Structure Learning in Graphical Models from Indirect Observations

Anonymous authors
Paper under double-blind review

Abstract

This paper considers learning of the graphical structure of a p-dimensional random vector $\mathbf{X} \in \mathbb{R}^p$ using both parametric and non-parametric methods. Unlike the previous works which observe $\boldsymbol{x}$ directly, we consider the indirect observation scenario in which samples $\boldsymbol{y}$ are collected via a sensing matrix $\mathbf{A} \in \mathbb{R}^{d\times p}$, and corrupted with some additive noise $\mathbf{w}$, i.e, $\mathbf{Y} = \mathbf{AX} + \mathbf{W}$. For the parametric method, we assume $\mathbf{X}$ to be Gaussian, i.e., $\boldsymbol{x} \in \mathbb{R}^p \sim \mathcal{N}(\mu, \boldsymbol{\Sigma})$, $\mu \in \mathbb{R}^p$, and $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$. For the first time, we show that the correct graphical structure can be correctly recovered under the indefinite sensing system ($d < p$) using insufficient samples ($n < p$). In particular, we show that for the exact recovery, we require dimension $d = \Omega\left(p^{0.8}\right)$ and sample number $n = \Omega\left(p^{0.8}\log^3 p\right)$. For the nonparametric method, we assume a nonparanormal distribution for $\mathbf{X}$ rather than Gaussian. Under mild conditions, we show that our graph-structure estimator can obtain the correct structure. We derive the minimum sample number n and dimension d as $n \gtrsim (\text{deg})^4 \log^4 n$ and $d \gtrsim p + (\text{deg} \cdot \log(d-p))^{\beta/4}$, respectively, where deg is the maximum Markov blanket in the graphical model and $\beta > 0$ is some fixed positive constant. Additionally, we obtain a non-asymptotic uniform bound on the estimation error of the CDF of $\mathbf{X}$ from indirect observations with inexact knowledge of the noise distribution. To the best of our knowledge, this bound is derived for the first time and may serve as an independent interest. Numerical experiments on both real-world and synthetic data are provided confirm the theoretical results.

1 Introduction

Graphical models provide a general framework of representing the dependency relations among random variables. They have a broad spectrum of applications in biology, natural language processing and computer vision Koller & Friedman (2009); Friedman (2004), etc. For an arbitrary random vector $\mathbf{X} \in \mathbb{R}^p$, we can construct a graphical model $G = (V, E)$ by associating each entry X_i with a node $v_i \in V$ and adding an edge $e = (v_i, v_j)$ to the edge set E if X_i and X_j are conditionally dependent given other random variables, where X_i and X_j denote the ith and jth entry of $\mathbf{X}$, respectively.

Discovery of the graph structure from a collection of direct observations of $\mathbf{X}$ has been studied in the past Ravikumar et al. (2011); Cai et al. (2011); Liu et al. (2009; 2012); Zhao et al. (2014); Xu & Gu (2016); Fan et al. (2017). However, direct observations of the desired signal is not always possible. Instead, the signal has to be measured indirectly. Such a measurement model arises frequently in many practical applications, such as biomedical sensing Müller et al. (2008) where direct measurements of some desired molecules such as miRNAs are too expensive. Further, one commonly encountered problem is that the observations are contaminated with measurement noise, which leads to inaccurate estimation of the graphical structure. Inspired by these challenges, we consider the graph structure recovery under an indirect linear measurement scenario from the desired signal as

$$\mathbf{Y}^{(s)} = \mathbf{AX}^{(s)} + \mathbf{W}^{(s)}, \quad 1 \le s \le n, \tag{1}$$

where $\mathbf{Y}^{(s)}$ denotes the sth measurement, $\mathbf{A} \in \mathbb{R}^{d\times p}$ denotes the sensing matrix, and $\mathbf{W} \in \mathbb{R}^d$ denotes the sensing noise. Our goal is to learn the pair-wise independence relation (structure) of an undirected

graphical model, i.e., Markov random fields, from the observations $\{\mathbf{Y}^{(s)}\}_{1\leq s\leq n}$ using both parametric and non-parametric approaches.

In the parametric scenario, random vector $\mathbf{X} \in \mathbb{R}^p$ satisfies the Gaussian distribution, i.e., $\mathbf{X} \sim \mathsf{N}(\mathbf{0}, \mathbf{\Sigma})$, and hence a *Gaussian Graphical Model* (GGM). While for the nonparametric method, we assume random vector $\mathbf{X}$ follows the nonparanormal distribution, i.e., the joint distribution $\boldsymbol{g}(\mathbf{X})$ exhibits a Gaussian distribution $\mathsf{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$, i.e., $\mathbf{X}$ behaves as multivariate Gaussian after some transformation $\boldsymbol{g}(\mathbf{X}) = [g_i(X_i) \ \cdots \ g_p(X_p)]^\top$. Let $\mathbf{\Theta}$ be the inverse of the covariance matrix $\mathbf{\Sigma}$. It is known that conditional independency relation of $\mathbf{X}$ is completely incorporated into matrix $\mathbf{\Theta}$. In other words, we have $\Theta_{i,j} = 0$ iff X_i and X_j are independent given the rest of entries $X_{\setminus\{i,j\}}$. Hence our goal of learning the graphical structure reduces to detecting the support set S of $\mathbf{\Theta}$, where $S \triangleq \{(i,j) \mid \Theta_{i,j} \neq 0, \ \forall i \neq j\}$. This property of precision matrix lays the foundation for many algorithms that learn the graphical structure.

1.1 Related Works.

To the best of our knowledge, our work is the first to consider learning the undirected graphical model structure of a high dimensional signal $\mathbf{X}$ from possibly low dimensional indirect observations of $\mathbf{X}$. In the following, we separately review the related work regarding the parametric and non-parametric scenarios.

Parametric Scenario. Learning GGM when the data samples are *directly* observed has been extensively studied in the literature, e.g., Ravikumar et al. (2011); Cai et al. (2011). Both works assumed an sparse precision matrix $\mathbf{\Theta}$ and propose to recover its support set with ℓ_1-relaxation. In Ravikumar et al. (2011), the authors proposed gLasso, which can be regarded as an M-estimator Wainwright (2019) (Chap. 9). The basic assumption is that the ground-truth signal lies within a certain low-dimensional space. By minimizing the negative log-likelihood plus some specific regularizers, one can reconstruct the signal from insufficient samples. Different from Ravikumar et al. (2011), Cai et al. (2011) proposed CLIME, a constrained convex optimization framework to estimate sparse precision matrix $\mathbf{\Theta}$. Instead of maximizing the likelihood, the authors first obtain an empirical covariance matrix $\widehat{\mathbf{\Sigma}}_n$ and then force $\mathbf{\Theta}$ to approximate its inverse. Other variations of GGM-learning such as GGM with hidden variables Chandrasekaran et al. (2010); Mazumder & Hastie (2012), and GGM with an unknown block structure Marlin & Murphy (2009) have been proposed when some prior information or constraints on the structure exist. For detailed discussions, please refer to Tibshirani et al. (2015); Wainwright (2019).

Non-parametric Scenario. The most related works on the nonparametric learning of graphical models are Liu et al. (2009; 2011; 2012); Zhao et al. (2014); Xu & Gu (2016); Fan et al. (2017).

The works by Liu et al. (2009; 2012); Xue et al. (2012); Zhao et al. (2014) considered a similar problem as ours, that is the discovery of the graphical structure with nonparanormal distributed random vector $\mathbf{X}$. However, they assumed direct observations of $\mathbf{X}$ and without any measurement noise. The differences across the above works lie in the estimation method of the covariance matrix. In Liu et al. (2009), the covariance matrix is estimated via the CDF estimator; while in Xue et al. (2012), it is estimated by the Spearman's rho estimator. In an independent work, Liu et al. (2012) pointed out the covariance matrix can be estimated by Kendall's tau estimator as well. Later, Zhao et al. (2014) proposed a projection based algorithm to accelerate the estimator in Liu et al. (2012). Apart from the graphical models following the nonparanormal distribution, other types of works include the graph with forest structure Liu et al. (2012), the graph with the elliptical distribution Xu & Gu (2016), and latent Gaussian copula model Fan et al. (2017), etc.

Another line of research is the density deconvolution which dates at least back to Zhang (1990); Fan (1991); Masry (1991), where kernel-based estimators are proposed for the noise with infinite support set. To improve the performance for the noise with finite support, a ridge-parameter method is proposed in Hall & Meister (2007). Similar work includes Dattner et al. (2011); Trong & Phuong (2019). The work by Youndjé & Wells (2008) generalized Masry (1991) and gave a data-driven method to select the optimal bandwidth. Apart from the kernel-based methods, Pensky & Vidakovic (1999) presented a projection-based method based on the Meyer-type wavelets, which adapts to the super-smooth noise. Notice that the above works all assume perfect knowledge of the noise distribution. Later, Dattner & Reiser (2013); Kappus & Mabon (2014) and Phuong (2020) studied the unknown noisy case but required repeated measurements to estimate the distribution of

the noise. In contrast, our setting assumes inexact knowledge of the noise and requires no extra step for the noise estimation. More importantly, our setting is focused on indirect measurements.

1.2 Contributions

To the best of our knowledge, our work is the first on estimating the graphical structure under an indirect measurement scenario. We also provide the theoretical analysis of our estimation.

- For the parametric scenario, we show the correct graphical structure can be recovered under the indefinite sensing system ($d < p$) using insufficient samples ($n < p$). Specifically, we require the sample number $n = \Omega\left(p^{0.8}\log^3 p\right)$ and the dimension $d = \Omega\left(p^{0.8}\right)$, which suggests with fewer samples than signal length p, we can still recover the graph structure of a high dimensional signal from its low dimensional observations ($d < p$).
- For the non-parametric scenario, we propose an estimator for the graphical structure together with the sufficient conditions for the correct recovery. We show that the sample number n must be at least $n \gg (\deg)^4 \log^4 n$, where deg denotes the maximum Markov blanket in the graph. Further, we obtain a lower bound on the dimension as $d \gg p + (\deg)^{\beta/4}\log^{\beta/4}(d-p)$, where $\beta > 0$ is some fixed positive constant to be defined.
- Additionally, our work is the first to consider the deconvolution estimator for the CDF with limited knowledge of the noise distribution. Compared with the previous work Dattner & Reiser (2013); Kappus & Mabon (2014); Phuong (2020), our work does not assume perfect knowledge of noise distribution, and our estimator requires no additional steps to estimate the noise distribution. Moreover, we give a non-asymptotic uniform bound on the estimation error, that is, $\sup_x \left|\widehat{F}(x) - F(x)\right| \lesssim \frac{\log^2(np)}{(d-p)^{\beta/2}} + \frac{1}{\sqrt{n}}$ (cf. Thm. 4), where $\widehat{F}(\cdot)$ and $F(\cdot)$ denote the estimated marginal CDF and ground-truth marginal CDF, respectively. These analytical results might be of interest to some other areas of statistics.

2 Problem Formulation

We start with a formal restatement of the sensing relation reading as

$$\mathbf{Y}^{(s)} = \mathbf{A}\mathbf{X}^{(s)} + \mathbf{W}^{(s)}, \quad 1 \le s \le n, \tag{2}$$

where $\mathbf{Y}^{(s)} \in \mathbb{R}^d$ denotes the sth reading, $\mathbf{A} \in \mathbb{R}^{d\times p}$ is the sensing matrix with each entry A_{ij} being a standard normal RV, i.e., $A_{ij} \sim \mathsf{N}(0,1)$, and $\mathbf{W}^{(s)}$ denotes the measurement noise with each entry $\mathbf{W}_i$ being a Gaussian RV with zero mean and variance σ^2, i.e., $\mathbf{W}_i \sim \mathsf{N}(0,\sigma^2)$.

We assume that most entries of $\mathbf{X}$ are pair-wise conditionally independent, which is widely used in previous works such as Ravikumar et al. (2011); Cai et al. (2011); Liu et al. (2009). Our goal is to uncover the undirected graphical structure (or pair-wise independence) of $\mathbf{X}$ from the samples $\left\{\mathbf{Y}^{(s)}\right\}_{1\le s\le n}$. Before proceeding, we first introduce the notations.

Notations. We denote $c, c_0, c_i > 0$ as arbitrary fixed positive constants. Notice that the specific values may not be necessarily identical even if they share the same name. For arbitrary real numbers a and b, we denote $a \lesssim b$ if there exists some $c_0 > 0$ such that $a \le c_0 b$. Similarly we define $a \gtrsim b$. We write $a \asymp b$ when $a \lesssim b$ and $a \gtrsim b$ hold simultaneously. The maximum of a and b is denoted as $a \vee b$; while the minimum is denoted as $a \wedge b$.

For an arbitrary matrix $\mathbf{M} \in \mathbb{R}^{m_1\times m_2}$, we denote $\mathbf{M}_i$ as its i-th column and M_{ij} as the (i,j)-th entry. Its Frobenius norm $\|\mathbf{M}\|_{\mathrm{F}}$ is defined as $\sqrt{\sum_{i,j} M_{i,j}^2}$ and the operator norm $\|\mathbf{M}\|_{\mathrm{OP}}$ is defined as $\max_{\|\boldsymbol{u}\|_2=1}\|\mathbf{M}\boldsymbol{u}\|_2$. Furthermore, we define $\|\mathbf{M}\|_{a,b}$ as $\max_{\|\boldsymbol{z}\|_a=1}\|\mathbf{M}\boldsymbol{z}\|_b$. Two special cases are $\|\mathbf{M}\|_{1,1}$ and $\|\mathbf{M}\|_{\infty,\infty}$, which can be written as $\|\mathbf{M}\|_{1,1} \triangleq \max_j \sum_i |M_{i,j}|$, and $\|\mathbf{M}\|_{\infty,\infty} \triangleq \max_i \sum_j |M_{i,j}|$, respectively. Moreover, we define $\|\mathbf{M}\|_{1,\mathrm{off}} \triangleq \sum_{i\neq j}|M_{i,j}|$, $\|\mathbf{M}\|_{\mathrm{off,F}} \triangleq \sqrt{\sum_{i\neq j} M_{i,j}^2}$, and $\|\mathbf{M}\|_{\infty} \triangleq \max_{i,j}|M_{i,j}|$, respectively.

Algorithm 1 Estimation of the Graphical Structure via the Parametric Method.

Input: Samples $\left\{\mathbf{Y}^{(i)}\right\}_{i=1}^{n} \in \mathbb{R}^p$ and sensing matrix $\mathbf{A} \in \mathbb{R}^{d\times p}$.

Stage I: Estimate the covariance matrix $\widehat{\mathbf{\Sigma}}_n^{\text{param}}$ as

$$\widehat{\mathbf{\Sigma}}_n^{\text{param}} = \mathbf{I} + \frac{1}{d+1}\left[\mathbf{A}^\top\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{Y}^{(i)}\mathbf{Y}^{(i)\top}\right)\mathbf{A}\right]_{\text{off}}, \tag{3}$$

where $[\cdot]_{\text{off}}$ denotes the operation of picking non-diagonal entries,

Stage II: Obtain $\widehat{\mathbf{\Theta}}^{\text{param}}$ as

$$\widehat{\mathbf{\Theta}}^{\text{param}} = \underset{\mathbf{\Theta}\succ\mathbf{0}}{\operatorname{argmin}} - \log\det(\mathbf{\Theta}) + \operatorname{Tr}(\widehat{\mathbf{\Sigma}}_n^{\text{param}}\mathbf{\Theta}) + \lambda_{\text{param}}\|\mathbf{\Theta}\|_{1,\text{off}}, \tag{4}$$

where $\lambda_{\text{param}} > 0$ is some positive constants.

