{\max}^2 B_{\max}^2 s_{\max}^2 \sigma_\varepsilon^4, \sigma_\varepsilon^4 \right\}} \right) \right\}. \end{aligned}$$

Applying the union bound, we get

$$\Pr(\|\widehat{\mathbf{S}}_{yy} - \mathbf{S}_{yy}\|_{\max} \geq t) \leq 4q^2 C \left\{ \exp\left(- \frac{cnt^2 m_{\min}^2}{\sigma_W^4 \max\left\{ X_{\max}^4 B_{\max}^4 s_{\max}^4, X_{\max}^2 B_{\max}^2 s_{\max}^2 \sigma_\varepsilon^4, \sigma_\varepsilon^4 \right\}} \right) \right\}.$$

□

Proof of Lemma 16

Proof. The deviation $\|\widehat{\mathbf{S}}_{\varepsilon\varepsilon} - \mathbf{\Sigma}^*_{\varepsilon\varepsilon}\|_{\max}$ can be decomposed and upper bounded as

$$\begin{aligned}
\|\widehat{\mathbf{S}}_{\varepsilon\varepsilon} - \mathbf{\Sigma}^*_{\varepsilon\varepsilon}\|_{\max} &= \|\widehat{\mathbf{S}}_{yy} - \widehat{\mathbf{B}}^{(1)\top}\mathbf{S}_{xx}\widehat{\mathbf{B}}^{(1)} - \mathbf{\Sigma}^*_{\varepsilon\varepsilon}\|_{\max} \\
&= \|\widehat{\mathbf{S}}_{yy} - \mathbf{S}_{yy} + \mathbf{S}_{yy} - \widehat{\mathbf{B}}^{(1)\top}\mathbf{S}_{xx}\widehat{\mathbf{B}}^{(1)} - \mathbf{\Sigma}^*_{\varepsilon\varepsilon}\|_{\max} \\
&= \|\widehat{\mathbf{S}}_{yy} - \mathbf{S}_{yy} + (1/n)\mathbf{Y}^\top\mathbf{Y} - \widehat{\mathbf{B}}^{(1)\top}\mathbf{S}_{xx}\widehat{\mathbf{B}}^{(1)} - \mathbf{\Sigma}^*_{\varepsilon\varepsilon}\|_{\max} \\
&\leq \|\widehat{\mathbf{S}}_{yy} - \mathbf{S}_{yy}\|_{\max} + \|(1/n)(\mathbf{X}\mathbf{B}^* + \boldsymbol{\varepsilon})^\top(\mathbf{X}\mathbf{B}^* + \boldsymbol{\varepsilon}) - \widehat{\mathbf{B}}^{(1)\top}\mathbf{S}_{xx}\widehat{\mathbf{B}}^{(1)} - \mathbf{\Sigma}^*_{\varepsilon\varepsilon}\|_{\max} \\
&\leq \underbrace{\|(\widehat{\mathbf{S}}_{yy} - \mathbf{S}_{yy})\|_{\max}}_{T_1} + \underbrace{\|\widehat{\mathbf{B}}^{(1)\top}\mathbf{S}_{xx}\widehat{\mathbf{B}}^{(1)} - \mathbf{B}^{*\top}\mathbf{S}_{xx}\mathbf{B}^*\|_{\max}}_{T_2} \\
&\quad + \underbrace{\|(2/n)\boldsymbol{\varepsilon}^\top\mathbf{X}\mathbf{B}^*\|_{\max}}_{T_3} + \underbrace{\|(1/n)\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon} - \mathbf{\Sigma}^*_{\varepsilon\varepsilon}\|_{\max}}_{T_4},
\end{aligned}$$

by applying a series of triangular inequalities. In the following, we bound each term on the right hand side of the above inequality separately. The first term T_1 can be bound by applying Lemma 15 by setting

$$\begin{aligned}
t_1 &\coloneqq \sigma_W^2 \max\left\{X_{\max}^2 B_{\max}^2 s_{\max}^2/m_{\min}, X_{\max}B_{\max}s_{\max}\sigma_\varepsilon^2/m_{\min}, \sigma_\varepsilon^2/m_{\min}\right\}\sqrt{\frac{\log(q^2)}{n}} \\
&\leq t_0^{(2)} \coloneqq \min(t_0^{(a)}, t_0^{(b)}, t_0^{(c)})
\end{aligned}$$

we obtain the following tail bound

$$\begin{aligned}
&\Pr(T_1 \leq t_1) \\
\geq &\Pr(\|\widehat{\mathbf{S}}_{yy} - \mathbf{S}_{yy}\|_{\max} \leq t_1) \\
\geq &1 - C\exp(-c\log q^2)
\end{aligned}$$

The second term T_2 can be simplified as follows:

$$\begin{aligned}
\|(\widehat{\mathbf{B}}^{(1)\top}\mathbf{S}_{xx}\widehat{\mathbf{B}}^{(1)} - \mathbf{B}^{*\top}\mathbf{S}_{xx}\mathbf{B}^*)\|_{\max} &= \|(\widehat{\mathbf{B}}^{(1)} - \mathbf{B}^*)^\top\mathbf{S}_{xx}(\widehat{\mathbf{B}}^{(1)} - \mathbf{B}^*) + 2[(\widehat{\mathbf{B}}^{(1)} - \mathbf{B}^*)^\top\mathbf{S}_{xx}\mathbf{B}^*]\|_{\max} \\
&\leq \|(\widehat{\mathbf{B}}^{(1)} - \mathbf{B}^*)^\top\mathbf{S}_{xx}(\widehat{\mathbf{B}}^{(1)} - \mathbf{B}^*)\|_{\max}
\end{aligned}$$

$$+ 2|||(\widehat{\mathbf{B}}^{(1)} - \mathbf{B}^*)^\top \mathbf{S}_{xx}\mathbf{B}^*|||_{\max}.$$

We treat each term on the right hand side in sequence. Starting with the first term, we have

$$\begin{aligned}
|||(\widehat{\mathbf{B}}^{(1)} - \mathbf{B}^*)^\top \mathbf{S}_{xx}(\widehat{\mathbf{B}}^{(1)} - \mathbf{B}^*)|||_{\max} &\overset{\text{(i)}}{\leq} |||(\widehat{\mathbf{B}}^{(1)} - \mathbf{B}^*)^\top|||_\infty |||\mathbf{S}_{xx}(\widehat{\mathbf{B}}^{(1)} - \mathbf{B}^*)|||_{\max} \\
&\leq |||(\widehat{\mathbf{B}}^{(1)} - \mathbf{B}^*)^\top|||_\infty |||(\widehat{\mathbf{B}}^{(1)} - \mathbf{B}^*)^\top \mathbf{S}_{xx}^\top|||_{\max} \\
&\leq |||(\widehat{\mathbf{B}}^{(1)} - \mathbf{B}^*)^\top|||_\infty |||(\widehat{\mathbf{B}}^{(1)} - \mathbf{B}^*)^\top|||_\infty |||\mathbf{S}_{xx}|||_{\max} \\
&\overset{\text{(ii)}}{=} |||\widehat{\mathbf{B}}^{(1)} - \mathbf{B}^*|||_1 |||\widehat{\mathbf{B}}^{(1)} - \mathbf{B}^*|||_1 |||\mathbf{S}_{xx}|||_{\max} \\
&= |||\mathbf{S}_{xx}|||_{\max} |||\widehat{\mathbf{B}}^{(1)} - \mathbf{B}^*|||_1^2 \\
&\leq X_{\max}^2 |||\widehat{\mathbf{B}}^{(1)} - \mathbf{B}^*|||_1^2 \\
&= X_{\max}^2 \left(\max_{l \in [q]} \|\widehat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l^*\|_1 \right)^2 \\
&\leq X_{\max}^2 \left(\max_{l \in [q]} 12 s_l \lambda_l / \kappa_l \right)^2
\end{aligned}$$