Output: Estimated precision matrix $\widehat{\mathbf{\Theta}}_{\text{param}}$.

For the covariance matrix $\mathbf{\Sigma}$, we denote its inverse as $\mathbf{\Theta}$. The support set S is defined as $\{(i,j) \mid \mathbf{\Theta}_{i,j} \neq 0,\ \forall\ i \neq j\}$ and its complement is denoted as S^c. Moreover, we define the maximum Markov blanket in the graphical model as deg, or equivalently the maximum of the non-zero entries among columns of $\mathbf{\Theta}$, i.e., $\text{deg} \triangleq \max_i \|\mathbf{\Theta}_i\|_0$, where $\mathbf{\Theta}_i$ denotes the ith column of $\mathbf{\Theta}$. The parameter $\kappa_{\mathbf{\Sigma}}$ is defined as $\max_i \sum_j |\mathbf{\Sigma}_{i,j}|$. Furthermore, we define the Fisher information matrix of $\mathbf{\Sigma}$ as $\mathbf{\Gamma} \triangleq \mathbf{\Sigma}\otimes\mathbf{\Sigma}$, where $\otimes$ is the kronecker product Golub & Van Loan (2012). The parameter $\kappa_{\mathbf{\Gamma}}$ is defined as $\left\|\left(\mathbf{\Gamma}_{SS}\right)^{-1}\right\|_{\infty,\infty}$, where S is the support set and $\mathbf{\Gamma}_{SS}$ is the sub-matrix of $\mathbf{\Gamma}$ such that its rows and columns lie within the set S.

3 Parametric Method

For the parametric scenario, we assume random vector $\mathbf{X} \in \mathbb{R}^p$ follows the Gaussian distribution with zero mean and covariance $\mathbf{\Sigma}^\natural$, i.e., $\mathbf{X} \sim \mathsf{N}(\mathbf{0}, \mathbf{\Sigma}^\natural)$. We focus on the scenario where the sensing matrix $\mathbf{A} \in \mathbb{R}^{d\times p}$ is under-determined, namely, $d < p$. In other words, we wish to recover the graphical model of a high dimensional $\mathbf{X}$ from its low dimensional observations $\mathbf{Y}$. To avoid measurements being dominated by one or a small group of X_i's, we assume that $\text{Var}(X_i) = 1$, $1 \leq i \leq p$. Our proposed estimator to reconstruct the graphical structure is illustrated in Alg. 1. The analysis of the estimator is given in the following.

3.1 Properties of Graphical Structure Estimator: Parametric Case

We will show that a correct graphical structure can be obtained when the penalty coefficient λ_{param} in equation 4 is properly chosen. The core of the analysis lies in the following lemma.

Lemma 1. *The property $\|\widehat{\mathbf{\Sigma}}_n^{\text{param}} - \mathbf{\Sigma}^\natural\|_\infty \leq \tau_\infty$ holds with probability $1 - o(1)$, where $\widehat{\mathbf{\Sigma}}_n^{\text{param}}$ is defined in equation 3, and the threshold parameter τ_∞ is written as*

$$\tau_\infty \triangleq \frac{c_0\sqrt{d\log p}}{d+1}\max_i\left\|\mathbf{\Sigma}_i^\natural\right\|_2 + \frac{c_1\log p}{d+1}\left(1+\frac{c_2 p}{d}\right)\left\|\mathbf{\Sigma}^\natural\right\|_{\text{off,F}} + \frac{c_3\log p\sqrt{dp}}{\sqrt{n}(d+1)} + \frac{c_4 p(\log p)^{3/2}}{\sqrt{n}(d+1)}\left(1+\frac{c_2 p}{d}\right) + \frac{c_5 p\sqrt{\log p}}{\sqrt{d}(d+1)} + \frac{c_6\sigma^2\log p}{d}\left[1 + c_7\left(\sqrt{\frac{d}{n}}\vee\frac{d}{n}\right)\right]. \tag{5}$$

Having obtained the covariance matrix $\widehat{\mathbf{\Sigma}}_n^{\text{param}}$, we estimate the graphical structure of $\mathbf{X}$ by plugging $\widehat{\mathbf{\Sigma}}_n^{\text{param}}$ into the graphical lasso estimator Tibshirani et al. (2015) as in equation 4. To analyze the properties of the graphical structure estimator, we assume the *irrepresentable condition*, a widely-used assumption in the previous works (cf. Ravikumar et al. (2011); Liu et al. (2009)), which control the influence of the non-edge terms, i.e., $\mathbf{\Gamma}_{S^cS}$, on the edge terms, i.e., $\mathbf{\Gamma}_{SS}$ (see Section 2 for definitions of these matrices).

Assumption 1 (irrepresentable condition)**.** *Let the support set S and matrix $\mathbf{\Gamma}$ be defined in Section 2. Then, there exists a positive constant $\theta \in (0, 1]$ such that $\left\|\!\left|\mathbf{\Gamma}_{S^cS}(\mathbf{\Gamma}_{SS})^{-1}\right|\!\right\|_{1,1} \leq 1 - \theta$.*

Adopting a similar proof method as in Ravikumar et al. (2011), we can obtain the conditions (cf. Thm. 1) for the correct recovery of graphical structure under Assumption 1. For the conciseness of presentation, we refer the details to Ravikumar et al. (2011) and omit them in this work.

Theorem 1. *Let $\tau_\infty \lesssim \frac{\theta}{(\theta+8)(\deg)\cdot\kappa_{\mathbf{\Sigma}}^3\kappa_{\mathbf{\Gamma}}}$, and set $\lambda_{\text{param}} = 8\tau_\infty/\theta$. Then the property $(\widehat{\mathbf{\Theta}}^{\text{param}})_{i,j} = 0$ holds for all (i, j) outside the support set S, i.e., $(i, j) \in S^c$, with probability $1 - o(1)$. Furthermore, if $\min_{(i,j)\in S} |\mathbf{\Theta}^\natural_{i,j}| \geq 2\kappa_{\mathbf{\Gamma}}(1 + 8\theta^{-1})\tau_\infty$, then $\text{sign}(\widehat{\mathbf{\Theta}}^{\text{param}}) = \text{sign}(\mathbf{\Theta}^\natural)$ with probability $1 - o(1)$.*

This theorem proves that our proposed Alg. 1 can detect all conditionally independent pairs (the graph edges) provided that the non-zero elements $\mathbf{\Theta}^\natural_{i,j}$, $(i, j) \in S$, are not too small, i.e., the absolute value of non-zero elements in $\mathbf{\Theta}^\natural$ are above some fixed threshold.

3.2 Discussions

In the following, we provide more insights regarding: (i) the minimum sample size n and (ii) the minimum projection dimension d. An illustration of the infeasible region of n and d is plotted in Fig. 1, where the infeasible region is marked with color red.

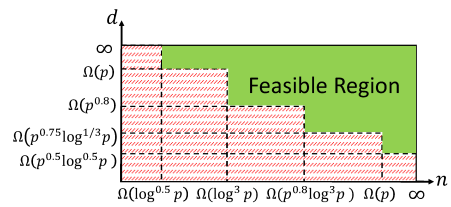

Figure 1: Infeasible region w.r.t number n and dimension d.

Minimum Sample Size n. In the high-dimensional setting, it is desirable to reduce the sample number n to less than the dimension of signal p, i.e., $n < p$.

Treating parameters θ, $\kappa_{\mathbf{\Sigma}}$, and $\kappa_{\mathbf{\Gamma}}$ as some positive constants, Thm. 1 requires that $\deg \cdot \tau_\infty$ to be some positive constant, from which we can obtain the minimum sample size $n \gtrsim \sqrt{\deg^2 \cdot \log p}$. For the case of indefinite sensing matrix $\mathbf{A} \in \mathbb{R}^{d\times p}$ $(d < p)$, we need to increase the sample number n to the order of $\Omega\left(\deg^2 \cdot (\log p)^3\right)$, which is obtained by setting $d = cp$, where $0 < c < 1$ is some positive constant. When d reduces to $\Omega(p^{0.8})$, we need to further inflate the sample number n to be at least $\deg^2(\log p)^3 p^{0.8}$, which is still less than the order p.

Minimum Projection Dimension d. We investigate the minimum projection dimension d and proves the possibility of exact recovery of graph with an indefinite sensing matrix $\mathbf{A}$.

Under the high-dimensional setting where we require $n \leq p$, we need d to be at least

$$d \gtrsim \sqrt{\deg \cdot p \log p \cdot \left\|\!\left|\mathbf{\Sigma}^\natural\right|\!\right\|_{\text{off,F}}} \vee \sigma^2 \deg \cdot \log p \vee p^{3/4} (\log p)^{1/3} \sqrt{\deg} \vee \deg^2 \cdot \log p \cdot \left(\max_i \left\|\mathbf{\Sigma}^\natural_i\right\|_2^2\right).$$

Using above, we can confirm that d is less than p given that $\left\|\!\left|\mathbf{\Sigma}^\natural\right|\!\right\|_{\text{off,F}} = o(p)$ and $\max_i \left\|\mathbf{\Sigma}^\natural_i\right\|_2 = o(\sqrt{p})$. If $\left\|\!\left|\mathbf{\Sigma}^\natural\right|\!\right\|_{\text{off,F}}$ is some fixed positive constant, we can reduce the dimension d to $\sqrt{\deg}\cdot p^{3/4}(\log p)^{1/3}$, approximately of the order $(\deg)^{1/2} p^{3/4}$.

4 Non-parametric Method

We begin the discussion with some background knowledge on the nonparanormal distribution.

4.1 Background

Definition 1 (Nonparanormal)**.** *We call a set of random variables $\mathbf{X} = [X_i \ \cdots \ X_p]^\top$ follows the* nonparanomral distribution*, namely, $\mathbf{X} \sim \text{NPN}(\boldsymbol{g}, \boldsymbol{\mu}, \mathbf{\Sigma})$, if there exists a set of functions $\{g_j\}_{1\leq j\leq p}$ such that $\mathbf{Z} = [g_1(X_1) \ \cdots \ g_p(X_p)]^\top$ has Gaussian distribution, i.e., $\mathbf{Z} \sim \mathsf{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$.*

Assume that $\mathbf{X}$ satisfies the nonparanormal assumption, i.e., $\mathbf{X} \sim \mathrm{NPN}(\boldsymbol{g}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $\boldsymbol{\Theta}$ be the inverse matrix of the covariance matrix $\boldsymbol{\Sigma}$.

Lemma 2 (Lemma 3 in Liu et al. (2009))**.** *Random variables X_i and X_j are pairwise conditionally independent, i.e., $X_i \perp X_j \left| X_{\setminus i,j} \right.$, iff $\boldsymbol{\Theta}_{ij} = 0$.*

Thus, the pair-wise independence relation across the entries of $\mathbf{X}$ is fully incorporated into the matrix $\boldsymbol{\Theta}$. Denote the marginal distribution function of the ith entry X_i as $F_i(\cdot)$ and define the function $h_i(x) = \Phi^{-1}(F_i(x))$, where $\Phi(\cdot)$ is the CDF of the standard normal RV, namely, $\Phi(\cdot) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(\cdot)} e^{-t^2/2} dt$. According to Liu et al. (2009), we can estimate function $g_i(\cdot)$ by estimating function $h_i(\cdot)$, from which we can conclude the bottleneck in estimating the graphical structure lies in the estimation of the CDF functions, i.e., $F_i(\cdot)$, $1 \leq i \leq p$. Different from the previous work Liu et al. (2009), our sensing relation assumes noisy indirect measurements instead of noiseless direct measurements, which bring extra difficulties.

4.2 Estimator Design Intuition

Compared with the parametric learning method, we need to have a reliable estimation of the transform function $\boldsymbol{g}(\cdot)$ in the presence of the measurement noise $\mathbf{W}$. To handle such an issue, we relax the constraint that sensing matrix $\mathbf{A}$ is under-determined, where extra measurements are conducted ($d > p$) to suppress the sensing noise.

Remark 2. *Consider the direct noisy measurement reading as $\widetilde{\mathbf{Y}}^{(s)} = \mathbf{X}^{(s)} + \mathbf{W}^{(s)}$. Previous works Fan (1991); Zhang (1990) suggest that the error of estimating the CDF X_i from $\{\widetilde{\mathbf{Y}}^{(s)}\}$ is at least $\inf_{\widehat{F}} \left| \widehat{F}_i(x) - F_i(x) \right| \geq c_0 (\log n)^{-c_1}$ for a fixed point x, which means significant number of samples are required.*

Remark 3. *In fact, whether a compressive sensing system, i.e., $d < p$, can be used to estimate the transform function $\boldsymbol{g}(\cdot)$ still remains an open-problem.*

The basic idea of our estimator is to first transform the sensing relation in equation 2 to the additive model such that $\widehat{\mathbf{X}}^{(s)} = \mathbf{X}^{(s)} + \widehat{\mathbf{W}}^{(s)}$, where $\widehat{\mathbf{W}}^{(s)}$ is defined as $\left(\mathbf{A}^\top \mathbf{A}\right)^{-1} \mathbf{A}^\top \mathbf{W}^{(s)}$. Then our task reduces to estimating the marginal CDF from the samples $\widehat{\mathbf{X}}^{(s)}$, which is contaminated by the noise $\widehat{\mathbf{W}}^{(s)}$. This problem can be broadly categorized as the *density deconvolution* problem in statistics. When comparing with the previous work, instead of relying on full knowledge, we only use some inexact knowledge of the distribution of $\widehat{\mathbf{W}}^{(s)}$. To put more specifically, we can compute its mean and the approximated value of its variance as

$$\mathbb{E}(\widehat{w}_i) = 0, \quad \mathrm{Var}(\widehat{w}_i) \approx \frac{\sigma^2}{d-p}, \tag{6}$$

where $\widehat{w}_i$ denotes the ith entry of the noise $\widehat{\mathbf{W}}$. Whereas the previous work such as Zhang (1990); Fan (1991); Masry (1991); Hall & Meister (2007) assumes perfect knowledge of the distributions of $\widehat{\mathbf{W}}$, and the work Dattner & Reiser (2013); Kappus & Mabon (2014); Phuong (2020) disregards the distribution but requires extra steps to estimate the variance of noise $\{\widehat{\mathbf{W}}^{(s)}\}_{1 \leq s \leq n}$. The computation procedure of equation 6 is deferred to the supplementary material and the details of the estimator is summarized in Alg. 2.

4.3 Properties of CDF Estimator

This subsection investigates the properties of the marginal CDF estimator in equation 8 where each entry X_i is within the region $[0, 1]$. We begin the discussion by presenting the assumptions.

Definition 2 (Density family $\mathcal{F}_{\alpha,L}$)**.** *The density family $\mathcal{F}_{\alpha,L}$ is defined as the set of all distributions whose density functions $f(\cdot)$ and characteristic functions $\phi(\cdot)$ possess the following properties:*

- *The functions $f(\cdot)$ satisfy $\int_{-\infty}^{\infty} x^2 f(x) dx < \infty$;*
- *The characteristic functions $\phi(\cdot)$ satisfy $\int_{-\infty}^{\infty} |\phi(t)|^2 (1 + t^2)^\alpha dt \leq L$, where α is a positive constant controlling the smoothness of the corresponding PDF.*

Algorithm 2 Estimation of the Graphical Structure via the Non-parametric Method.

Input: Samples $\{\mathbf{Y}^{(i)}\}_{i=1}^n \in \mathbb{R}^p$ and sensing matrix $\mathbf{A} \in \mathbb{R}^{d\times p}$.

Stage I. We reconstruct the values $\widehat{\mathbf{X}}^{(s)}$ via the *least-square* (LS) estimator as

$$\widehat{\mathbf{X}}^{(s)} = \text{argmin}_{\mathbf{X}} \left\| \mathbf{Y}^{(s)} - \mathbf{A}\mathbf{X} \right\|_2. \tag{7}$$

Stage II. We estimate the marginal distribution function $\widehat{F}_i(\cdot)$ for the ith entry from the samples $\{\widehat{X}_i^{(s)}\}_{1\le s\le n}$ as

$$\widehat{F}_i(x) = \frac{1}{2} - \frac{1}{n\pi}\sum_{s=1}^n \int_0^\infty \frac{\sin\left[t(\widehat{X}_i^{(s)} - x)\right]}{t} \times \frac{\exp\left(-\frac{\sigma^2 t^2}{2(d-p)}\right)}{\exp\left(-\frac{\sigma^2 t^2}{d-p}\right) \vee \gamma t^a} dt, \tag{8}$$

where $\gamma > 0$ and $a > 1$ are some fixed positive constants. Then we truncate the estimated CDF function as

$$\widehat{F}_i^{\text{tr}}(x) = \begin{cases} \delta_{n,d,p}, & \widehat{F}_i(x) \le \delta_{n,d,p}; \\ \widehat{F}_i(x), & \delta_{n,d,p} \le \widehat{F}_i(x) \le 1 - \delta_{n,d,p}; \\ 1 - \delta_{n,d,p}, & \widehat{F}_i(x) \ge 1 - \delta_{n,d,p}, \end{cases} \tag{9}$$

where $\delta_{n,d,p} > 0$ is some pre-determined parameter.

Stage III. First, we estimate the mean $\widehat{m}_i$ and the variance $\widehat{v}_i$ as

$$\widehat{m}_i = \frac{1}{n}\sum_{s=1}^n \widehat{X}_i^{(s)};$$

$$\widehat{v}_i = \sqrt{\frac{1}{n-1}\sum_{s=1}^n \left(\widehat{X}_i^{(s)} - \widehat{m}_i\right)^2 - \frac{n}{n-1}\frac{\sigma^2}{d-p}}.$$

Then we estimate the covariance matrix $\widehat{\mathbf{\Sigma}}_n^{\text{non-param}}$ as

$$\widehat{\mathbf{\Sigma}}_n^{\text{non-param}} = \frac{1}{n}\sum_{s=1}^n (\widehat{\boldsymbol{h}}(\widehat{X}^{(s)}) - \widehat{\boldsymbol{\mu}})(\widehat{\boldsymbol{h}}(\widehat{X}^{(s)}) - \widehat{\boldsymbol{\mu}})^\top, \tag{10}$$

where the ith entry of $\widehat{\boldsymbol{h}}$ is defined as $\widehat{h}_i(x) = \widehat{m}_i + \widehat{v}_i \Phi^{-1}(\widehat{F}_i^{\text{tr}}(x))$, and $\widehat{\boldsymbol{\mu}}$ is the estimated mean of $\widehat{\boldsymbol{h}}(\cdot)$, namely, $n^{-1}\sum_{s=1}^n \widehat{\boldsymbol{h}}(\mathbf{X}^{(s)})$.

Stage IV. We reconstruct the matrix $\widehat{\mathbf{\Theta}}$ as

$$\widehat{\mathbf{\Theta}}^{\text{non-param}} = \text{argmin}_{\mathbf{\Theta}} - \log\det(\mathbf{\Theta}) + \text{Tr}\left(\mathbf{\Theta}\widehat{\mathbf{\Sigma}}_n^{\text{non-param}}\right) + \lambda_{\text{non-param}} \|\mathbf{\Theta}\|_{1,\text{off}}, \tag{11}$$

where $\lambda_{\text{non-param}} > 0$ is some positive constant for the regularizer $\|\mathbf{\Theta}\|_1$.

Output: Estimated matrix $\widehat{\mathbf{\Theta}}^{\text{non-param}}$.

Notice that these assumptions may have been widely used in previous work Trong & Phuong (2019); Phuong (2020). Many usual distributions belong to this density family, e.g., Cauchy distribution, Gaussian distribution, etc. In addition, we need the following two assumptions.

Assumption 2. *For an arbitrary entry X_i, we assume its distribution belongs to the density family $\mathcal{F}_{\alpha,L}$ such that $\alpha > -\frac{1}{2}$.*

Assumption 3. *For an arbitrary entry X_i, we assume its density function $f_i(\cdot)$ is bounded by some constant L_f, i.e., $|f_i(\cdot)| \le L_f$.*

Setting the parameter a as some fixed positive constant and using $\gamma \asymp \log(np)\left(\frac{\sigma^2}{d-p}\right)^{a/4}$, we can prove that the estimation error of the CDF estimator converges to zero with high probability, which is formally stated in the following theorem.

Theorem 4. *Under the Assumptions 2 and 3, the estimator error can be bounded as*

$$\sup_{x\in[0,1]}\left|F_i(x)-\widehat{F}_i(x)\right| \leq (\log n)\varepsilon_x + \ c_1\sqrt{\varepsilon_x}+\frac{c_2}{\sqrt{n}}, \tag{12}$$

for $1 \leq i \leq p$ with probability exceeding $1-o(1)$ when setting $\gamma \asymp \log(np)\left(\frac{\sigma^2}{d-p}\right)^{\frac{a}{4}}$. The parameter ε_x is defined as

$$\varepsilon_x \triangleq \frac{\log^{2/a}(np)\sigma}{\sqrt{d-p}}+\frac{\log^2(np)}{(d-p)^{\frac{a}{4}}}+\left(\frac{\sigma^2}{d-p}\right)^{\frac{2\alpha+1}{4}}+\frac{1}{n}, \tag{13}$$

where $a>1$ is some pre-determined positive constant.

To the best of our knowledge, this is the first non-asymptotic uniform bound on the estimation error of a CDF deconvolution estimator with inexact knowledge of the noise variance.

4.4 Properties of the Graphical Structure Estimator: Non-Parametric Case

The technical challenges of this analysis can be divided as two parts: (i) choosing the appropriate truncation parameter $\delta_{n,d,p}$ in equation 9; and (ii) estimating the covariance matrix with the noisy samples $\left\{\mathbf{Y}^{(s)}\right\}_{1\leq s\leq n}$.

Denote the oracle empirical covariance matrix $\mathbf{\Sigma}_n^{\text{non-param}}$ as

$$\mathbf{\Sigma}_n^{\text{non-param}} \triangleq \frac{1}{n}\sum_{s=1}^{n}\boldsymbol{h}(\mathbf{X}^{(s)})\boldsymbol{h}(\mathbf{X}^{(s)})^\top - \big(n^{-1}\sum_{s=1}^{n}\boldsymbol{h}(\mathbf{X}^{(s)})\big)\big(n^{-1}\sum_{s=1}^{n}\boldsymbol{h}(\mathbf{X}^{(s)})\big)^\top, \tag{14}$$

where $h(\cdot)$ denotes the oracle estimator of the transform functions in Def. 1. The core of the analysis lies on bounding the estimation error of the covariance matrix in terms of the ℓ_∞. In comparison with the previous work Liu et al. (2009), we cannot directly access the samples $\left\{\mathbf{X}^{(s)}\right\}_{1\leq s\leq n}$. Instead, we have to use the perturbed samples $\{\widehat{\mathbf{X}}^{(s)}\}_{1\leq s\leq n}$, which will lead to additional errors in the estimation of the covariance matrix. How to bound these errors constitutes the technical bottleneck.

Define β as $\frac{1}{2}\wedge\frac{a}{4}\wedge\frac{2\alpha+1}{4}$ and set $\delta_{n,d,p}$ in equation 9 as

$$\delta_{n,d,p} = \frac{c_0}{(\log n)n^{1/4}}+\frac{c_1\log^2(np)}{\sqrt{\log(d-p)}(d-p)^{\beta/4}}. \tag{15}$$

Then we conclude,

Theorem 5. *Under the Assumptions 2 and 3, the following bound*

$$\left\|\mathbf{\Sigma}_n^{\text{non-param}}-\widehat{\mathbf{\Sigma}}_n^{\text{non-param}}\right\|_\infty \lesssim \sqrt{\log n\vee\log(d-p)}\cdot\left(\frac{\sqrt{\log n}}{n^{1/4}}\vee\frac{\sqrt{\log(d-p)}}{(d-p)^{\beta/4}}\right)$$

holds with probability exceeding $1-o(1)$, where $\delta_{n,d,p}$ is set as equation 15, β is defined as $\frac{1}{2}\wedge\frac{a}{4}\wedge\frac{2\alpha+1}{4}$, and $\mathbf{\Sigma}_n^{\text{non-param}}$ and $\widehat{\mathbf{\Sigma}}_n^{\text{non-param}}$ are defined in equation 14 and equation 10, respectively.

Having obtained the covariance matrix $\widehat{\mathbf{\Sigma}}_n^{\text{non-param}}$, we estimate the graphical structure of $\mathbf{X}$ by plugging $\widehat{\mathbf{\Sigma}}_n^{\text{non-param}}$ into the graphical lasso estimator Tibshirani et al. (2015), which is put in equation 11. Adopting the same strategy as in Thm. 1, we can obtain the conditions for the correct recovery of graphical structure under Assumptions 1, 2, and 3. With a slight abuse of notations, we redefine the support set S, the parameters deg, $\kappa_{\mathbf{\Gamma}}$, and $\kappa_{\mathbf{\Sigma}}$ w.r.t the covariance matrix $\mathbf{\Sigma}^\natural$ and matrix $\mathbf{\Theta}^\natural$ that are correspond to the nonparanormal distribution. Then we can show the following results on the graph recovery.

Theorem 6. *Let* $\lambda_{\text{non-param}} \asymp \theta^{-1}\left(\frac{\log n}{n^{1/4}} \vee \frac{\log(d-p)}{(d-p)^{\beta/4}}\right)$. *Provided that*

$$\sqrt{\log n \vee \log(d-p)}\left(\frac{\sqrt{\log n}}{n^{1/4}} \vee \frac{\sqrt{\log(d-p)}}{(d-p)^{\beta/4}}\right) \lesssim \frac{\theta}{(\theta+8)(\deg)\cdot \kappa_{\mathbf{\Sigma}}^3 \kappa_{\mathbf{\Gamma}}}, \tag{16}$$

the property $(\widehat{\mathbf{\Theta}}^{\text{non-param}})_{i,j} = 0$ *for all* (i,j) *outside the support set* S*, i.e.,* $(i,j) \in S^c$*, holds with probability* $1 - o(1)$*, where* $\beta \triangleq \frac{1}{2} \wedge \frac{a}{4} \wedge \frac{2\alpha+1}{4}$*. Furthermore, if*

$$\min_{(i,j)\in S}\left|\mathbf{\Theta}^{\natural}_{i,j}\right| \geq 2\kappa_{\Gamma}(1+8\theta^{-1})\sqrt{\log n \vee \log(d-p)} \times \left(\frac{\sqrt{\log n}}{n^{1/4}} \vee \frac{\sqrt{\log(d-p)}}{(d-p)^{\beta/4}}\right),$$

then $\operatorname{sign}(\widehat{\mathbf{\Theta}}^{\text{non-param}}) = \operatorname{sign}(\mathbf{\Theta}^{\natural})$ *with probability* $1 - o(1)$.

4.5 Discussions

Following the same logic as in the discussion of the parametric method, we can obtain the minimum sample number n in terms of dimension d, and length p from equation 16. Treating parameters θ, deg, $\kappa_{\mathbf{\Sigma}}$, and $\kappa_{\mathbf{\Gamma}}$ as some constants, the condition equation 16 requires the left-hand side to be constant.

Minimum Sample Size n. We can show that

$$n \gtrsim (\deg)^4 \log^2 n \left(\log^2 n \vee \log^2(d-p)\right).$$

In contrast, the previous work Liu et al. (2009) only requires the sample number n to satisfy $n \gtrsim (\deg)^4 \log^4(n)$. Hence our result experience a loss of up to $\frac{\log^2(d-p)}{\log^2(n)} \vee 1$. Since this inflation is closely related to the dimension d, we conclude the loss is due to the indirect measurement scheme.

Minimum Projection Dimension d. For the dimension d, we require

$$d \geq p + (\deg)^{\beta/4} \log^{\beta/8}(d-p)\left(\log^{\beta/8} n \vee \log^{\beta/8}(d-p)\right),$$

which is a slightly larger than the dimension p. To the best of our knowledge, this is the first condition involving the dimension d for the nonparametric learning of the graphical structure. Whether we can use a compressive sensing system, namely, $d < p$, for the nonparametric method still remains an open-problem.

Having showed that the correct graphical structure can be obtained under mild conditions, next we will present some numerical experiments to validate our theoretical analysis.

5 Simulation Results

This section presents the numerical results, which applies to both the synthetic and the real-world data. Due to the space limit, we only put a subset of our numerical experiments and leave the rest to the supplementary material.

5.1 Synthetic Data

We adopt the classical setting as Tibshirani et al. (2015) (9.5, P 252), where the ground-truth matrix $\mathbf{\Theta}^{\natural}$ is set as (i) $\mathbf{\Theta}^{\natural}_{ij} = \rho_1$ if $i = j$; (ii) $\mathbf{\Theta}^{\natural}_{ij} = \rho_2$ if $|i-j| = 1$; and (iii) $\mathbf{\Theta}^{\natural}_{ij} = 0$ otherwise. The corresponding edge set is denoted as $E^{\natural}$. Simulations with other types of graphs are referred to the supplementary material.

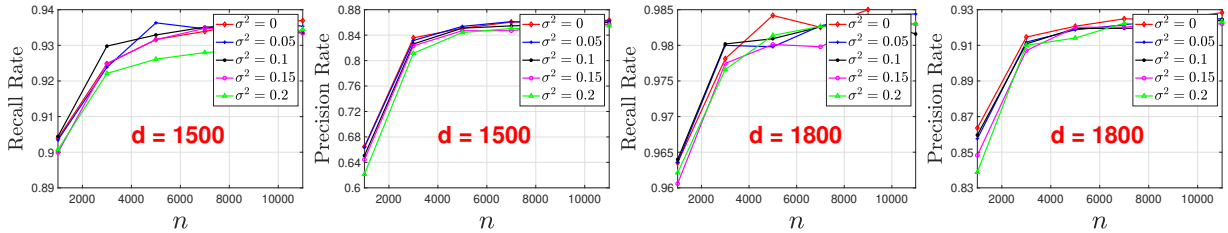

Figure 2: The signal dimension p is fixed as 2000. We study the impact of sample size n for $d = 1500$ (left two plots), and $d = 1800$ (right two plots).