Inquality (i) follows from that for two matrices $\mathbf{A}$ and $\mathbf{B}$

$$\begin{aligned}
|||\mathbf{AB}|||_{\max} &= \max_{i,k} |\sum_j A_{ij} B_{jk}| \\
&\leq \max_{i,k} |\sum_j A_{ij} (\max_j |B_{jk}|)| \\
&= \max_i |\sum_j A_{ij}| \max_{j,k} |B_{jk}| \\
&= \max_k [\max_j |B_{jk}|] \max_i |\sum_j A_{ij}| \\
&= |||\mathbf{A}|||_\infty |||\mathbf{B}|||_{\max},
\end{aligned}$$

and equality (ii) follows from the relationship $|||\mathbf{A}|||_1 = |||\mathbf{A}^\top|||_\infty$. Since by Proposition 1

$$\Pr(\|\widehat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l^*\|_1 \leq 12 s_l \lambda_l / \kappa_l) \geq 1 - C \exp(-c \log p),$$

Hence

$$
\begin{aligned}
&\Pr\left[\|\!|(\widehat{\mathbf{B}}^{(1)}-\mathbf{B}^*)^\top\mathbf{S}_{xx}(\widehat{\mathbf{B}}^{(1)}-\mathbf{B}^*)|\!\|_{\max}\le X_{\max}^2\left(\max_{l\in[q]}12s_l\lambda_l/\kappa_l\right)^2\right]\\
&\ge\Pr\left[X_{\max}^2\left(\max_{l\in[q]}\|\widehat{\boldsymbol{\beta}}_l-\boldsymbol{\beta}_l^*\|_1\right)^2\le X_{\max}^2\left(\max_{l\in[q]}12s_l\lambda_l/\kappa_l\right)^2\right]\\
&=\Pr\left[\max_{l\in[q]}\|\widehat{\boldsymbol{\beta}}_l-\boldsymbol{\beta}_l^*\|_1\le\max_{l\in[q]}12s_l\lambda_l/\kappa_l\right]\\
&=\Pr\left[\cap_{l=1}^q\{\|\widehat{\boldsymbol{\beta}}_l-\boldsymbol{\beta}_l^*\|_1\le\max_{l\in[q]}12s_l\lambda_l/\kappa_l\}\right]\\
&=\Pr\left[\left[\cup_{l=1}^q\{\|\widehat{\boldsymbol{\beta}}_l-\boldsymbol{\beta}_l^*\|_1\ge\max_{l\in[q]}12s_l\lambda_l/\kappa_l\}\right]^c\right]\\
&=1-\Pr\left[\cup_{l=1}^q\{\|\widehat{\boldsymbol{\beta}}_l-\boldsymbol{\beta}_l^*\|_1\ge\max_{l\in[q]}12s_l\lambda_l/\kappa_l\}\right]\\
&\ge1-\sum_{l=1}^q\Pr\left[\|\widehat{\boldsymbol{\beta}}_l-\boldsymbol{\beta}_l^*\|_1\ge\max_{l\in[q]}12s_l\lambda_l/\kappa_l\right]\\
&\ge1-\sum_{l=1}^q\Pr\left[\|\widehat{\boldsymbol{\beta}}_l-\boldsymbol{\beta}_l^*\|_1\ge12s_l\lambda_l/\kappa_l\right]\\
&\ge1-qC\exp(-c\log p)\\
&=1-C\exp(-c\log(pq)).
\end{aligned}
$$

Next, applying Proposition 1 again, we get

$$
\begin{aligned}
|\!\|(\widehat{\mathbf{B}}^{(1)}-\mathbf{B}^*)^\top\mathbf{S}_{xx}\mathbf{B}^*|\!\|_{\max}&\le|\!\|(\widehat{\mathbf{B}}^{(1)}-\mathbf{B}^*)^\top\mathbf{S}_{xx}|\!\|_{\max}|\!\|\mathbf{B}^*|\!\|_1\\
&\le|\!\|(\widehat{\mathbf{B}}^{(1)}-\mathbf{B}^*)^\top|\!\|_\infty|\!\|\mathbf{S}_{xx}|\!\|_{\max}|\!\|\mathbf{B}^*|\!\|_1\\
&\le|\!\|\widehat{\mathbf{B}}^{(1)}-\mathbf{B}^*|\!\|_1|\!\|\mathbf{S}_{xx}|\!\|_{\max}|\!\|\mathbf{B}^*|\!\|_1\\
&\le X_{\max}^2s_{\max}B_{\max}|\!\|\widehat{\mathbf{B}}^{(1)}-\mathbf{B}^*|\!\|_1\\
&=X_{\max}^2s_{\max}B_{\max}\left(\max_{l\in[q]}\|\widehat{\boldsymbol{\beta}}_l-\boldsymbol{\beta}_l^*\|_1\right)\\
&\le X_{\max}^2s_{\max}B_{\max}\max_{l\in[q]}12s_l\lambda_l/\kappa_l
\end{aligned}
$$

with probability

$$
\begin{aligned}
&\Pr\left[\|\!|(\widehat{\mathbf{B}}^{(1)}-\mathbf{B}^*)^\top\mathbf{S}_{xx}\mathbf{B}^*|\!\|_{\max}\le X_{\max}^2 s_{\max}B_{\max}\max_{l\in[q]}12s_l\lambda_l/\kappa_l\right]\\
&\ge\Pr\left[X_{\max}^2 s_{\max}B_{\max}\max_{l\in[q]}\|\widehat{\boldsymbol{\beta}}_l-\boldsymbol{\beta}_l^*\|_1\le X_{\max}^2 s_{\max}B_{\max}\max_{l\in[q]}12s_l\lambda_l/\kappa_l\right]\\
&\ge\Pr\left[\max_{l\in[q]}\|\widehat{\boldsymbol{\beta}}_l-\boldsymbol{\beta}_l^*\|_1\le\max_{l\in[q]}12s_l\lambda_l/\kappa_l\right]\\
&=\Pr\left[\cap_{l=1}^q\{\|\widehat{\boldsymbol{\beta}}_l-\boldsymbol{\beta}_l^*\|_1\le\max_{l\in[q]}12s_l\lambda_l/\kappa_l\}\right]\\
&=1-\Pr\left[\cup_{l=1}^q\{\|\widehat{\boldsymbol{\beta}}_l-\boldsymbol{\beta}_l^*\|_1\ge\max_{l\in[q]}12s_l\lambda_l/\kappa_l\}\right]\\
&\ge1-\sum_{l=1}^q\Pr\left[\|\widehat{\boldsymbol{\beta}}_l-\boldsymbol{\beta}_l^*\|_1\ge\max_{l\in[q]}12s_l\lambda_l/\kappa_l\right]\\
&\ge1-\sum_{l=1}^q\Pr\left[\|\widehat{\boldsymbol{\beta}}_l-\boldsymbol{\beta}_l^*\|_1\ge12s_l\lambda_l/\kappa_l\right]\\
&\le1-C\exp(-c\log(pq)).
\end{aligned}
$$

Now, we bound the third term T_3. Let us consider the jth row and kth column for $j,k=1,\ldots,q$,

$$
\left|\left(\frac{2}{n}\boldsymbol{\varepsilon}^\top\mathbf{X}\mathbf{B}^*\right)_{jk}\right|=\left|\frac{2}{n}\sum_{i=1}^n\sum_{k'\in S_k}X_{ik'}\beta^*_{k'k}\varepsilon_{ij}\right|\le\left|2X_{\max}s_{\max}B_{\max}\sum_{i=1}^n(1/n)\varepsilon_{ij}\right|.
$$

Since ε_{ij} is sub-Gaussian with parameter σ_ε^2, we can bound it by applying Theorem 2.6.3 general Hoeffding inequality on page 27 of Vershynin (2018)

$$
\begin{aligned}
\Pr\left[\Big|\sum_{i=1}^n(1/n)\varepsilon_{ij}\Big|\ge t\right]&<2\exp\left\{-\frac{ct^2}{(\max_i\|\varepsilon_{ij}\|_{\psi_2})^2\sum_{i=1}^n(1/n)^2}\right\}\\
&<2\exp\left\{-\frac{cnt^2}{\sigma_\varepsilon^2}\right\}.
\end{aligned}
$$

Hence,

$$\Pr\left[\Big|\sum_{i=1}^{n}(2X_{\max}s_{\max}B_{\max}/n)\varepsilon_{ij}\Big| \geq t\right] < 2\exp\left\{-\frac{cnt^2}{\sigma_\varepsilon^2 X_{\max}^2 s_{\max}^2 B_{\max}^2}\right\}.$$

Applying the union bound, we get

$$\Pr(|||(2/n)\boldsymbol{\varepsilon}^\top \mathbf{X}\mathbf{B}^*|||_{\max} \geq t) \leq q^2 C\exp\left\{-\frac{cnt^2}{\sigma_\varepsilon^2 X_{\max}^2 s_{\max}^2 B_{\max}^2}\right\}.$$

By setting

$$t_2 \coloneqq \sigma_\varepsilon X_{\max}s_{\max}B_{\max}\sqrt{\frac{\log(q^2)}{n}}$$

we get the tail bound for T_3

$$\Pr(|||(2/n)\boldsymbol{\varepsilon}^\top \mathbf{X}\mathbf{B}^*|||_{\max} \leq t_2) \geq 1 - C\exp(-c\log(q^2)).$$

Now, we bound T_4. For the jth row and kth column of $(1/n)\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon} - \boldsymbol{\Sigma}^*_{\varepsilon\varepsilon}$, we have

$$((1/n)\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon} - \boldsymbol{\Sigma}^*_{\varepsilon\varepsilon})_{jk} = \sum_{i=1}^{n}(1/n)[\varepsilon_{ij}\varepsilon_{ik} - \mathbb{E}(\varepsilon_j\varepsilon_k)]$$

where ε_{ij} and ε_{ik} are sub-Gaussian with parameter σ_ε^2. Their product $\varepsilon_{ij}\varepsilon_{ik}$ is sub-exponential with

$$\|\varepsilon_{ij}\varepsilon_{ik}\|_{\psi_1} \leq \|\varepsilon_{ij}\|_{\psi_2}\|\varepsilon_{ik}\|_{\psi_2}.$$

Therefore, define

$$V_{ikj} \coloneqq \varepsilon_{ij}\varepsilon_{ik} - \mathbb{E}(\varepsilon_j\varepsilon_k)$$

as a centered sub-exponential random variable, which has

$$\|V_{ikj}\|_{\psi_1} = \|\varepsilon_{ij}\varepsilon_{ik} - \mathbb{E}(\varepsilon_j\varepsilon_k)\|_{\psi_1} = C\|\varepsilon_{ij}\varepsilon_{ik}\|_{\psi_1} \leq C\|\varepsilon_{ij}\|_{\psi_2}\|\varepsilon_{ik}\|_{\psi_2}.$$

Therefore

$$\begin{aligned}
\Pr\left(\left|\sum_{i=1}^{n}((1/n)\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}-\boldsymbol{\Sigma}^*_{\varepsilon\varepsilon})_{jk}\right|\geq t\right) &= \Pr\left(\left|\sum_{i=1}^{n}(1/n)V_{ikj}\right|\geq t\right)\\
&\leq 2\exp\left[-cn\min\left(\frac{t^2}{\max_i\|V_{ikj}\|^2_{\psi_1}},\frac{t}{\max_i\|V_{ikj}\|_{\psi_1}}\right)\right]\\
&\leq C\exp\left[-cn\min\left(\frac{t^2}{\max_i\|\varepsilon_{ij}\|^2_{\psi_2}\|\varepsilon_{ik}\|^2_{\psi_2}},\frac{t}{\max_i\|\varepsilon_{ij}\|_{\psi_2}\|\varepsilon_{ik}\|_{\psi_2}}\right)\right]\\
&\leq C\exp\left[-cn\min\left(\frac{t^2}{\sigma^4_\varepsilon},\frac{t}{\sigma^2_\varepsilon}\right)\right]\\
&\leq C\exp\left[-cn\frac{t^2}{\sigma^4_\varepsilon}\right]
\end{aligned}$$

The last inequality holds if we assume that t is chosen satisfying $t\leq t_0^{(3)}\coloneqq\sigma^2_\varepsilon$. Therefore, applying union bound, we get

$$\Pr(|||(1/n)\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}-\boldsymbol{\Sigma}^*_{\varepsilon\varepsilon}|||_{\max}\geq t)\leq q^2C\exp\left[-cn\left(\frac{t^2}{\sigma^4_\varepsilon}\right)\right].$$

Define,

$$t_3\coloneqq\sigma^2_\varepsilon\sqrt{\frac{\log(q^2)}{n}}$$

the requirement $t_3\leq t_0^{(3)}$ implies that $\sqrt{\frac{\log(q^2)}{n}}\leq 1$ must be satisfied. Hence, we get the tail bound for T_4

$$\Pr(|||(1/n)\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}-\boldsymbol{\Sigma}^*_{\varepsilon\varepsilon}|||_{\max}\leq t_3)\geq 1-C\exp(-c\log(q^2)).$$