Parametric Method. We set $\rho_1 = 1$ and $\rho_2 = 0.4$. First we create samples $\mathbf{X}^{(i)}$ and then mask it by the sensing relation $\mathbf{Y}^{(i)} = \mathbf{A}\mathbf{X}^{(i)} + \mathbf{W}^{(i)}$, where $A_{ij} \overset{\text{i.i.d}}{\sim} \mathsf{N}(0,1)$ and $W^{(i)} \sim \mathsf{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{n\times n})$. [1] Using Alg. 1, we reconstruct $\widehat{\mathbf{\Theta}}^{\text{param}}$ and evaluate it with the recall rate and precision rate with the results being put in Fig. 5.

We confirm that the edge of graphical model can be selected correctly with high probability even when the dimension of the projection space is much lower than the dimension of the signal. In addition, we notice a threshold effect on n when $\sigma^2 = 1$. This can be explained by the parameter τ_∞ in equation 5, which contains the term $\frac{c_6\sigma^2 \log p}{d}(1 + c_7(\sqrt{\frac{d}{n}} \vee \frac{d}{n}))$. If d is not high enough, this term is still lower-bounded by $c_6\sigma^2 \log p/d$ even as $n \to \infty$, which means a large τ_∞ and further the violation of the conditions in Thm. 1.

From Fig. 5, we confirm that the edge of graphical model can be selected correctly with high probability even when the dimension of the projection space is much lower than the dimension of the signal. In addition, we notice a threshold effect on n when $\sigma^2 = 1$ in 5. This can be explained by the parameter τ_∞ in equation 5, which contains the term $\frac{c_8\sigma^2 \log p}{d}(1 + c_9(\sqrt{\frac{d}{n}} \vee \frac{d}{n}))$. If d is not high enough, this term is still lower-bounded by $c_8\sigma^2 \log p/d$ even as $n \to \infty$, which means a large τ_∞ and further the violation of the assumption in Thm. 1.

Generally speaking, we find larger sample size n, higher dimension d, and lower noise variance σ^2 contribute to the more accurate edge selection, which is consistent with our intuition and verifies Thm. 1.

Non-Parametric Method. In Fig. 3, we compare the algorithms for the uniform distribution where the marginal distribution of each entry is uniformly within the region $[0, 1]$. We can see that our method has a significant improvement when comparing with the method in Liu et al. (2009) (a widely-used baseline).

In addition, we evaluate the performance with the following three types of marginal distribution for the RV $\mathbf{X}$:

- uniform distribution within the region $[0, 1]$;
- exponential distribution, i.e., e^{-z} for $z \geq 0$;
- Gaussian mixture, i.e., $0.25\sum_{i=1}^{4} \mathsf{N}(\mu_i, 10^{-2})$, where $\mu_i \in \{\pm 0.25, \pm 0.5\}$.

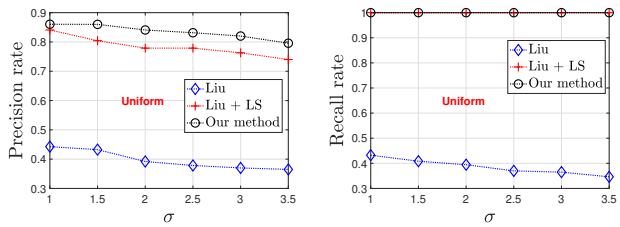

Figure 3: We study the impact of noise with $p = 50$ and $d = 200$. **Liu** refers to the method in Liu et al. (2009); **Liu + LS** refers to first performing denoising with our method (i.e. least square) followed by Liu et al. (2009), which estimates the CDF without density deconvolution. We believe the performance gap will increase with a larger noise variance.

Due to the spatial limit, we leave the numerical results to the supplementary material, which suggests our algorithm has an improvement in both the recall rate and precision rate.

[1] Notice that $\Sigma^{\natural}_{ii} \neq 1$ in this setting. Hence we need to adapt the estimation of the diagonal elements of $\widehat{\Sigma}$.

Table 1: Recall rate and precision rate of edge selection on real-world databases, namely, GSE148426 Bhattacharya et al. (2020) and GSE137140 Asakura et al. (2020). The precision matrix $\mathbf{\Theta}$ learned by direct observations is assumed to be the ground-truth. For more results, please refer to the supplementary material.

d/p	Healthy group		Unhealthy group	
	GSE 148426	GSE 137140	GSE 148426	GSE 137140
Recall Rate				
2	0.9494	0.8892	0.9424	0.9692
5	1	0.9950	1	0.9692
10	1	1	1	1
12	1	1	1	1
15	1	1	1	1
20	1	1	1	1
Precision Rate				
2	1	1	0.9704	0.9692
5	0.9080	0.9900	0.9205	0.9692
10	0.9080	0.9341	0.9205	0.9420
12	0.9080	0.9475	0.9329	0.9420
15	0.9080	0.9566	0.9456	0.9420
20	0.9518	0.9613	0.9586	0.9420

5.2 Real-World Data

We now consider the real-world databases, which consists of 5 databases: Carolina Breast Cancer (GSE148426) with 2497 samples (patients) Bhattacharya et al. (2020), Lung Cancer (GSE137140) with 3924 samples Asakura et al. (2020), Ovarian Cancer (GSE106817) with 4046 samples Yokoi et al. (2018), Colorectal Cancer (GSE115513) with 1513 samples Slattery et al. (2016), and Esophageal Squamous Cell Carcinoma (GSE122497) with 5531 samples Sudo et al. (2019). Each database is divided into two categories, i.e., **Healthy group** and **Patients**, where the measurements are given as the concentration of miRNAs. The miRNAs are known to have dependency among each other, i.e., a non-diagonal precision matrix, and hence there is an underlying graphical model describing these dependency structure based on the associated precision matrix.

The sensing matrix $\mathbf{A} \in \mathbb{R}^{d\times p}$ is assumed to be $A_{ij} \overset{\text{i.i.d}}{\sim} \mathsf{N}(0,1)$ and the variance of the measurement noise is set to one, $W_i \overset{\text{i.i.d}}{\sim} \mathsf{N}(0,1)$. The goal is to reconstruct the underlying dependency graph among miRNAs.

Evaluation. We adopt the nonparametric method for the evaluation due to its wider applications. The precision matrix $\mathbf{\Theta}$ learned with noiseless direct measurements using the method in Liu et al. (2009) is assumed to be the ground-truth. We evaluate the performance of our estimator in both the recall rate and precision rate of the edge selection. The results are shown in Tab. 4. [2] These experiments confirm that our estimator can obtain the correct dependency relation with high-probability.

6 Conclusions

This is the first work on learning of the graphical structure with noisy indirect measurements, where both the parametric and non-parametric methods are investigated. For the parametric method, we considered the Gaussian graphical model and learned the graphical structure by using only indirect low-dimensional observations, i.e., the observations obtained via compressive sensing of the desired signal. For the non-parametric method, we relaxed the gaussian distribution to the nonparanormal distribution. We established a non-asymptotic uniform bound on the errors of the CDF estimation. To the best of our knowledge, this

[2] Only a part of the results are put here and the rest results are left to the supplementary due to the spatial limit.

is the first such results on the CDF error bounds when only limited information exists regarding the noise distribution. For both scenarios, we showed our estimator can generate the correct graphical structure under mild conditions, from which the relation between the sample number n, the signal dimension p, and the projection dimension d are obtained. In addition, we provided numerical experiments, using both synthetic and real world miRNA data, to corroborate the correctness of our theoretical results.

References

Keisuke Asakura, Tsukasa Kadota, Juntaro Matsuzaki, Yukihiro Yoshida, Yusuke Yamamoto, Kazuo Nakagawa, Satoko Takizawa, Yoshiaki Aoki, Eiji Nakamura, Junichiro Miura, et al. A mirna-based diagnostic model predicts resectable lung cancer in humans with high accuracy. *Communications biology*, 3(1):1–9, 2020.

Arjun Bhattacharya, Alina M. Hamilton, Melissa A. Troester, and Michael I. Love. Decompress: tissue compartment deconvolution of targeted mrna expression panels using compressed sensing. *bioRxiv*, 2020. doi: 10.1101/2020.08.14.250902. URL `https://www.biorxiv.org/content/early/2020/08/17/2020.08.14.250902`.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Tony Cai, Weidong Liu, and Xi Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

Venkat Chandrasekaran, Pablo A Parrilo, and Alan S Willsky. Latent variable graphical model selection via convex optimization. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1610–1613. IEEE, 2010.

Itai Dattner and B Reiser. Estimation of distribution functions in measurement error models. *Journal of Statistical Planning and Inference*, 143(3):479–493, 2013.

Itai Dattner, Alexander Goldenshluger, and Anatoli Juditsky. On deconvolution of distribution functions. *The Annals of Statistics*, pp. 2477–2501, 2011.

Jianqing Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, pp. 1257–1272, 1991.

Jianqing Fan, Han Liu, Yang Ning, and Hui Zou. High dimensional semiparametric latent graphical model for mixed data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):405–421, 2017.

Nir Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805, 2004.

Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

Peter Hall and Alexander Meister. A ridge-parameter approach to deconvolution. *The Annals of Statistics*, 35(4):1535–1558, 2007.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

Svante Janson et al. *Gaussian hilbert spaces*, volume 129. Cambridge university press, 1997.

Johanna Kappus and Gwennaëlle Mabon. Adaptive density estimation in deconvolution problems with unknown error distribution. *Electronic journal of statistics*, 8(2):2879–2904, 2014.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. 2009.

Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(10), 2009.

Han Liu, Min Xu, Haijie Gu, Anupam Gupta, John Lafferty, and Larry Wasserman. Forest density estimation. *Journal of Machine Learning Research*, 12(Mar):907–951, 2011.

Han Liu, Fang Han, Ming Yuan, John Lafferty, Larry Wasserman, et al. High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326, 2012.

Benjamin M Marlin and Kevin P Murphy. Sparse gaussian graphical models with unknown block structure. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 705–712, 2009.

Elias Masry. Multivariate probability density deconvolution for stationary random processes. *IEEE Transactions on Information Theory*, 37(4):1105–1115, 1991.

Rahul Mazumder and Trevor Hastie. The graphical lasso: New insights and alternatives. *Electronic journal of statistics*, 6:2125, 2012.

Johannes Müller, Christina Kuttler, and Burkhard A Hense. Sensitivity of the quorum sensing system is achieved by low pass filtering. *Biosystems*, 92(1):76–81, 2008.

Marianna Pensky and Brani Vidakovic. Adaptive wavelet estimator for nonparametric density deconvolution. *The Annals of Statistics*, 27(6):2033–2053, 1999.

Cao Xuan Phuong. Deconvolution of cumulative distribution function with unknown noise distribution. *Acta Applicandae Mathematicae*, 170(1):483–514, 2020.

Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, Bin Yu, et al. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5: 935–980, 2011.

R Tyrrell Rockafellar. *Convex analysis*. Number 28. Princeton university press, 1970.

Martha L Slattery, Jennifer S Herrick, Daniel F Pellatt, John R Stevens, Lila E Mullany, Erica Wolff, Michael D Hoffman, Wade S Samowitz, and Roger K Wolff. Microrna profiles in colorectal carcinomas, adenomas and normal colonic mucosa: variations in mirna expression and disease progression. *Carcinogenesis*, 37(3): 245–261, 2016.

Kazuki Sudo, Ken Kato, Juntaro Matsuzaki, Narikazu Boku, Seiichiro Abe, Yutaka Saito, Hiroyuki Daiko, Satoko Takizawa, Yoshiaki Aoki, Hiromi Sakamoto, et al. Development and validation of an esophageal squamous cell carcinoma detection model by large-scale microrna profiling. *JAMA network open*, 2(5): e194573–e194573, 2019.

Michel Talagrand. *Upper and lower bounds for stochastic processes: modern methods and classical problems*, volume 60. Springer Science & Business Media, 2014.

Robert Tibshirani, Martin Wainwright, and Trevor Hastie. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC, 2015.

Dang Duc Trong and Cao Xuan Phuong. Deconvolution of a cumulative distribution function with some non-standard noise densities. *Vietnam Journal of Mathematics*, 47(2):327–353, 2019.

Antonia M Tulino, Sergio Verdú, and Sergio Verdu. *Random matrix theory and wireless communications*. Now Publishers Inc, 2004.

Roman Vershynin. High dimensional probability, 2016.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Pan Xu and Quanquan Gu. Semiparametric differential graph models. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 1072–1080, 2016.

Lingzhou Xue, Hui Zou, et al. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics*, 40(5):2541–2571, 2012.

Akira Yokoi, Juntaro Matsuzaki, Yusuke Yamamoto, Yutaka Yoneoka, Kenta Takahashi, Hanako Shimizu, Takashi Uehara, Mitsuya Ishikawa, Shun-ichi Ikeda, Takumi Sonoda, et al. Integrated extracellular microrna profiling for ovarian cancer screening. *Nature communications*, 9(1):1–10, 2018.

Élie Youndjé and Martin T Wells. Optimal bandwidth selection for multivariate kernel deconvolution density estimation. *Test*, 17(1):138–162, 2008.

Cun-Hui Zhang. Fourier methods for estimating mixing densities and distributions. *The Annals of Statistics*, pp. 806–831, 1990.

Tuo Zhao, Kathryn Roeder, and Han Liu. Positive semidefinite rank-based correlation matrix estimation with application to semiparametric graph estimation. *Journal of Computational and Graphical Statistics*, 23(4):895–922, 2014.

A Analysis of Covariance Matrix Estimation via Parametric Method

For the convenience of analysis, we rescale the sensing matrix $\widetilde{\mathbf{A}}$ such that $\widetilde{A}_{ij} \sim \mathsf{N}(0, d^{-1})$.

Lemma 3. *Consider the covariance estimator $\widehat{\mathbf{\Sigma}}_n^{\text{param}}$ which reads*

$$\widehat{\mathbf{\Sigma}}_n^{\text{param}} = \mathbf{I} + \frac{d}{d+1}\left[\widetilde{\mathbf{A}}^\top \left(\frac{1}{n}\sum_{i=1}^n \mathbf{Y}^{(i)}\mathbf{Y}^{(i)\top}\right)\widetilde{\mathbf{A}}\right]_{\text{off}}, \tag{17}$$

where $(\cdot)_{\text{off}}$ denotes the operation of picking non-diagonal entries. We then have $\left\|\widehat{\mathbf{\Sigma}}_n^{\text{param}} - \mathbf{\Sigma}^\natural\right\|_\infty \le \tau_\infty$ holding with probability at least $1 - c_0 p^{-1} - c_1 p^2 e^{-c_2 d} - c_3 p^2 e^{-c_4 p}$.

We begin the analysis by redefining the following events,

$$\begin{aligned}
\mathcal{E}_1 &\triangleq \left\{1 - c_0\sqrt{\frac{\log p}{d}} \le \left\|\widetilde{\mathbf{A}}_i\right\|_2 \le 1 + c_0\sqrt{\frac{\log p}{d}}, \quad \forall\, i\right\};\\
\mathcal{E}_2 &\triangleq \left\{|\left\langle \widetilde{\mathbf{A}}_i, \widetilde{\mathbf{A}}_j\right\rangle| \lesssim \sqrt{\frac{\log p}{d}} \vee \frac{\log p}{d}, \quad \forall\, i \neq j\right\};\\
\mathcal{E}_3 &\triangleq \left\{\left\|\widehat{\mathbf{\Sigma}}_n^{\text{param}} - \mathbf{\Sigma}^\natural\right\|_\infty \lesssim \sqrt{\frac{\log p}{n}}\right\};\\
\mathcal{E}_4(\mathbf{B}) &\triangleq \left\{\left\|\sum_{\ell\neq 1} B_\ell \widetilde{\mathbf{A}}_\ell\right\|_2 \lesssim \sqrt{\sum_{\ell\neq 1} B_\ell^2}\right\}, \quad \text{where } \mathbf{B}\in\mathbb{R}^p \text{ is a fixed vector.}
\end{aligned}$$

Additionally, we define $\Psi(\mathcal{E})$ as $\mathbb{E}\mathbb{1}(\mathcal{E})$. Moreover, our analysis focuses on the region when $d \gg \log p$.