Define the event

$$\begin{aligned}
\mathcal{B}_1&=\left\{|||\widehat{\mathbf{S}}_{yy}-\mathbf{S}_{yy}|||_{\max}\leq t_1\right\}\\
\mathcal{B}_2&=\left\{|||(\widehat{\mathbf{B}}^{(1)}-\mathbf{B}^*)^\top\mathbf{S}_{xx}(\widehat{\mathbf{B}}^{(1)}-\mathbf{B}^*)|||_{\max}\leq X^2_{\max}\left(\max_{l\in[q]}12s_l\lambda_l/\kappa_l\right)^2\right\}\\
\mathcal{B}_3&=\left\{|||(\widehat{\mathbf{B}}^{(1)}-\mathbf{B}^*)^\top\mathbf{S}_{xx}\mathbf{B}^*|||_{\max}\leq X^2_{\max}s_{\max}B_{\max}\max_{l\in[q]}12s_l\lambda_l/\kappa_l\right\}
\end{aligned}$$

$$\mathcal{B}_4 = \left\{ |||(2/n)\boldsymbol{\varepsilon}^\top \mathbf{X}\mathbf{B}^*|||_{\max} \leq t_2 \right\}$$
$$\mathcal{B}_5 = \left\{ |||(1/n)\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} - \boldsymbol{\Sigma}^*_{\varepsilon\varepsilon}|||_{\max} \leq t_3 \right\},$$

and

$$\Delta = t_1 + X^2_{\max}\left(\max_{l\in[q]} 12 s_l\lambda_l/\kappa_l\right)^2 + X^2_{\max} s_{\max} B_{\max} \max_{l\in[q]} 12 s_l\lambda_l/\kappa_l \tag{4.45}$$

$$\begin{aligned} & t_2 + t_3 \\ &= \sigma_W^2 \max\left\{X^2_{\max}B^2_{\max}s^2_{\max}/m_{\min}, X_{\max}B_{\max}s_{\max}\sigma^2_\varepsilon/m_{\min}, \sigma^2_\varepsilon/m_{\min}\right\}\sqrt{\frac{\log(q^2)}{n}} \\ &\quad + X^2_{\max}\left(\max_{l\in[q]} 12 s_l\lambda_l/\kappa_l\right)^2 + X^2_{\max} s_{\max} B_{\max} \max_{l\in[q]} 12 s_l\lambda_l/\kappa_l + \sigma_\varepsilon X_{\max} s_{\max} B_{\max}\sqrt{\frac{\log(q^2)}{n}} \\ &\quad + \sigma^2_\varepsilon\sqrt{\frac{\log(q^2)}{n}} \end{aligned}$$

and the event $\mathcal{H} = \{|||\widehat{\mathbf{S}}_{\varepsilon\varepsilon} - \boldsymbol{\Sigma}^*_{\varepsilon\varepsilon}|||_{\max} < \Delta\}$. Then by Boole's inequality

$$\begin{aligned} \Pr(\mathcal{H}) &= \Pr(|||\widehat{\mathbf{S}}_{\varepsilon\varepsilon} - \boldsymbol{\Sigma}^*_{\varepsilon\varepsilon}|||_{\max} \leq \Delta) \\ &\geq \Pr(|||\widehat{\mathbf{S}}_{yy} - \mathbf{S}_{yy}|||_{\max} + |||(\widehat{\mathbf{B}}^{(1)} - \mathbf{B}^*)^\top \mathbf{S}_{xx}(\widehat{\mathbf{B}}^{(1)} - \mathbf{B}^*)|||_{\max} \\ &\qquad + |||(\widehat{\mathbf{B}}^{(1)} - \mathbf{B}^*)^\top \mathbf{S}_{xx}\mathbf{B}^*|||_{\max} + |||(2/n)\boldsymbol{\varepsilon}^\top\mathbf{X}\mathbf{B}^*|||_{\max} + |||(1/n)\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon} - \boldsymbol{\Sigma}^*_{\varepsilon\varepsilon}|||_{\max} \leq \Delta) \\ &\geq \Pr(\mathcal{B}_1 \cap \mathcal{B}_2 \cap \mathcal{B}_3 \cap \mathcal{B}_4 \cap \mathcal{B}_5) \\ &= \Pr[(\mathcal{B}_1^c \cup \mathcal{B}_2^c \cup \mathcal{B}_3^c \cup \mathcal{B}_4^c \cup \mathcal{B}_5^c)^c] \\ &\geq 1 - \Pr[(\mathcal{B}_1^c \cup \mathcal{B}_2^c \cup \mathcal{B}_3^c \cup \mathcal{B}_4^c \cup \mathcal{B}_5^c)] \\ &\geq 1 - \Pr(\mathcal{B}_1^c) - \Pr(\mathcal{B}_2^c) - \Pr(\mathcal{B}_3^c) - \Pr(\mathcal{B}_4^c) - \Pr(\mathcal{B}_5^c) \\ &\geq 1 - C\exp(-c\log q^2) - C\exp(-c\log(pq)) - C\exp(-c\log(pq)) \\ &\qquad - C\exp(-c\log(q^2)) - C\exp(-c\log(q^2)) \\ &= 1 - C\exp(-c\log q^2) - C\exp(-c\log(pq)) \end{aligned}$$

□

Chapter 5

Discussion

In this chapter, we summarize and discuss the work developed in the earlier chapters of this thesis. In Chapter 2 we provided a comprehensive literature review of the works that lead to Chapter 3 and 4. Specifically, we discuss the classical literature and methods developed for precision matrix estimation in a high-dimensional setting when the data are fully observed. Next, we discussed the consequences of having measurement error being present in the data, in additive form or multiplicative form. Specifically, when there is additive noise or missing data which is a special case of multiplicative measurement error, the objective function tends to be unbounded from below and the problem does not remain convex anymore. Moreover, the sample covariance matrix may not remain positive semi-definite either and as a consequence, might have zero or negative eigenvalues. Many approaches have been suggested to tackle this problem. A noise-corrected non-convex approach is popularly used to estimate the precision matrix where unbiased surrogate estimates are proposed while the objective function still remains non-convex, but an additional side constraint is added to it and solved using projected gradient descent method (Fan et al., 2019; Loh and Wainwright, 2012).

In this thesis, we have proposed an approach to estimate the precision matrix in the presence of corrupted data while preserving the convexity of the objective function.

Inspired by the CoCoLasso methods in the regression setting in the presence of noisy data (Datta and Zou, 2017), we proposed the CoGlasso algorithm to estimate the precision matrix by projecting the unbiased surrogate estimate of the sample covariance matrix to the nearest positive semi-definite cone and use it as a plug-in estimate in the objective function which essentially converts an unbounded objective function problem to a convex optimization problem. The idea of projecting the surrogate covariance matrix has been proposed before in estimating precision matrix when the data are corrupted, but the theoretical guarantees have not been studied properly in this setting for high-dimensional data. On the other hand, the theoretical guarantees for the fully observed data scenario for precision matrix estimation have been extensively studied in Ravikumar et al. (2011).

In this thesis, in Chapter 3, we have laid down the framework of our method and performed a rigorous theoretical study along with deviation bounds of the estimated precision matrix from the truth in elementwise maximum norm under four different scenarios. We have considered two different tail conditions for both the unobserved data, $\mathbf{X}$ and the measurement error, $\mathbf{W}$, namely, exponential-type and polynomial-type tails. We also considered the cases when the measurement error is additive as well as when there is missing data for two different tail conditions, resulting into four distinct scenarios. The main result of this chapter is presented in Theorem 1. In this theorem, we can see that the deviation bound between the estimated and the true precision matrix looks similar to the clean data case as shown in Ravikumar et al. (2010). However, the bound varies by the quantity $\bar{\delta}_{f_*}(n, p^\gamma)$, for the aforementioned four scenarios and we provided expressions for this quantity when $\mathbf{X}$ and $\mathbf{W}$ follows a multivariate Gaussian case (exponential-type tail with $a = 2$) and in the case of polynomial tails. To prove Theorem 1, we required to prove Lemmas 1, 2, 3 and 4. To our knowledge, the proofs of these Lemmas and parts of Theorem 1 where we modified the steps to incorporate the deviation between the surrogate estimate for the corrupted data to the sample covariance matrix for the clean data are original contributions of this thesis.

In our work, following Ravikumar et al. (2010), we derived the consistency bounds under the mutual incoherence condition. Assuming the sizes of the entries in the true covariance matrix, $\kappa_{\mathbf{\Gamma}^*}$ and the inverse Hessian, $\kappa_{\mathbf{\Sigma}^*}$ and the incoherence parameter, α defined in Section 3.5.1 to be constants as a function of the sample size, n, number of nodes, p and the number of maximum number of non-zero elements per row/column, d, we have the elementwise ℓ_∞ bound $|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_{\max} = \mathcal{O}(\bar{\delta}_{f_*}(n, p^\gamma))$, so that the inverse tails functions $\bar{\delta}_{f_*}(n, p^\gamma)$ defined in the Remarks under Theorem 1 specify the rate of convergence in the elementwise ℓ_∞-norm. We also derived the model selection consistency bound in Theorem 2, which does not differ from the clean case scenario as shown in Ravikumar et al. (2010) other than in terms of the expression of $\bar{\delta}_{f_*}(n, p^\gamma)$. The rates in terms of the Frobenius and spectral norm are also established in this chapter. For completeness, we also provided consistency results in the Technical Details section, however these results were similar to the results shown in Ravikumar et al. (2010) for the clean graphical Lasso case.

We assumed the irrepresentability condition or mutual incoherence condition for the graphical Lasso problem with measurement error similar to Ravikumar et al. (2011). This is a necessary assumption to establish the model selection consistency of the estimator. However, it is a strong assumption and hard to check in practice. Some alternative approaches to that can be explored to impose rather weaker conditions for this estimation problem. For example, Johnson et al. (2012) proposed two greedy approaches which learn the full structure of the model with high probability given just $\mathcal{O}(d \log p)$ samples, whereas graphical Lasso requires $\mathcal{O}(d^2 \log p)$ samples. They also showed that their imposed restricted eigenvalue and smoothness conditions were weaker than the irrepresentable condition. Zhang and Zou (2012) proposed the D-trace loss for the estimation of precision matrix under slightly different irrepresentability condition and compared their work with Ravikumar et al. (2011). The choice of irrepresentability condition is an open problem even in the fully observed data. The aforementioned works deal with complete data case,

therefore, it would be an interesting direction to study these techniques to impose a weaker condition for the corrupted data scenarios.

Our CoGlasso approach, is easy to understand and implement, has solid theoretical foundations and shares many properties with the clean graphical Lasso method which is well studied in literature. Specifically, our algorithm can be solved using any graphical lasso algorithm, such as `GLASSO` (Friedman et al., 2008) and `QUIC` (Hsieh et al., 2014). Therefore, the numerical stability of these algorithms are shared by our proposed method.

We have assumed that the certain parameters of the measurement error model are known for the purpose of simplification of our model. However, in practice, they may not be known and an estimate based on the data would be required to proceed with the methodology. Moreover, our assumed measurement error structure is quite simplified but in practice, more complex model based measurement error models might be required. As demonstrated in Loh and Wainwright (2012), one simplified method is to assume that the covariance structure of the measurement error, $\mathbf{\Sigma}_W$ is estimated from independent observations of the noise and the sample covariance matrix is used as an estimate of the unknown covariance structure. They also showed that the theoretical guarantees continue to hold under such estimation. More sophisticated ways to estimate the unknown parameters in different measurement error models are well studied by Carroll et al. (2006). Specifically, an estimator can be formed for $\mathbf{\Sigma}_W$ by assuming that we observe k_i replicate measurements of the corrupted observations $Z_{i1}, \ldots, Z_{ik}$ for each x_i and form an estimator

$$\widehat{\mathbf{\Sigma}}_W = \frac{\sum_{i=1}^{n} \sum_{j=1}^{k_i} (Z_{ij} - \bar{Z}_{i\bullet})(Z_{ij} - \bar{Z}_{i\bullet})^\top}{\sum_{i=1}^{n} (k_i - 1)}.$$

Based on this estimator, we can form the surrogate estimators and proceed with the analysis. Sensitivity analysis of the estimators can be performed by considering different degrees of mismeasurement. Another alternative Bayesian approach can be taken by imposing a

prior distribution on the parameter. However, this must be carefully studied under the context of our problem since our methodological development is mainly frequentist.

Comparing the convex and non-convex approaches, we noticed that the two methods are fairly competitive. When the signal is stronger, the projected methods tend to perform better in terms of model selection compared to the non-convex method (Fan et al., 2019). In terms of norm error, with strong signal of the data, CoGlasso performs significantly better than the non-convex approach. When the signal is weaker, the performace of both the methods deteriorate. Specifically, CoGlasso tend to produce a high false negative rate whereas the alternative non-convex approach tend to produce a high false positive rate. One of the reasons that could deteriorate the performance of CoGlasso could be that it depends on the positive semidefinite projection of the surrogate estimator of the sample covariance matrix and therefore pays a cost in terms of efficiency due to the loss of information in the projection.

We have demonstrated the superior performance of our method over the non-convex approach in Fan et al. (2019) by simulation studies for a number of simulation settings. This could be due to the fact that no additional prior information of the true parameter is required for our method as opposed to the non-convex type approaches. The non-convex approach depends on some crucial information on the hidden parameters to satisfy the restricted eigenvalue condition and in order to have the desirable bounds. In terms of algorithm, both the non-convex methods proposed in Loh and Wainwright (2012) and Fan et al. (2019), use iterative algorithms which heavily depend on the choice of such hidden parameters as well as some knowledge of the step sizes for the iterations, which makes the convergence process complicated. Despite the theoretical guarantees, the implementation remains difficult for such non-convex approaches (Datta and Zou, 2017).

From the simulation study, we see that the method performs well when the graph is very sparse, for example when the true precision matrix represented a chain graph, in both additive error and missing data setting. However, when the graph becomes denser,

our method tended to give a high false negative rate. However, the other method we compared with tends to show poor performance in the case of a denser graph as well.

We have utilized a few simulation settings to demonstrate our method in this thesis, however, the theory we proposed allows for more general settings. We have only shown simulations when the data are multivariate Gaussian. Many other settings where the data are not sub-Gaussian can be explored in the simulation settings. We have only explored two types of graph structures in this thesis, namely the chain graph and the Erdös-Rényi graphs. A set of more complicated settings for the precision matrix can be explored both in terms of the graph structure of the original problem, or by adding a more complicated structure for the noise. A more extensive comparative study can be performed to see how our method performs compared to classical ways to solve missing data problems such as imputations or EM-algorithm based techniques. We demonstrated our method using a real data in the missing data scenario which has a similar sample size and the dimension as shown in the simulation settings. Since the underlying graph structure of the problem is unknown, it was difficult to compare the performance of the proposed methods to the truth.