A.1 Main Structure

Having collected all the lemmas, we turn to the proof of Lemma 1. Notice that by the definition of our estimator, $\left\|\widehat{\mathbf{\Sigma}}_n^{\text{param}} - \mathbf{\Sigma}^\natural\right\|_\infty = \max_{i\neq j}\left|\left(\widehat{\mathbf{\Sigma}}_n^{\text{param}}\right)_{i,j} - \mathbf{\Sigma}^\natural_{i,j}\right|$. Hence, we only consider the off-diagonal entries, which gives

$$\begin{aligned}
&\left\|\widehat{\mathbf{\Sigma}}_n^{\text{param}} - \mathbf{\Sigma}^\natural\right\|_{\text{off},\infty} = \left\|\frac{d}{d+1}\widetilde{\mathbf{A}}^\top\left(\sum_{\ell=1}^n \mathbf{Y}^{(\ell)}\mathbf{Y}^{(\ell)\top}\right)\widetilde{\mathbf{A}} - \mathbf{\Sigma}^\natural\right\|_{\text{off},\infty}\\
&\le 2\left\|\frac{d}{d+1}\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\mathbf{\Sigma}_n^{\text{param}}\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}} - \mathbf{\Sigma}^\natural\right\|_{\text{off},\infty} + \frac{2d}{n(d+1)}\left\|\widetilde{\mathbf{A}}^\top\left(\sum_{\ell=1}^n \mathbf{W}^{(\ell)}\mathbf{W}^{(\ell)\top}\right)\widetilde{\mathbf{A}}\right\|_{\text{off},\infty}\\
&\overset{①}{=} \frac{2d}{d+1}\underbrace{\left\|\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\mathbf{\Sigma}_n^{\text{param}}\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}} - \mathbb{E}_{\widetilde{\mathbf{A}}}\left(\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\mathbf{\Sigma}_n^{\text{param}}\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\right)\right\|_{\text{off},\infty}}_{\vartheta_1}\\
&+ \frac{2d}{d+1}\underbrace{\left\|\mathbb{E}_{\widetilde{\mathbf{A}}}\left(\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\mathbf{\Sigma}_n^{\text{param}}\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\right) - \mathbb{E}_{\widetilde{\mathbf{A}},\mathbf{X}}\left(\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\mathbf{\Sigma}_n^{\text{param}}\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\right)\right\|_{\text{off},\infty}}_{\vartheta_2}\\
&+ \frac{2d}{n(d+1)}\underbrace{\left\|\widetilde{\mathbf{A}}^\top\left(\sum_{\ell=1}^n \mathbf{W}^{(\ell)}\mathbf{W}^{(\ell)\top}\right)\widetilde{\mathbf{A}}\right\|_{\text{off},\infty}}_{\vartheta_3},
\end{aligned}$$

where $\mathbf{\Sigma}_n^{\text{param}}$ is defined as $n^{-1}\left(\sum_{\ell=1}^n \mathbf{X}^{(\ell)}\mathbf{X}^{(\ell)\top}\right)$, and in ① we use $d/(d+1)\mathbb{E}_{\widetilde{\mathbf{A}},\mathbf{X}}\left(\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\mathbf{\Sigma}_n^{\text{param}}\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\right) = \mathbf{\Sigma}^\natural$. Then we separately upper-bound ϑ_1, ϑ_2, and ϑ_3 conditional on the event $\bigcap_{i=1}^3 \mathcal{E}_i$. For the conciseness in

notations, define $\mathbf{Z} = \widetilde{\mathbf{A}}^\top \widetilde{\mathbf{A}} \mathbf{\Sigma}_n^{\text{param}} \widetilde{\mathbf{A}}^\top \widetilde{\mathbf{A}}$. Therefore $\vartheta_1 = \|\mathbf{Z} - \mathbb{E}\mathbf{Z}\|_{\text{off},\infty}$. For an arbitrary entry $Z_{i,j}$, we can expand it as

$$Z_{i,j} = \sum_{\ell_1,\ell_2} (\mathbf{\Sigma}_n^{\text{param}})_{\ell_1,\ell_2} \left\langle \widetilde{\mathbf{A}}_i, \widetilde{\mathbf{A}}_{\ell_1} \right\rangle \left\langle \widetilde{\mathbf{A}}_j, \widetilde{\mathbf{A}}_{\ell_2} \right\rangle, \quad i \neq j.$$

Compared to the existing work Ravikumar et al. (2011), our analysis of Z_{ij} involves fourth-order Gaussian chaos Talagrand (2014), which exhibits heavy tails, and constructs the major obstacle.

Stage I: Bounding ϑ_1. To obtain upper bound on ϑ_1, we first adopt the union bound and obtain

$$\mathbb{P}\left(\vartheta_1 \geq \delta\right) = \mathbb{P}\left(\max_{i \neq j} |Z_{i,j} - \mathbb{E}Z_{i,j}| \geq \delta\right) \leq \sum_{i \neq j} \mathbb{P}\left(|Z_{i,j} - \mathbb{E}Z_{i,j}| \geq \delta,\ i \neq j\right). \tag{18}$$

Then our focus is to bound the probability $\mathbb{P}\left(|Z_{i,j} - \mathbb{E}Z_{i,j}| \geq \delta,\ i \neq j\right)$. Without loss of generality, assume that $i = 1$ and $j = 2$ and expand $Z_{1,2}$ as $\sum_{i=1}^5 T_i$, which reads

$$\begin{aligned}
T_1 &\triangleq (\mathbf{\Sigma}_n^{\text{param}})_{1,2} \left\|\widetilde{\mathbf{A}}_1\right\|_2^2 \left\|\widetilde{\mathbf{A}}_2\right\|_2^2; \\
T_2 &\triangleq (\mathbf{\Sigma}_n^{\text{param}})_{2,1} \left(\left\langle \widetilde{\mathbf{A}}_1, \widetilde{\mathbf{A}}_2 \right\rangle\right)^2; \\
T_3 &\triangleq \sum_{\ell \neq 1} (\mathbf{\Sigma}_n^{\text{param}})_{2,\ell} \left\|\widetilde{\mathbf{A}}_2\right\|_2^2 \left\langle \widetilde{\mathbf{A}}_1, \widetilde{\mathbf{A}}_\ell \right\rangle + \sum_{\ell \neq 2} (\mathbf{\Sigma}_n^{\text{param}})_{\ell,1} \left\|\widetilde{\mathbf{A}}_1\right\|_2^2 \left\langle \widetilde{\mathbf{A}}_2, \widetilde{\mathbf{A}}_\ell \right\rangle; \\
T_4 &\triangleq \sum_{\ell \neq 1,2} (\mathbf{\Sigma}_n^{\text{param}})_{\ell,\ell} \left\langle \widetilde{\mathbf{A}}_1, \widetilde{\mathbf{A}}_\ell \right\rangle \left\langle \widetilde{\mathbf{A}}_2, \widetilde{\mathbf{A}}_\ell \right\rangle; \\
T_5 &\triangleq \sum_{\substack{\ell_1,\ell_2 \neq 1,2 \\ \ell_1 \neq \ell_2}} (\mathbf{\Sigma}_n^{\text{param}})_{\ell_1,\ell_2} \left\langle \widetilde{\mathbf{A}}_1, \widetilde{\mathbf{A}}_{\ell_1} \right\rangle \left\langle \widetilde{\mathbf{A}}_2, \widetilde{\mathbf{A}}_{\ell_2} \right\rangle.
\end{aligned}$$

Now, we separately bound the deviations $|T_i - \mathbb{E}T_i|$, $1 \leq i \leq 5$. First, we have

$$\begin{aligned}
&|T_1 - \mathbb{E}T_1| = |\,(\mathbf{\Sigma}_n^{\text{param}})_{1,2}\,| \times \left| \left\|\widetilde{\mathbf{A}}_1\right\|_2^2 \left\|\widetilde{\mathbf{A}}_2\right\|_2^2 - 1 \right| \\
&= |\,(\mathbf{\Sigma}_n^{\text{param}})_{1,2}\,| \times \left| \left\|\widetilde{\mathbf{A}}_1\right\|_2 \left\|\widetilde{\mathbf{A}}_2\right\|_2 - 1 \right| \times \left(\left\|\widetilde{\mathbf{A}}_1\right\|_2 \left\|\widetilde{\mathbf{A}}_2\right\|_2 + 1 \right) \\
&\overset{②}{\leq} |\,(\mathbf{\Sigma}_n^{\text{param}})_{1,2}\,| \times \left(\left| \left\|\widetilde{\mathbf{A}}_1\right\|_2 - 1 \right| \left\|\widetilde{\mathbf{A}}_2\right\|_2 + \left| \left\|\widetilde{\mathbf{A}}_2\right\|_2 - 1 \right| \right) \times \left(\left(1 + c_0\sqrt{\log p/d}\right)^2 + 1 \right) \\
&\overset{③}{\leq} |\,(\mathbf{\Sigma}_n^{\text{param}})_{1,2}\,| \times c_0\sqrt{\log p/d}(2 + c_0\sqrt{\log p/d}) \times \left(\left(1 + c_0\sqrt{\log p/d}\right)^2 + 1 \right) \\
&\overset{④}{\lesssim} |\,(\mathbf{\Sigma}_n^{\text{param}})_{1,2}\,| \sqrt{\log p/d} \overset{⑤}{\lesssim} \sqrt{\frac{\log p}{d}} \left(\left|\Sigma_{1,2}^{\natural}\right| + \sqrt{\frac{\log p}{n}} \right),
\end{aligned} \tag{19}$$

where in ② and ③ we condition on event $\mathcal{E}_1$, ④ is due to $d \gg \log p$, and ⑤ is by the definition of event $\mathcal{E}_3$. For $|T_2 - \mathbb{E}T_2|$, by invoking Lemma 8,

$$\begin{aligned}
|T_2 - \mathbb{E}T_2| &\lesssim |\,(\mathbf{\Sigma}_n^{\text{param}})_{2,1}\,| \times \left(\frac{\log p}{d} + \frac{\sqrt{\log p}}{d^{3/2}} \right) \overset{⑥}{\lesssim} \left|(\mathbf{\Sigma}_n^{\text{param}})_{2,1}\right| \frac{\log p}{d} \\
&\overset{⑦}{\lesssim} \frac{\log p}{d} \left(\left|\Sigma_{2,1}^{\natural}\right| + \sqrt{\frac{\log p}{n}} \right),
\end{aligned} \tag{20}$$

holds with probability exceeding $1 - 2p^{-3}$, where ⑥ is because $d \gg \log p$, and ⑦ is due to event $\mathcal{E}_3$.

We continue to bound $|T_3 - \mathbb{E}T_3|$ by

$$
\begin{aligned}
|T_3 - \mathbb{E}T_3| \leq & \left\|\widetilde{\mathbf{A}}_1\right\|_2^2 \times \left|\sum_{\ell\neq 2}(\mathbf{\Sigma}_n^{\text{param}})_{1,\ell}\left\langle\widetilde{\mathbf{A}}_2, \widetilde{\mathbf{A}}_\ell\right\rangle\right| + \left\|\widetilde{\mathbf{A}}_2\right\|_2^2 \times \left|\sum_{\ell\neq 1}(\mathbf{\Sigma}_n^{\text{param}})_{\ell,2}\left\langle\widetilde{\mathbf{A}}_1, \widetilde{\mathbf{A}}_\ell\right\rangle\right| \\
\leq & \left(1 + c_0\sqrt{\frac{\log p}{d}}\right)^2 \left|\sum_{\ell\neq 2}(\mathbf{\Sigma}_n^{\text{param}})_{1,\ell}\left\langle\widetilde{\mathbf{A}}_2, \widetilde{\mathbf{A}}_\ell\right\rangle\right| + \left(1 + c_0\sqrt{\frac{\log p}{d}}\right)^2 \left|\sum_{\ell\neq 1}(\mathbf{\Sigma}_n^{\text{param}})_{\ell,2}\left\langle\widetilde{\mathbf{A}}_1, \widetilde{\mathbf{A}}_\ell\right\rangle\right| \\
\overset{⑧}{\lesssim} & \left(1 + c_0\sqrt{\frac{\log p}{d}}\right)^2 \times \sqrt{\frac{\log p}{d}}\left(\sqrt{\sum_{\ell\neq 2}(\mathbf{\Sigma}_n^{\text{param}})_{1,\ell}^2} + \sqrt{\sum_{\ell\neq 1}(\mathbf{\Sigma}_n^{\text{param}})_{\ell,2}^2}\right) \\
\overset{⑨}{\lesssim} & \sqrt{\frac{\log p}{d}}\left(\sqrt{\sum_{\ell\neq 2}(\mathbf{\Sigma}_n^{\text{param}})_{1,\ell}^2} + \sqrt{\sum_{\ell\neq 1}(\mathbf{\Sigma}_n^{\text{param}})_{\ell,2}^2}\right) \overset{Ⓐ}{\lesssim} \sqrt{\frac{\log p}{d}}\left(\left\|\mathbf{\Sigma}_1^\natural\right\|_2 + \left\|\mathbf{\Sigma}_2^\natural\right\|_2 + \sqrt{\frac{p\log p}{n}}\right)
\end{aligned} \tag{21}
$$

which holds for probability exceeding $1 - c_0p^{-3} - c_1e^{-c_2d}$, where in ⑧ we invoke Lemma 9, in ⑨ we use that $\mathbf{\Sigma}_n^{\text{param}}$ is symmetric and $d \gg \log p$, and Ⓐ is due to event $\mathcal{E}_3$.

To bound $|T_4 - \mathbb{E}T_4|$, we invoke Lemma 10,

$$
\begin{aligned}
|T_4 - \mathbb{E}T_4| \lesssim & \frac{\left|\sum_{\ell>2}(\mathbf{\Sigma}_n^{\text{param}})_{\ell,\ell}\right|}{d}\sqrt{\frac{\log p}{d}} + \frac{\sqrt{\log p}}{d}\left[\sqrt{\sum_{\ell>2}(\mathbf{\Sigma}_n^{\text{param}})_{\ell,\ell}^2} \vee \left(\sqrt{\log p}(\max_\ell|(\mathbf{\Sigma}_n^{\text{param}})_{\ell,\ell}|)\right)\right] \\
\overset{Ⓑ}{\lesssim} & \frac{p\sqrt{\log p}}{d^{3/2}}\left(1 + c_0\sqrt{\frac{\log p}{n}}\right) + \frac{\sqrt{p\log p}}{d}\underbrace{\left[\left(1 + c_0\sqrt{\frac{\log p}{n}}\right) \vee \sqrt{\frac{\log p}{p}}\left(1 + c_0\sqrt{\frac{\log p}{n}}\right)\right]}_{1+c_0\sqrt{\frac{\log p}{n}}},
\end{aligned} \tag{22}
$$

holds with probability exceeding $1 - 2p^{-3}$, where Ⓑ is by definition of event $\mathcal{E}_3$.