In order to tune the regularization parameters, we used the cross-validation and the BIC criterion, however, only cross-validation technique seemed to have performed well. It would be interesting to study the role of BIC and propose a modification of the BIC criterion for these corrupted data scenarios, especially when some portion of the data are missing. Since the BIC criterion penalizes the negative log likelihood on the number of observations used in the analysis, it seems natural to impose some sort of adjustments in the penalty part when we do not observe the whole data.

The theme of precision matrix estimation in the presence of noisy data are shared between the two chapters, even though it has been specifically developed in Chapter 3. In Chapter 4, we have explored a multivariate regression setting when the covariates are fully observed and there is missing data in the responses only when the random error and

the measurement error are assumed to be obtained from sub-Gaussian distributions. In Chapter 2, we provided a thorough literature review of the methods developed to jointly estimate precision matrix and the regression coefficients for multivariate regression in a high-dimensional setting. The regression case of a single response variable with corrupted covariates in sparse high-dimensional setting is well studied (Datta and Zou, 2017; Loh and Wainwright, 2012). For the multivariate regression problem with correlated errors, the problem can be solved by ℓ_1-penalization methods. In the case of fully observed data, the most prominent computational approach, popularly known as MRCE, to jointly estimate the precision matrix and regression coefficients is developed by Rothman et al. (2010). There are multiple approaches to solving the joint estimation problem in the fully observed data case and theoretical guarantees of many of these approaches had been studied in literature which we have discussed in Chapter 2.

We looked at a multivariate regression problem when there is missing data in the responses only. We assume that the responses are correlated and therefore an estimate of the precision matrix is of interest along with the regression coefficients. The problem with missing data may lead the empirical covariance matrix of the error to not be positive semi-definite and consequently the objective function may become unbounded from below and non-convex. We tackle the problem of non-convex objective function by converting it to a convex problem at a stage. To do so, first we replace the empirical estimate with an unbiased surrogate estimate that takes into account of the proportion of missing data, and then we project it onto the nearest positive semi-definite cone proposed by Datta and Zou (2017). As a result, the overall problem becomes convex and enjoys many nice properties of a convex optimization problem.

Although, not a missing data in responses scenario, but a similar type of solution to a different problem has been studied by Zhao and Genest (2019) which motivated us to adapt their approach to find the solution of our problem. They looked at the estimation of the joint dependence between all the observed variables (responses and covariates) characterized

by an elliptical copula and used non-parametric estimators of the input matrix for the covariates. As a result of the underlying structure, their estimated covariance for the covariates were not positive semi-definite and therefore the objective function became non-convex. They used the projection method proposed by Datta and Zou (2017) to convert the optimization problem to be convex for further analysis. They also established the theoretical properties of their estimators. Our method shares many similarities with the work by Zhao and Genest (2019), especially in terms of the steps of solving the problems, but also has many dissimilarities in terms of plug-in estimates, events of interest discussed in Section 4.3 and theoretical bounds. In our work, we assume the responses, covariates and the errors to have sub-Gaussian distributions.

Specifically, we assumed that there is no correlation among the response variables in the first stage of the estimation and performed a column-by-column Lasso estimation for each response variable and the fully observed covariates. No projection was necessary at this stage since there is no missing data in the covariates. However, we could not use any standard solver such as the `R` package `glmnet` directly that treats the responses as fully observed. Instead, we performed a projected gradient descent algorithm and provided the unbiased surrogate estimate (Loh and Wainwright, 2012) of the covariance between the observed covariates and each column of the corrupted response as an input. This stage provided us with a preliminary estimate of the regression coefficients. We proved the recovery rate of the first stage estimation of regression coefficients in Section 4.3.1. We assumed that the true covariance matrix of the covariates satisfy the restricted eigenvalue condition as required for classical Lasso estimation. To our knowledge, the proofs of Lemmas 13 and 14 are original contributions to knowledge which we required to prove Proposition 1. Proposition 1 is adapted for our problem from the Proposition 3.1 of Zhao and Genest (2019) and serves as an original contribution to the knowledge in this field.

In the next step, we used the CoGlasso algorithm established in Chapter 3 to estimate the precision matrix of the error of the model. The empirical covariance matrix of the error

is a function of the empirical covariance matrix of the responses, empirical covariance matrix of the covariates and the initial estimates of the regression coefficients estimated in the first stage. Since the responses are corrupted by missing data, we replaced the empirical covariance estimate with an unbiased surrogate estimate as proposed in Loh and Wainwright (2012), we used the estimated regression coefficients as a plug-in estimate and used the empirical covariance estimate for the covariates directly since those were fully observed. The estimated covariance matrix of the error may not be positive semi-definite since it is a function of a non-positive semi-definite matrix due to missingness, therefore, we project it to the nearest positive semi-definite cone to get a positive semi-definite input for the CoGlasso objective function. This step provided us with an estimate of a precision matrix that takes into account of the overall correlation structure of the error. We provided the recovery rate for the precision matrix estimator in Section 4.3.2 in elementwise maximum and operator norm. Following the theoretical development of CoGlasso estimator, we made the mutual incoherence assumption, which is also known as irrepresentability condition and proved Lemmas 15 and 16 to prove Proposition 2. Similar to the first stage, the proof of Proposition 2 is similar to the Proposition 4.2 of Zhao and Genest (2019) in terms of steps taken, but required different events of interest and resulted in different probabilistic bounds.

In the final step, we estimated a refined and final version of the regression coefficients by using the precision matrix estimate as a plug-in estimate from the previous stage by solving a ℓ_1-penalized regression problem. This stage only required the empirical covariance matrix of the covariates and the surrogate estimate for the covariance between the covariates and the responses as inputs. We performed an iterative soft-thresholding algorithm (ISTA) to get the final estimates of the regression coefficients. The recovery rates of the final estimate of regression coefficients are established in Section 4.3.3 in terms of Frobenius and ℓ_1-norm. We required that the loss function satisfies the restricted eigenvalue condition. We proved Theorem 3 using Proposition 3 which is similar in

flavour to Theorem 5.2 in Zhao and Genest (2019), but different in terms of the underlying events of interest and probabilistic bounds. To our knowledge, in the case of missing data in the responses for multivariate regression problems, our method is an original contribution to the existing field of work. We provide a step by step method to estimate the regression coefficients and precision matrix jointly that is easy to implement and comes with theoretical guarantees.

We also performed modest simulations to demonstrate the three-step method. We have not compared our methods with the state-of-the-art methods in all stages. The results shown only vary in the second step for the precision matrix matrix estimation where our method was compared with the method proposed by Fan et al. (2019) using the ADMM algorithm. We also have not studied different covariance structures for the precision matrix, for example, by varying the sparsity of the graph structures. These are some of the areas that can be explored in future.

To close the discussion, we want to emphasize that our contribution in this work is essentially by filling in the theoretical gap that existed in these types of projection based methods to estimate precision matrix in a noisy setting. Specifically, in Chapter 3, we have extended the theory on precision matrix estimation in the presence of two types of measurement errors. Our theory is established on top of the results proposed by Ravikumar et al. (2011) for the fully observed data, but it is significantly different in a sense that we have contributed to bridge the gap between estimation of undirected graphical model in the presence of measurement error for two different tail conditions imposed on the distribution of the data. We have used the idea of projection proposed by Datta and Zou (2017), originally shown in the case of penalized Lasso regression in the presence of measurement errors in the data. We have extended the problem for the graphical Lasso problem for corrupted data. We have also borrowed inspiration from the method proposed by Loh and Wainwright (2012), who perform a non-convex analysis in a nodewise-regression setup for the graphical Lasso in a measurement error model setup.

We contrast with this work in a sense that we have proposed a convex solution to the problem using the graphical Lasso objective function using a projection based method.

In Chapter 4, our contribution is novel and significant in a sense that to our knowledge, all the existing conditional graphical Lasso type problems are solved for full observed data. Therefore, we propose a convex solution to the joint estimation of the regression coefficients and the precision matrix in the presence of missing data, both in terms of proposing an algorithm as well as providing the theoretical guarantee for the estimation. We certainly borrowed inspiration and tools from Zhao and Genest (2019) for the theoretical derivations, but the problem solved in this thesis is completely different from the Zhao and Genest (2019) paper, therefore establishing the significance of our work.

Chapter 6

Conclusion and Future Work

The broad objective of this study was to develop and explore techniques to analyze high-dimensional data in a graphical model setting in the presence of corrupted data. We have proposed methods to solve problems of graph discovery (undirected) and also in a multivariate regression setting where the responses could be corrupted, thereby estimating the regression coefficients and the precision matrix jointly. We provided results to show theoretical guarantees for the two research problems that we studied in the previous two chapters along with some practical applications using synthetic data.

For Chapter 3, in terms of future work, both theoretical and computational aspects can be studied for Ising model type graph structures. Another direction to explore theoretically would be other types of penalties that are non-convex, such as Smoothly Clipped Absolute Deviation (SCAD) or minimax concave penalty (MCP). Fan et al. (2019) showed some of the computational aspects of such comparison but did not provide a rigorous theoretical work. As we have seen from the simulation studies, that the criterion for tuning parameter selection can certainly be improved for different types of measurement error, specifically for missing data scenarios, since only cross-validation seemed to have performed well. Theoretical guarantees can be studied by specifically imposing restrictions in the number of true edges, s, that is, by changing the sparsity pattern of the graphs.

In Chapter 4, both theoretical and simulation studies can be performed by imposing measurement error into both the covariates and the responses. Since we only explored the missing data scenario for the responses under sub-Gaussian assumptions, the work can be easily extended to additive noise setup as well as for polynomial-type tail conditions. In terms of the distribution of the data, other dependence structures among the covariates and/or response would be interesting to investigate in the presence of noisy data as an extension of Chapter 4. Another direction that can be explored is when the responses are time to event data and prone to censoring.

In this thesis, the general development of the estimation of the graph structure was formulated for undirected graphs. For directed acyclic graphs (DAGs), the problem of estimating the precision matrix has been studied in the literature from both frequentist and Bayesian point of view (Castelletti et al., 2018; Datta et al., 2019; Shojaie and Michailidis, 2010). Extending this case to incorporate the case when the data contain measurement error in a causal structural learning is worth investigating. Moreover, the joint estimation of the regression coefficients along with the precision matrix in a multivatiate regression setup when the responses and/or the covariates have an underlying causal structure is an intriguing and open problem. Another possible extension of the joint estimation in the multivariate regression problem under a directed acyclic graphs setting could be in the presence of measurement errors.

Finally, all the proposed methods can be applied in real life data in different domains where the data may have measurement errors of the two types that we discussed in this thesis. We performed all the simulation in `R 4.2.1` using multiple servers of Compute Canada clusters. A natural next step is to create `R` packages for efficient implementation of the two algorithms proposed in the two main chapters of this thesis.

Bibliography

Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika* **60,** 255–265.

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *The Journal of Machine Learning Research* **9,** 485–516.

Banerjee, S. and Ghosal, S. (2014). Posterior convergence rates for estimating large precision matrices using graphical models. *Electronic Journal of Statistics* **8,** 2111–2137.

Banerjee, S. and Ghosal, S. (2015). Bayesian structure learning in graphical models. *Journal of Multivariate Analysis* **136,** 147–162.

Benjamini, Y. and Speed, T. (2011). Estimation and correction for GC-content bias in high throughput sequencing. *Nucleic Acids Research* **40,** e72.

Berry, M. P., Graham, C. M., McNab, F. W., Xu, Z., Bloch, S. A., Oni, T., Wilkinson, K. A., Banchereau, R., Skinner, J., Wilkinson, R. J., et al. (2010). An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature* **466,** 973–977.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine learning* **3,** 1–122.

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics* **7,** 200–217.

Byrd, M., Nghiem, L. H., and McGee, M. (2021). Bayesian regularization of Gaussian graphical models with measurement error. *Computational Statistics & Data Analysis* **156,** 107085.

Cai, T., Liu, W., and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106,** 594–607.

Cai, T. T., Li, H., Liu, W., and Xie, J. (2013). Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika* **100,** 139–156.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC.

Castelletti, F., Consonni, G., Della Vedova, M. L., and Peluso, S. (2018). Learning Markov equivalence classes of directed acyclic graphs: an objective Bayes approach. *Bayesian Analysis* **13,** 1235–1260.

Censor, Y., Zenios, S. A., et al. (1997). *Parallel optimization: Theory, algorithms, and applications*. Oxford University Press on Demand.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95,** 759–771.

Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische Mathematik* **31,** 377–403.

Datta, A., Banerjee, S., Hodges, J. S., and Gao, L. (2019). Spatial disease mapping using directed acyclic graph auto-regressive (DAGAR) models. *Bayesian Analysis* **14,** 1221.

Datta, A. and Zou, H. (2017). CoCoLasso for high-dimensional error-in-variables regression. *The Annals of Statistics* **45,** 2400–2426.

Fan, J., Feng, Y., and Wu, Y. (2009). Network exploration via the adaptive Lasso and SCAD penalties. *The Annals of Applied Statistics* **3,** 521.

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

Fan, R., Jang, B., Sun, Y., and Zhou, S. (2019). Precision matrix estimation with noisy and missing data. *Proceedings of Machine Learning Research* .

Foygel, R. and Drton, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. In *Advances in Neural Information Processing Systems 22 (NIPS 2010)*, volume 23, pages 2020–2028.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics* **9,** 432–441.

Gao, X., Pu, D. Q., Wu, Y., and Xu, H. (2012). Tuning parameter selection for penalized likelihood estimation of Gaussian graphical model. *Statistica Sinica* **22,** 1123–1146.

Götze, F., Sambale, H., and Sinulis, A. (2021). Concentration inequalities for polynomials in α-sub-exponential random variables. *Electronic Journal of Probability* **26,** 1–22.

Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge University Press.

Hsieh, C.-J., Sustik, M. A., Dhillon, I. S., Ravikumar, P., et al. (2014). QUIC: quadratic approximation for sparse inverse covariance estimation. *The Journal of Machine Learning Research* **15,** 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,** 680–688.