The deviation $|T_5 - \mathbb{E}T_5|$ is bounded via Lemma 11, which gives

$$
|T_5 - \mathbb{E}T_5| \lesssim \frac{\log p}{d}\left(1 + c_0\sqrt{\frac{p}{d}}\right)^2 \left|\!\left|\!\left|\mathbf{\Sigma}^\natural\right|\!\right|\!\right|_{\text{off,F}} \overset{Ⓒ}{\lesssim} \frac{\log p}{d}\left(1 + \frac{c_1p}{d}\right)\left(\left|\!\left|\!\left|\mathbf{\Sigma}^\natural\right|\!\right|\!\right|_{\text{off,F}} + p\sqrt{\frac{\log p}{n}}\right), \tag{23}
$$

with probability exceeding $1 - 4p^{-3} - e^{-c_0p}$, where in Ⓒ we condition on event $\mathcal{E}_3$.

Combining (19), (20), (21), (22), and (23), we conclude that

$$
\begin{aligned}
|Z_{i,j} - \mathbb{E}Z_{i,j}| \lesssim & \sqrt{\frac{\log p}{d}}\left(\left\|\mathbf{\Sigma}_i^\natural\right\|_2 + \left\|\mathbf{\Sigma}_j^\natural\right\|_2\right) + \frac{\log p}{d}\left(1 + \frac{c_1p}{d}\right)\left|\!\left|\!\left|\mathbf{\Sigma}^\natural\right|\!\right|\!\right|_{\text{off,F}} \\
& + \frac{\log p\sqrt{p}}{\sqrt{nd}} + \frac{c_0p(\log p)^{3/2}}{n^{1/2}d}\left(1 + \frac{c_1p}{d}\right) + \frac{\sqrt{p\log p}}{d} + \frac{p\sqrt{\log p}}{d^{3/2}},
\end{aligned}
$$

with probability $1 - c_0p^{-3} - c_1e^{-c_2d} - c_3e^{-c_4p}$. We conclude the proof by plugging into (18) and $\left|\!\left|\!\left|\mathbf{\Sigma}_n^{\text{param}} - \mathbf{\Sigma}^\natural\right|\!\right|\!\right|_{\text{off},\infty} \leq \left\|\mathbf{\Sigma}_n^{\text{param}} - \mathbf{\Sigma}^\natural\right\|_\infty \lesssim \sqrt{\log p/n}$ according to event $\mathcal{E}_3$.

Stage II: Bounding ϑ_2. We rewrite ϑ_2 as $\left\|\mathbf{\Sigma}_n^{\text{param}} - \mathbf{\Sigma}^\natural\right\|_{\text{off},\infty}$ and invoke event $\mathcal{E}_3$.

Stage III: Bounding ϑ_3. This stage is completed by invoking Lemma 12. Combing the above three stages will then yield the proof.

A.2 Supporting Lemmas

We first compute the values of $\Psi(\mathcal{E}_i)$, $1 \leq i \leq 3$.

Lemma 4. $\Psi(\mathcal{E}_1) \geq 1 - 2p^{-1}$.

Proof. We conclude that

$$\mathbb{P}\left(\left|\left\|\widetilde{\mathbf{A}}_i\right\|_2 - 1\right| \geq c_0\sqrt{\frac{\log p}{d}},\ \exists 1 \leq i \leq p\right) \leq p\mathbb{P}\left(\left|\left\|\widetilde{\mathbf{A}}_i\right\|_2 - 1\right| \geq c_0\sqrt{\frac{\log p}{d}}\right) \overset{①}{\leq} 4p\exp\left(-2d \times \frac{\log p}{d}\right) = 2p^{-1},$$

where ① is due to the properties of χ^2 distribution.

□

Lemma 5. *Conditional on $\mathcal{E}_1$, we have $\Psi(\mathcal{E}_2) \geq 1 - 2p^{-1}$.*

Proof. Due to the independence between $\widetilde{\mathbf{A}}_i$ and $\widetilde{\mathbf{A}}_j$, where $i \neq j$. We can condition on $\widetilde{\mathbf{A}}_i$ and view $\left\langle \widetilde{\mathbf{A}}_i, \widetilde{\mathbf{A}}_j \right\rangle$ as a Gaussian RV with mean zero and variance $d^{-1}\left\|\widetilde{\mathbf{A}}_i\right\|_2^2$, namely, $\mathsf{N}\left(0, d^{-1}\left\|\widetilde{\mathbf{A}}_i\right\|_2^2\right)$. Then we conclude that

$$\Psi\left(\overline{\mathcal{E}}_2 \mid \mathcal{E}_1\right) \overset{①}{\leq} p^2\Psi\left(\left|\left\langle \widetilde{\mathbf{A}}_i, \widetilde{\mathbf{A}}_j \right\rangle\right| \geq \delta \mid \mathcal{E}_1\right) \overset{②}{\leq} 2p^2\left(\Phi\left(-\frac{\sqrt{d}\delta}{\left\|\widetilde{\mathbf{A}}_i\right\|_2}\right) \times \Psi(\mathcal{E}_1)\right)$$

$$\overset{③}{\leq} 2p^2\left(\exp\left(-\frac{d\delta^2}{2\left\|\widetilde{\mathbf{A}}_i\right\|_2^2}\right) \times \Psi(\mathcal{E}_1)\right) \overset{④}{\leq} 2p^2\exp\left(-\frac{d\delta^2}{2\left(1 + c_0\sqrt{\log p/d}\right)^2}\right),$$

where ① is due to the union bound, in ② we denote $\Phi(t) = 1/\sqrt{2\pi}\int_{-\infty}^{t} e^{-x^2/2}dx$, the CDF of the normal distribution, ③ is because $\Phi(x) \leq e^{-x^2/2}$, and ④ is according to event $\mathcal{E}_1$.

In the end, we complete the proof by setting δ as $c_1\left(\sqrt{\frac{\log p}{d}} \vee \frac{\log p}{d}\right)$, which yields $\Psi(\overline{\mathcal{E}}_2 \mid \mathcal{E}_1) \leq 2p^{-1}$. □

Lemma 6. $\Psi(\mathcal{E}_3) \geq 1 - 4p^{-1}$.

Proof. The proof can be found in the proof of Thm. 1 and Thm. 4 in Cai et al. (2011). □

Lemma 7. *For an arbitrary fixed vector $\mathbf{B} \in \mathbb{R}^p$, we have $\Psi(\mathcal{E}_4(\mathbf{B})) \geq 1 - e^{-0.8d}$.*

Proof. Notice that $\sum_{\ell\neq 1} B_\ell \widetilde{\mathbf{A}}_\ell$ is a vector satisfying $\mathsf{N}\left(\mathbf{0}, d^{-1}\left(\sum_{\ell\neq 1} B_\ell^2\right)\mathbf{I}\right)$. Hence, $\frac{d}{\sum_{\ell\neq 1} B_\ell^2}\|\sum_{\ell\neq 1} B_\ell \widetilde{\mathbf{A}}_\ell\|_2^2$ is a χ^2 RV with freedom d, which suggests

$$\mathbb{P}\left(\left\|\sum_{\ell\neq 1} B_\ell \widetilde{\mathbf{A}}_\ell\right\|_2^2 \geq 4\left(\sum_{\ell\neq 1} B_\ell^2\right)\right) = \mathbb{P}\left(\frac{d}{\sum_{\ell\neq 1} B_\ell^2}\left\|\sum_{\ell\neq 1} B_\ell \widetilde{\mathbf{A}}_\ell\right\|_2^2 \geq 4d\right) \overset{①}{\leq} \exp\left(\frac{d}{2}\left(\log 4 - 3\right)\right) \leq e^{-0.8d},$$

where ① is due to the properties of χ^2 distribution. □

Lemma 8. *Conditional on event $\mathcal{E}_1$,*

$$\left|\sum_{\ell\neq 1} B_\ell \left\langle \widetilde{\mathbf{A}}_1, \widetilde{\mathbf{A}}_\ell \right\rangle^2 - d^{-1}\left(\sum_{\ell\neq 1} B_\ell\right)\right| \lesssim \frac{1}{d}\left(\sqrt{\log p}\sqrt{\sum_{\ell\neq 1} B_\ell^2} \vee (\log p)\left(\max_{\ell\neq 1}|B_\ell|\right)\right) + \frac{\sum_{\ell\neq 1} B_\ell}{d}\sqrt{\frac{\log p}{d}}$$

holds with probability exceeding $1 - 2p^{-3}$, for an arbitrary fixed vector $\mathbf{B} \in \mathbb{R}^p$.

Proof. First, we decompose the above term as

$$
\begin{aligned}
&\left|\sum_{\ell\neq 1} B_\ell \left\langle \widetilde{\mathbf{A}}_1, \widetilde{\mathbf{A}}_\ell \right\rangle^2 - d^{-1}\left(\sum_{\ell\neq 1} B_\ell\right)\right| \\
\leq\ & \frac{\left\|\widetilde{\mathbf{A}}_1\right\|_2^2}{d}\left|\sum_{\ell\neq 1} B_\ell \left(\frac{\sqrt{d}\left\langle \widetilde{\mathbf{A}}_1, \widetilde{\mathbf{A}}_\ell \right\rangle}{\left\|\widetilde{\mathbf{A}}_1\right\|_2}\right)^2 - \sum_{\ell\neq 1} B_\ell\right| + d^{-1}\left|\sum_{\ell\neq 1} B_\ell\right| \left|\left\|\widetilde{\mathbf{A}}_1\right\|_2^2 - 1\right| \\
\overset{①}{\leq}\ & \frac{1}{d}\left(1 + c_0\sqrt{\frac{\log p}{d}}\right)\underbrace{\left|\sum_{\ell\neq 1} B_\ell \left(\frac{\sqrt{d}\left\langle \widetilde{\mathbf{A}}_1, \widetilde{\mathbf{A}}_\ell \right\rangle}{\left\|\widetilde{\mathbf{A}}_1\right\|_2}\right)^2 - \sum_{\ell\neq 1} B_\ell\right|}_{T} + \frac{\left|\sum_{\ell\neq 1} B_\ell\right|}{d}\sqrt{\frac{\log p}{d}} \\
\overset{②}{\lesssim}\ & \frac{T}{d} + \frac{\left|\sum_{\ell\neq 1} B_\ell\right|}{d}\sqrt{\frac{\log p}{d}}
\end{aligned}
$$

where ① is due to the definition of event $\mathcal{E}_1$, and ② is because $d \gg \log p$. Our following analysis focuses on upper-bounding T. First we define $\mathbf{\Lambda}$ as $\mathbf{\Lambda} = [\Lambda_2\ \cdots\ \Lambda_d]^\top$, where $\Lambda_i = \sqrt{d}\left\langle \widetilde{\mathbf{A}}_1, \widetilde{\mathbf{A}}_\ell \right\rangle / \left\|\widetilde{\mathbf{A}}_1\right\|_2$. Then we can rewrite T as

$$
T = \sum_{\ell\neq 1} B_\ell \Lambda_i^2 = \mathbf{\Lambda}^\top \mathrm{diag}(B_\ell)_{\ell\neq 2}\mathbf{\Lambda}.
$$

Due to the independence between $\widetilde{\mathbf{A}}_1$ and $\widetilde{\mathbf{A}}_\ell$, $\ell \neq 1$, we can condition $\widetilde{\mathbf{A}}_1$ and can view Λ_i as a Gaussian RV satisfying $\mathsf{N}(0,1)$. Invoking the Hanson-Wright inequality (Theorem 6.2.1 in Vershynin (2016)), we conclude that

$$
\mathbb{P}\left(T \geq \delta\right) \leq\ 2\exp\left(-c\left(\frac{\delta^2}{\left\|\mathrm{diag}(B_\ell)_{\ell\neq 1}\right\|_{\mathrm{F}}^2} \wedge \frac{\delta}{\left\|\mathrm{diag}(B_\ell)_{\ell\neq 1}\right\|_{\mathrm{OP}}}\right)\right).
$$

Setting δ as $\left(c_0\sqrt{\log p}\sqrt{\sum_{\ell\neq 1} B_\ell^2}\right) \vee c_1\left(\log p \times \max_{\ell\neq 1}|B_\ell|\right)$, we conclude that $\mathbb{P}(T \geq \delta) \leq 2p^{-3}$ and complete the proof. □

Lemma 9. *Given a fixed vector* $\mathbf{B} \in \mathbb{R}^p$,

$$
\left|\sum_{\ell\neq 1} B_\ell \left\langle \widetilde{\mathbf{A}}_1, \widetilde{\mathbf{A}}_\ell \right\rangle\right| \lesssim \sqrt{\frac{\log p}{d}}\sqrt{\sum_{\ell\neq 1} B_\ell^2},
$$

holds with probability at least $1 - p^{-3} - e^{-0.8d}$.

Proof. Due to the independence between $\widetilde{\mathbf{A}}_1$ and $\widetilde{\mathbf{A}}_\ell$, where $\ell \neq 1$. We condition on $\widetilde{\mathbf{A}}_\ell$ and view $\sum_{\ell\neq 1} B_\ell \left\langle \widetilde{\mathbf{A}}_1, \widetilde{\mathbf{A}}_\ell \right\rangle$ as a Gaussian distributed RV $\mathsf{N}\left(0, d^{-1}\| \sum_{\ell\neq 1} B_\ell \widetilde{\mathbf{A}}_\ell\|_2^2\right)$. Then we obtain

$$
\begin{aligned}
&\mathbb{P}\left(\left|\sum_{\ell\neq 1} B_\ell \left\langle \widetilde{\mathbf{A}}_1, \widetilde{\mathbf{A}}_\ell \right\rangle\right| \gtrsim \sqrt{\frac{\log p}{d}}\sqrt{\sum_{\ell\neq 1} B_\ell^2}\right) \\
= &\underbrace{\mathbb{P}\left(\left|\sum_{\ell\neq 1} B_\ell \left\langle \widetilde{\mathbf{A}}_1, \widetilde{\mathbf{A}}_\ell \right\rangle\right| \gtrsim \sqrt{\frac{\log p}{d}}\sqrt{\sum_{\ell\neq 1} B_\ell^2}\right) \times \Psi(\mathcal{E}_4(\mathbf{B}))}_{T_1} \\
+ &\underbrace{\mathbb{P}\left(\left|\sum_{\ell\neq 1} B_\ell \left\langle \widetilde{\mathbf{A}}_1, \widetilde{\mathbf{A}}_\ell \right\rangle\right| \gtrsim \sqrt{\frac{\log p}{d}}\sqrt{\sum_{\ell\neq 1} B_\ell^2}\right) \times \Psi(\overline{\mathcal{E}}_3(\mathbf{B}))}_{T_2}.
\end{aligned}
$$

The proof is then completed by separately bounding T_1 and T_2. For term T_1, we have

$$
T_1 \overset{①}{\leq} \mathbb{E}_{\widetilde{\mathbf{A}}_\ell} \exp\left(-\frac{c\log p(\sum_{\ell\neq 1} B_\ell^2)}{\|\sum_{\ell\neq 1} B_\ell \widetilde{\mathbf{A}}_1\|_2^2}\right) \times \Psi(\mathcal{E}_4(\mathbf{B})) \overset{②}{\leq} \exp\left(-\frac{3\log p(\sum_{\ell\neq 1} B_\ell^2)}{\sum_{\ell\neq 1} B_\ell^2}\right),
$$

where ① is due to the tail bound for the Gaussian RV $\sum_{\ell\neq 1} B_\ell \left\langle \widetilde{\mathbf{A}}_1, \widetilde{\mathbf{A}}_\ell \right\rangle$ conditional on $\widetilde{\mathbf{A}}_\ell$, and ② is according to the definition of $\mathcal{E}_4(\mathbf{B})$. For term T_2, we have $T_2 \leq \Psi(\overline{\mathcal{E}}_3(\mathbf{B})) \leq e^{-0.8d}$. Summaring the above analysis finishes the proof.