Iturria, S. J., Carroll, R. J., and Firth, D. (1999). Polynomial regression and estimating functions in the presence of multiplicative measurement error. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61,** 547–561.

Johnson, C., Jalali, A., and Ravikumar, P. (2012). High-dimensional sparse inverse covariance estimation using greedy methods. In *Artificial Intelligence and Statistics*, pages 574–582. PMLR.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics* **29,** 295–327.

Johnstone, I. M. and Lu, A. Y. (2009). Sparse principal components analysis. *arXiv preprint arXiv:0901.4392* .

Khondker, Z. S., Zhu, H., Chu, H., Lin, W., and Ibrahim, J. G. (2013). The Bayesian covariance Lasso. *Statistics and its Interface* **6,** 243.

Kolar, M. and Xing, E. P. (2012). Estimating sparse precision matrices from data with missing values. In *International Conference on Machine Learning*, pages 635–642.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics* **37,** 4254–4278.

Lauritzen, S. L. (1996). *Graphical models*, volume 17. Clarendon Press.

Lenkoski, A. and Dobra, A. (2011). Computational aspects related to inference in Gaussian graphical models with the G-Wishart prior. *Journal of Computational and Graphical Statistics* **20,** 140–157.

Lian, H. (2011). Shrinkage tuning parameter selection in precision matrices estimation. *Journal of Statistical Planning and Inference* **141,** 2839–2848.

Liang, F., Jia, B., Xue, J., Li, Q., and Luo, Y. (2018). An imputation–regularized optimization algorithm for high-dimensional missing data problems and beyond. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80,** 899–926.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Liu, H., Roeder, K., and Wasserman, L. (2010). Stability approach to regularization selection (StARS) for high dimensional graphical models. *Advances in Neural Information Processing Systems* **23,**.

Loh, P.-L. and Tan, X. L. (2018). High-dimensional robust precision matrix estimation: Cellwise corruption under ϵ-contamination. *Electronic Journal of Statistics* **12,** 1429–1467.

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

Loh, P.-L. and Wainwright, M. J. (2013). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Advances in Neural Information Processing Systems* **26,**.

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

Lounici, K. (2014). High-dimensional covariance matrix estimation with missing observations. *Bernoulli* **20,** 1029–1058.

Marlin, B., Schmidt, M., and Murphy, K. (2012). Group sparse priors for covariance estimation. *arXiv preprint arXiv:1205.2626* .

Marlin, B. M. and Murphy, K. P. (2009). Sparse Gaussian graphical models with unknown block structure. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 705–712.

Mazumder, R. and Hastie, T. (2012). Exact covariance thresholding into connected components for large-scale graphical Lasso. *The Journal of Machine Learning Research* **13,** 781–794.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics* **34,** 1436–1462.

Mohammadi, A. and Wit, E. C. (2015). Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Analysis* **10,** 109–138.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science* **27,** 538–557.

Öllerer, V. and Croux, C. (2015). Robust high-dimensional precision matrix estimation. In *Modern nonparametric, robust and multivariate methods*, pages 325–350. Springer.

Ortega, J. M. and Rheinboldt, W. C. (2000). *Iterative solution of nonlinear equations in several variables*. Society for Industrial and Applied Mathematics.

Park, S., Shedden, K., and Zhou, S. (2017). Non-separable covariance models for spatio-temporal data, with applications to neural encoding analysis. *arXiv preprint arXiv:1705.05265* .

Park, S., Wang, X., and Lim, J. (2021). Estimating high-dimensional covariance and precision matrices under general missing dependence. *Electronic Journal of Statistics* **15,** 4868–4915.

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association* **104,** 735–746.

Pourahmadi, M. (2011). Covariance estimation: The GLM and regularization perspectives. *Statistical Science* **26,** 369–387.

Purdom, E. and Holmes, S. P. (2005). Error distribution for gene expression data. *Statistical Applications in Genetics and Molecular Biology* **4,**.

Ravikumar, P., Wainwright, M. J., and Lafferty, J. D. (2010). High-dimensional ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics* **38,** 1287–1319.

Ravikumar, P., Wainwright, M. J., Raskutti, G., Yu, B., et al. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics* **5,** 935–980.

Rocha, G. V., Zhao, P., and Yu, B. (2008). A path following algorithm for sparse pseudo-likelihood inverse covariance estimation (SPLICE). *arXiv preprint arXiv:0807.3734* .

Rosenbaum, M. and Tsybakov, A. B. (2010). Sparse recovery under matrix uncertainty. *The Annals of Statistics* **38,** 2620–2651.

Rosenthal, H. P. (1970). On the subspaces of L^p $(p > 2)$ spanned by sequences of independent random variables. *Israel Journal of Mathematics* **8,** 273–303.

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* **2,** 494–515.

Rothman, A. J., Levina, E., and Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics* **19,** 947–962.

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

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC Press.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6,** 461–464.

Shi, W., Ghosal, S., and Martin, R. (2021). Bayesian estimation of sparse precision matrices in the presence of Gaussian measurement error. *Electronic Journal of Statistics* **15,** 4545–4579.

Shojaie, A. and Michailidis, G. (2010). Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika* **97,** 519–538.

Singhania, A., Verma, R., Graham, C. M., Lee, J., Tran, T., Richardson, M., Lecine, P., Leissner, P., Berry, M. P., Wilkinson, R. J., et al. (2018). A modular transcriptional signature identifies phenotypic heterogeneity of human tuberculosis infection. *Nature Communications* **9,** 1–17.

Slijepcevic, S., Megerian, S., and Potkonjak, M. (2002). Location errors in wireless embedded sensor networks: sources, models, and effects on applications. *ACM SIGMOBILE Mobile Computing and Communications Review* **6,** 67–78.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58,** 267–288.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027* .

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory* **55,** 2183–2202.

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

Wang, H. (2012). Bayesian graphical Lasso models and efficient posterior computation. *Bayesian Analysis* **7,** 867–886.

Wang, J. (2015). Joint estimation of sparse multivariate regression and conditional graphical models. *Statistica Sinica* **25,** 831–851.

Witten, D. M., Friedman, J. H., and Simon, N. (2011). New insights and faster computations for the graphical Lasso. *Journal of Computational and Graphical Statistics* **20,** 892–900.

Xu, Q. and You, J. (2007). Covariate selection for linear errors-in-variables regression models. *Communications in Statistics - Theory and Methods* **36,** 375–386.

Yin, J. and Li, H. (2011). A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *The Annals of Applied Statistics* **5,** 2630.

Yin, J. and Li, H. (2013). Adjusting for high-dimensional covariates in sparse precision matrix estimation by ℓ_1-penalization. *Journal of Multivariate Analysis* **116,** 365–381.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research* **11,** 2261–2286.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94,** 19–35.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38,** 894–942.

Zhang, T. and Zou, H. (2012). Sparse precision matrix estimation via lasso penalized d-trace loss. *Biometrika* **99,** 1–18.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *The Journal of Machine Learning Research* **7,** 2541–2563.

Zhao, Y. and Genest, C. (2019). Inference for elliptical copula multivariate response regression models. *Electronic Journal of Statistics* **13,** 911–984.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* **101,** 1418–1429.