□

Lemma 10. *Conditional on events $\mathcal{E}_1, \mathcal{E}_2$, we have*

$$
\left|\sum_{\ell\neq 1,2} B_\ell \left\langle \widetilde{\mathbf{A}}_1, \widetilde{\mathbf{A}}_\ell \right\rangle \left\langle \widetilde{\mathbf{A}}_2, \widetilde{\mathbf{A}}_\ell \right\rangle\right| \lesssim \frac{\left|\sum_{\ell>2} B_\ell\right|}{d}\sqrt{\frac{\log p}{d}} + \frac{\sqrt{\log p}}{d}\left[\sqrt{\sum_{\ell>2} B_\ell^2} \vee \left(\sqrt{\log p}(\max_\ell |B_\ell|)\right)\right],
$$

hold with probability at least $1 - 2p^{-3}$ for an arbitrary fixed vector $\mathbf{B} \in \mathbb{R}^p$.

Proof. First we rewrite the term $\sum_{\ell\neq 1,2} B_\ell \left\langle \widetilde{\mathbf{A}}_1, \widetilde{\mathbf{A}}_\ell \right\rangle \left\langle \widetilde{\mathbf{A}}_2, \widetilde{\mathbf{A}}_\ell \right\rangle$ as

$$
\sum_{\ell\neq 1,2} B_\ell \left\langle \widetilde{\mathbf{A}}_1, \widetilde{\mathbf{A}}_\ell \right\rangle \left\langle \widetilde{\mathbf{A}}_2, \widetilde{\mathbf{A}}_\ell \right\rangle = \sum_{\ell\neq 1,2} \widetilde{\mathbf{A}}_\ell^\top \left(B_\ell \widetilde{\mathbf{A}}_1 \widetilde{\mathbf{A}}_2^\top\right) \widetilde{\mathbf{A}}_\ell.
$$

Then we concatenate the vectors $\widetilde{\mathbf{A}}_\ell$, $\ell \neq 1, 2$ to a vector of length $d(p-2)$ and denote it as $\mathrm{vec}\left(\widetilde{\mathbf{A}}\right)$. Hence the summarization can be rewritten as

$$
\sum_{\ell\neq 1,2} \widetilde{\mathbf{A}}_\ell^\top \left(B_\ell \widetilde{\mathbf{A}}_1 \widetilde{\mathbf{A}}_2^\top\right) \widetilde{\mathbf{A}}_\ell = \mathrm{vec}\left(\widetilde{\mathbf{A}}\right)^\top \mathbf{\Lambda} \mathrm{vec}\left(\widetilde{\mathbf{A}}\right),
$$

where $\mathbf{\Lambda}$ is a block-diagonal matrix whose ith block is $B_{i+2}\widetilde{\mathbf{A}}_1\widetilde{\mathbf{A}}_2^\top$, $1 \leq i \leq p-2$. Due to the independence between $\widetilde{\mathbf{A}}_\ell$, $1 \leq \ell \leq p$, we first condition on $\widetilde{\mathbf{A}}_1, \widetilde{\mathbf{A}}_2$ and perform the following decomposition

$$
\left|\sum_{\ell\neq 1,2} B_\ell \left\langle \widetilde{\mathbf{A}}_1, \widetilde{\mathbf{A}}_\ell \right\rangle \left\langle \widetilde{\mathbf{A}}_2, \widetilde{\mathbf{A}}_\ell \right\rangle\right| \leq d^{-1}\Bigg(\underbrace{\left|\widetilde{\mathbf{A}}_1^\top \widetilde{\mathbf{A}}_2\right| \left|\sum_{\ell\neq 1,2} B_\ell\right|}_{\vartheta_1} + \underbrace{\left|d\mathrm{vec}\left(\widetilde{\mathbf{A}}\right)^\top \mathbf{\Lambda} \mathrm{vec}\left(\widetilde{\mathbf{A}}\right) - \sum_{\ell\neq 1,2} B_\ell \widetilde{\mathbf{A}}_1^\top \widetilde{\mathbf{A}}_2\right|}_{\vartheta_2}\Bigg).
$$

The upper-bound for ϑ_1 is relatively easy, which reads as

$$\vartheta_1 \lesssim \left|\sum_{\ell>2} B_\ell\right| \times \sqrt{\frac{\log p}{d}}.$$

The following analysis focus on bound ϑ_2. Since ϑ_2 also reads as

$$\vartheta_2 = \left| d\mathrm{vec}\left(\widetilde{\mathbf{A}}\right)^\top \mathbf{\Lambda} \mathrm{vec}\left(\widetilde{\mathbf{A}}\right) - \mathbb{E}_{\widetilde{\mathbf{A}}_\ell, \ell>2}\left(d\mathrm{vec}\left(\widetilde{\mathbf{A}}\right)^\top \mathbf{\Lambda} \mathrm{vec}\left(\widetilde{\mathbf{A}}\right)\right)\right|,$$

we can invoke the Hanson-Wright inequality (Theorem 6.2.1 in Vershynin (2016)) and obtain

$$\mathbb{P}\left(\vartheta_2 \geq \delta\right) \leq 2\exp\left(-c\left(\frac{\delta^2}{\|\mathbf{\Lambda}\|_{\mathrm{F}}^2} \wedge \frac{\delta}{\|\mathbf{\Lambda}\|_{\mathrm{OP}}}\right)\right),$$

where $\widetilde{\mathbf{A}}_1, \widetilde{\mathbf{A}}_2$ are viewed as constants. We complete the proof by setting δ as

$$\delta \asymp \left(\sqrt{\log p}\|\mathbf{\Lambda}\|_{\mathrm{F}}\right) \vee \left(\log p \|\mathbf{\Lambda}\|_{\mathrm{OP}}\right),$$

which yields $\mathbb{P}(\vartheta_2 \geq \delta) \leq 2p^{-3}$. The specific values of $\|\mathbf{\Lambda}\|_{\mathrm{F}}$ and $\|\mathbf{\Lambda}\|_{\mathrm{OP}}$ are computed as

$$\begin{aligned}
\|\mathbf{\Lambda}\|_{\mathrm{F}}^2 &= \left(\sum_{\ell>2} B_\ell^2\right) \left\|\widetilde{\mathbf{A}}_1 \widetilde{\mathbf{A}}_2^\top\right\|_{\mathrm{F}}^2 \overset{①}{=} \left(\sum_{\ell>2} B_\ell^2\right) \mathrm{Tr}\left(\widetilde{\mathbf{A}}_2 \widetilde{\mathbf{A}}_1^\top \widetilde{\mathbf{A}}_1 \widetilde{\mathbf{A}}_2^\top\right) \\
&\overset{②}{=} \left\|\widetilde{\mathbf{A}}_1\right\|_2^2 \left\|\widetilde{\mathbf{A}}_2\right\|_2^2 \left(\sum_{\ell>2} B_\ell^2\right) \overset{③}{\lesssim} \sum_{\ell>2} B_\ell^2 \\
\|\mathbf{\Lambda}\|_{\mathrm{OP}} &= \max_\ell |B_\ell| \left\|\widetilde{\mathbf{A}}_1 \widetilde{\mathbf{A}}_2^\top\right\|_{\mathrm{OP}} \overset{④}{=} \max_\ell |B_\ell| \times \left\|\widetilde{\mathbf{A}}_1\right\|_2 \left\|\widetilde{\mathbf{A}}_2\right\|_2 \lesssim \max_\ell |B_\ell|,
\end{aligned}$$

where in ① we use $\|\mathbf{M}\|_{\mathrm{F}}^2 = \mathrm{Tr}(\mathbf{M}^\top \mathbf{M})$ for arbitrary matrix $\mathbf{M}$, in ② we use $\mathrm{Tr}(\mathbf{M}_1\mathbf{M}_2) = \mathrm{Tr}(\mathbf{M}_2\mathbf{M}_1)$, in ③ we condition on event $\mathcal{E}_1$, and in ④ we use $\left\|\boldsymbol{u}\boldsymbol{v}^\top\right\|_{\mathrm{OP}} = \|\boldsymbol{u}\|_2 \|\boldsymbol{v}\|_2$ for arbitrary vectors $\boldsymbol{u}, \boldsymbol{v}$.

□

Lemma 11. *We have*

$$\left|\sum_{\substack{\ell_1,\ell_2\neq 1,2 \\ \ell_1\neq\ell_2}} B_{\ell_1,\ell_2} \left\langle \widetilde{\mathbf{A}}_1, \widetilde{\mathbf{A}}_{\ell_1}\right\rangle \left\langle \widetilde{\mathbf{A}}_2, \widetilde{\mathbf{A}}_{\ell_2}\right\rangle\right| \lesssim \frac{\log p}{d}\left(1 + c_0\sqrt{\frac{p}{d}}\right)^2 \sqrt{\sum_{\substack{i\neq 1,2 \\ j\neq 1,2,i}} B_{ij}^2},$$

holds with probability exceeding $1 - 4p^{-3} - e^{-c_0 p}$ *for a fixed matrix* $\mathbf{B}$.

Proof. We begin the proof by first rewriting $\sum_{\substack{\ell_1,\ell_2\neq 1,2 \\ \ell_1\neq\ell_2}} B_{\ell_1,\ell_2} \left\langle \widetilde{\mathbf{A}}_1, \widetilde{\mathbf{A}}_{\ell_1}\right\rangle \left\langle \widetilde{\mathbf{A}}_2, \widetilde{\mathbf{A}}_{\ell_2}\right\rangle$ as $\widetilde{\mathbf{A}}_1^\top \mathbf{\Lambda} \widetilde{\mathbf{A}}_2$, where $\mathbf{\Lambda}$ is defined as

$$\mathbf{\Lambda} \triangleq \sum_{\ell_1\neq 1,2} \sum_{\ell_2\neq 1,2,\ell_1} B_{\ell_1,\ell_2} \widetilde{\mathbf{A}}_{\ell_1} \widetilde{\mathbf{A}}_{\ell_2}^\top.$$

The whole proof procedure can be divided into the following stages.

Stage I. Due to the independence across $\widetilde{\mathbf{A}}_\ell$, we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\widetilde{\mathbf{A}}_1^\top \mathbf{\Lambda} \widetilde{\mathbf{A}}_2\right| \geq \delta\right) \leq \mathbb{P}\left(\left|\widetilde{\mathbf{A}}_1^\top \mathbf{\Lambda} \widetilde{\mathbf{A}}_2\right| \geq \delta,\ \left\|\mathbf{\Lambda}\widetilde{\mathbf{A}}_2\right\|_2 \leq \delta_1\right) + \mathbb{P}\left(\left\|\mathbf{\Lambda}\widetilde{\mathbf{A}}_2\right\|_2 \geq \delta_1\right) \\
&\overset{①}{\leq} 2\exp\left(-\frac{d\delta^2}{2\delta_1^2}\right) + \mathbb{P}\left(\left\|\mathbf{\Lambda}\widetilde{\mathbf{A}}_2\right\|_2 \geq \delta_1\right),
\end{aligned} \tag{24}$$

where in ① we first condition on $\widetilde{\mathbf{A}}_\ell$, $(\ell > 2)$ and view $\widetilde{\mathbf{A}}_1^\top \mathbf{\Lambda} \widetilde{\mathbf{A}}_2$ as a Gaussian RV satisfying $\mathsf{N}\left(0, d^{-1}\|\mathbf{\Lambda}\widetilde{\mathbf{A}}_2\|_2^2\right)$.

Stage II. To bound the probability $\mathbb{P}\left(\left\|\mathbf{\Lambda}\widetilde{\mathbf{A}}_2\right\|_2 \geq \delta_1\right)$, we need to upper bound the Frobenius norm $\|\mathbf{\Lambda}\|_{\mathrm{F}}$. First we define two matrices, namely, $\widetilde{\widetilde{\mathbf{A}}}$ and $\widetilde{\mathbf{B}}$, for the conciseness of notation, which reads

$$\widetilde{\widetilde{\mathbf{A}}} \triangleq \left[\widetilde{\mathbf{A}}_3 \ \cdots \ \widetilde{\mathbf{A}}_p\right], \quad (\widetilde{\mathbf{B}})_{i,j} \triangleq \begin{cases} B_{i+2,j+2}, & \text{if } \ i \neq j; \\ 0, & \text{otherwise.} \end{cases}$$

Easily we can verify $\mathbf{\Lambda}$ is equivalent to $\mathbf{\Lambda}$ as $\widetilde{\widetilde{\mathbf{A}}}\widetilde{\mathbf{B}}\widetilde{\widetilde{\mathbf{A}}}^\top$, which gives

$$\|\mathbf{\Lambda}\|_{\mathrm{F}} = \left\|\widetilde{\widetilde{\mathbf{A}}}\widetilde{\mathbf{B}}\widetilde{\mathbf{A}}^\top\right\|_{\mathrm{F}} \overset{②}{\leq} \left\|\widetilde{\widetilde{\mathbf{A}}}\right\|_{\mathrm{OP}}^2 \left\|\widetilde{\mathbf{B}}\right\|_{\mathrm{F}} \overset{③}{\leq} \left\|\widetilde{\mathbf{A}}\right\|_{\mathrm{OP}}^2 \left\|\widetilde{\mathbf{B}}\right\|_{\mathrm{F}},$$

where in ② we adopt the relation $\|\mathbf{M}_1\mathbf{M}_2\|_{\mathrm{F}} \leq \|\mathbf{M}_1\|_{\mathrm{OP}}\|\mathbf{M}_2\|_{\mathrm{F}}$ such that $\mathbf{M}_1, \mathbf{M}_2$ are arbitrary matrix, and in ③ we use the relation $\left\|\widetilde{\widetilde{\mathbf{A}}}\right\|_{\mathrm{OP}} \leq \left\|\widetilde{\mathbf{A}}\right\|_{\mathrm{OP}}$ in Corol. 2.4.2 Golub & Van Loan (2012) since $\widetilde{\widetilde{\mathbf{A}}}$ can be viewed as a sub-matrix of $\widetilde{\mathbf{A}}$.

Hence we conclude that

$$\begin{aligned}
&\mathbb{P}\left(\|\mathbf{\Lambda}\|_{\mathrm{F}} \geq \left(1 + c_0\sqrt{\frac{p}{d}}\right)^2 \left\|\widetilde{\mathbf{B}}\right\|_{\mathrm{F}}\right) \\
&= \mathbb{P}\left(\|\mathbf{\Lambda}\|_{\mathrm{F}} \gtrsim \frac{p}{d}\left\|\widetilde{\mathbf{B}}\right\|_{\mathrm{F}}, \ \left\|\widetilde{\mathbf{A}}\right\|_{\mathrm{OP}} \leq \sqrt{\frac{c_0 p}{d}}\right) + \mathbb{P}\left(\|\mathbf{\Lambda}\|_{\mathrm{F}} \gtrsim \frac{p}{d}\left\|\widetilde{\mathbf{B}}\right\|_{\mathrm{F}}, \ \left\|\widetilde{\mathbf{A}}\right\|_{\mathrm{OP}} > \sqrt{\frac{c_0 p}{d}}\right) \\
&\leq \underbrace{\mathbb{P}\left(\left\|\widetilde{\mathbf{A}}\right\|_{\mathrm{OP}}^2 \gtrsim \frac{p}{d}, \left\|\widetilde{\mathbf{A}}\right\|_{\mathrm{OP}} \leq \sqrt{\frac{c_0 p}{d}}\right)}_{0} + \mathbb{P}\left(\left\|\widetilde{\mathbf{A}}\right\|_{\mathrm{OP}} \geq \sqrt{\frac{c_0 p}{d}}\right) \overset{④}{\leq} e^{-c_0 p},
\end{aligned}$$

where ④ is due to Thm. 6.1 in Wainwright (2019).

Stage III. We bound $\mathbb{P}\left(\left\|\mathbf{\Lambda}\widetilde{\mathbf{A}}_2\right\|_2 \geq \delta_1\right)$ by splitting it as

$$\begin{aligned}
\mathbb{P}\left(\left\|\mathbf{\Lambda}\widetilde{\mathbf{A}}_2\right\|_2 \geq \delta_1\right) = &\underbrace{\mathbb{P}\left(\left\|\mathbf{\Lambda}\widetilde{\mathbf{A}}_2\right\|_2 \geq \delta_1, \ \|\mathbf{\Lambda}\|_{\mathrm{F}} \ \geq \left(1 + c_0\sqrt{\frac{p}{d}}\right)^2 \left\|\widetilde{\mathbf{B}}\right\|_{\mathrm{F}}\right)}_{\leq \mathbb{P}\left(\|\mathbf{\Lambda}\|_{\mathrm{F}} \ \geq \left(1+c_0\sqrt{\frac{p}{d}}\right)^2 \left\|\widetilde{\mathbf{B}}\right\|_{\mathrm{F}}\right)} \\
&+ \mathbb{P}\left(\left\|\mathbf{\Lambda}\widetilde{\mathbf{A}}_2\right\|_2 \geq \delta_1, \ \|\mathbf{\Lambda}\|_{\mathrm{F}} \ < \left(1 + c_0\sqrt{\frac{p}{d}}\right)^2 \left\|\widetilde{\mathbf{B}}\right\|_{\mathrm{F}}\right).
\end{aligned} \tag{25}$$

Notice that the first term is bounded in Stage II, we focus on bounding the second term in this stage, which proceeds as

$$
\begin{aligned}
&\mathbb{P}\left(\left\|\mathbf{\Lambda}\widetilde{\mathbf{A}}_2\right\|_2 \geq \delta_1,\ \|\mathbf{\Lambda}\|_{\mathrm{F}} < \left(1+c_0\sqrt{\frac{p}{d}}\right)^2 \left\|\widetilde{\mathbf{B}}\right\|_{\mathrm{F}}\right) \\
\overset{⑤}{\leq}\ & \mathbb{E}\mathbb{1}\left(\left|d\left\|\mathbf{\Lambda}\widetilde{\mathbf{A}}_2\right\|_2^2 - \|\mathbf{\Lambda}\|_{\mathrm{F}}^2\right| \geq \delta_2,\ \|\mathbf{\Lambda}\|_{\mathrm{F}} < \left(1+c_0\sqrt{\frac{p}{d}}\right)^2 \left\|\widetilde{\mathbf{B}}\right\|_{\mathrm{F}}\right) \\
\overset{⑥}{\leq}\ & 2\exp\left(-c_0\left(\frac{\delta_2}{\|\mathbf{\Lambda}^\top\mathbf{\Lambda}\|_{\mathrm{OP}}} \wedge \frac{\delta_2^2}{\|\mathbf{\Lambda}^\top\mathbf{\Lambda}\|_{\mathrm{F}}^2}\right)\right) \times \mathbb{1}\left(\|\mathbf{\Lambda}\|_{\mathrm{F}} < \left(1+c_0\sqrt{\frac{p}{d}}\right)^2 \left\|\widetilde{\mathbf{B}}\right\|_{\mathrm{F}}\right) \\
\leq\ & 2\exp\left(-c_0\left(\frac{\delta_2}{\|\mathbf{\Lambda}^\top\mathbf{\Lambda}\|_{\mathrm{F}}} \wedge \frac{\delta_2^2}{\|\mathbf{\Lambda}^\top\mathbf{\Lambda}\|_{\mathrm{F}}^2}\right)\right) \times \mathbb{1}\left(\|\mathbf{\Lambda}\|_{\mathrm{F}} < \left(1+c_0\sqrt{\frac{p}{d}}\right)^2 \left\|\widetilde{\mathbf{B}}\right\|_{\mathrm{F}}\right) \\
\overset{⑦}{\leq}\ & 2p^{-3},
\end{aligned}
\tag{26}
$$

where in ⑤ we require $d\delta_1^2 \geq \|\mathbf{\Lambda}\|_{\mathrm{F}}^2 + \delta_2$, ⑥ is due to the Hanson-Wright inequality (cf. Thm. 6.2.1 in Vershynin (2016)), and in ⑦ we set δ_2 as $(1+c_0\sqrt{p/d})^4 \log p \left\|\widetilde{\mathbf{B}}\right\|_{\mathrm{F}}^2$.

Combining equation 24, equation 25, and equation 26, we set δ_1, δ as $c\sqrt{\log p/d}(1+c_0\sqrt{p/d})^2\left\|\widetilde{\mathbf{B}}\right\|_{\mathrm{F}}$ and $c_3\delta \asymp \sqrt{\log p/d}\delta_1$, respectively, which yields

$$
\mathbb{P}\left(\left|\widetilde{\mathbf{A}}_1^\top\mathbf{\Lambda}\widetilde{\mathbf{A}}_2\right| \gtrsim \frac{\log p}{d}\left(1+c_0\sqrt{\frac{p}{d}}\right)^2 \left\|\widetilde{\mathbf{B}}\right\|_{\mathrm{F}}\right) \leq 4p^{-3} + e^{-c_0 p},
$$

and completes the proof. □

Lemma 12. *Conditional on the event $\mathcal{E}_1$, we have*

$$
\mathbb{P}\left[n^{-1}\left|\widetilde{\mathbf{A}}_i^\top\left(\sum_{\ell=1}^n \mathbf{W}^{(i)}\mathbf{W}^{(i)\top}\right)\widetilde{\mathbf{A}}_j\right| \geq c_0\sigma^2\left(1+c_1\sqrt{\frac{\log p}{d}}\right)\sqrt{\frac{\log p}{d}}\left(1+c_2\left(\sqrt{\frac{d}{n}} \vee \frac{d}{n}\right)\right)\right] \leq p^{-4} + c_3 e^{-c_4 d}
$$

with fixed i, j.

Proof. First we define the matrix $\widehat{\mathbf{\Xi}}_n$ as $\frac{1}{n}\sum_{\ell=1}^n \mathbf{W}^{(i)}\mathbf{W}^{(i)\top}$ for the conciseness of notation. Then we upper-bound $\widetilde{\mathbf{A}}_i\widehat{\mathbf{\Xi}}_n\widetilde{\mathbf{A}}_j$, $i \neq j$ as

$$
\begin{aligned}
&\mathbb{P}\left(\left|\widetilde{\mathbf{A}}_i^\top\widehat{\mathbf{\Xi}}_n\widetilde{\mathbf{A}}_j\right| \geq \delta\right) \overset{①}{\leq} \mathbb{P}\left(\|\widehat{\mathbf{\Xi}}_n\widetilde{\mathbf{A}}_j\|_2 \geq \delta_1\right) + \exp\left(-\frac{d\delta^2}{2\delta_1^2}\right) \\
\overset{②}{\leq}\ & \mathbb{P}\left(\left\|\widehat{\mathbf{\Xi}}_n\right\|_{\mathrm{OP}}\left\|\widetilde{\mathbf{A}}_j\right\|_2 \geq \delta_1\right) + p^{-4} \overset{③}{\leq} \mathbb{P}\left(\left\|\widehat{\mathbf{\Xi}}_n\right\|_{\mathrm{OP}} \geq \delta_2\right) + p^{-4} \\
\overset{④}{\leq}\ & \mathbb{P}\left(\left\|\widehat{\mathbf{\Xi}}_n - \sigma^2\mathbf{I}\right\|_{\mathrm{OP}} \geq c_2\sigma^2\left(\frac{d}{n} \vee \sqrt{\frac{d}{n}}\right)\right) + p^{-4} \overset{⑤}{\leq} p^{-4} + c_0 e^{-c_1 d},
\end{aligned}
$$

where in ① we exploit the independence between $\widetilde{\mathbf{A}}_i$ and $\widetilde{\mathbf{A}}_j$ when $i \neq j$ and treat $\widetilde{\mathbf{A}}_i^\top\widehat{\mathbf{\Xi}}_n\widetilde{\mathbf{A}}_j$ as a Gaussian RV with mean zero and variance $d^{-1}\|\widehat{\mathbf{\Xi}}_n\widetilde{\mathbf{A}}_j\|_2^2$, in ② we use the fact $\delta = 2\delta_1\sqrt{2\log p/d}$, and in ③ we condition on event $\mathcal{E}_1$ and set $\delta_1 \geq \delta_2\left(1+c_0\sqrt{\log p/d}\right)$, ④ is because $\delta_2 = \sigma^2\left(1+c_2\left(d/n \vee \sqrt{d/n}\right)\right)$, and in ⑤ we use Thm. 6.5 in Wainwright (2019). The proof is then completed by setting δ as

$$
\delta = c_0\sigma^2\left(1+c_1\sqrt{\frac{\log p}{d}}\right)\sqrt{\frac{\log p}{d}}\left(1+c_2\left(\sqrt{\frac{d}{n}} \vee \frac{d}{n}\right)\right).
$$

□

A.3 Insight Behind the Design of the Covariance Matrix Estimator

We now explain the rational behind the covariance matrix estimators of $\mathbf{X}$ that we use in the graphical structure estimation via the parametric method. For this purpose, we exploit the statistical properties of $\widetilde{\mathbf{A}}$. We approximate data samples, $\widehat{\mathbf{X}}^{(i)}$'s, as $\widehat{\mathbf{X}}^{(i)} = \widetilde{\mathbf{A}}^\top \widetilde{\mathbf{A}} \mathbf{X}^{(i)} = \widetilde{\mathbf{A}}^\top \mathbf{Y}^{(i)}$, $1 \le i \le n$. Due to the assumption on sensing matrix $\widetilde{\mathbf{A}}$, we have $\mathbb{E}_{\widetilde{\mathbf{A}}} \widehat{\mathbf{X}}^{(i)} = \mathbf{X}^{(i)}$. Hence, we can view the samples $\{\widehat{\mathbf{X}}^{(i)}\}_{i=1}^n$ as a "perturbed" version of the true data points $\left\{\mathbf{X}^{(i)}\right\}_{i=1}^n$. Hence, we propose to estimate the covariance matrix of $\mathbf{X}$ from $\{\widehat{\mathbf{X}}^{(i)}\}_{i=1}^n$. Following the above approach, a naive covariance estimator is given as

$$\widehat{\mathbf{\Sigma}}_{n,1} = \frac{1}{n} \sum_{i=1}^n \widehat{\mathbf{X}}^{(i)} \widehat{\mathbf{X}}^{(i)\top} = \widetilde{\mathbf{A}}^\top \left(\frac{1}{n} \sum_{i=1}^n \mathbf{Y}^{(i)} \mathbf{Y}^{(i)\top} \right) \widetilde{\mathbf{A}}.$$

However, through numerical experiments, we observed that this estimator performs poorly. To improve the performance of $\widehat{\mathbf{\Sigma}}_{n,1}$, we first analyze its properties thoroughly and then refine the estimator.

A.3.1 Theoretical Properties

First, we evaluate the mean and variance of the naive covariance estimator $\widehat{\mathbf{\Sigma}}_{n,1}$.

Lemma 13. *The mean of the naive covariance estimator is given by*

$$\mathbb{E}_{\widetilde{\mathbf{A}},\mathbf{X}}[\widehat{\mathbf{\Sigma}}_{n,1}] = \frac{d+1}{d} \mathbf{\Sigma}^\natural + \frac{p}{d} \mathbf{I} + \sigma^2 \mathbf{I}.$$

Proof. Due to the independence between $\mathbf{X}$ and $\widetilde{\mathbf{A}}$, we first condition on $\widetilde{\mathbf{A}}$ and take expectation w.r.t $\mathbf{X}, \mathbf{W}$, which gives

$$\mathbb{E}_{\mathbf{X},\mathbf{W}} \widehat{\mathbf{\Sigma}}_{n,1} = \widetilde{\mathbf{A}}^\top \widetilde{\mathbf{A}} \mathbf{\Sigma}^\natural \widetilde{\mathbf{A}}^\top \widetilde{\mathbf{A}} + \sigma^2 \widetilde{\mathbf{A}}^\top \widetilde{\mathbf{A}}.$$

Then we conclude that

$$\begin{aligned}
&\mathbb{E}_{\widetilde{\mathbf{A}}} \left(\widetilde{\mathbf{A}}^\top \widetilde{\mathbf{A}} \mathbf{\Sigma}^\natural \widetilde{\mathbf{A}}^\top \widetilde{\mathbf{A}} \right)_{ij} = \mathbb{E}_{\widetilde{\mathbf{A}}} \sum_{\ell_1,\ell_2} \mathbf{\Sigma}^\natural_{\ell_1,\ell_2} \left\langle \widetilde{\mathbf{A}}_{\ell_1}, \widetilde{\mathbf{A}}_i \right\rangle \left\langle \widetilde{\mathbf{A}}_{\ell_2}, \widetilde{\mathbf{A}}_j \right\rangle + \sigma^2 \mathbb{1}(i=j) \\
&= \mathbb{E}_{\widetilde{\mathbf{A}}} \sum_{\ell_1,\ell_2} \mathbf{\Sigma}^\natural_{\ell_1,\ell_2} \left(\sum_{\ell_3} \widetilde{\mathbf{A}}_{\ell_3,\ell_1} \widetilde{\mathbf{A}}_{\ell_3,i} \right) \cdot \left(\sum_{\ell_4} \widetilde{\mathbf{A}}_{\ell_4,\ell_2} \widetilde{\mathbf{A}}_{\ell_4,j} \right) + \sigma^2 \mathbb{1}(i=j) \\
&= \sum_{\ell_1,\ell_2,\ell_3,\ell_4} \mathbf{\Sigma}^\natural_{\ell_1,\ell_2} \, \mathbb{E}_{\widetilde{\mathbf{A}}} \left(\widetilde{\mathbf{A}}_{\ell_3,\ell_1} \widetilde{\mathbf{A}}_{\ell_3,i} \widetilde{\mathbf{A}}_{\ell_4,\ell_2} \widetilde{\mathbf{A}}_{\ell_4,j} \right) + \sigma^2 \mathbf{I} \\
&\overset{①}{=} d^{-2} \Bigg[\sum_{\ell_1,\ell_2,\ell_3,\ell_4} \mathbf{\Sigma}^\natural_{\ell_1,\ell_2} \Big(\mathbb{1}(\ell_1 = i) \, \mathbb{1}(\ell_2 = j) + \mathbb{1}(\ell_3 = \ell_4) \, \mathbb{1}(\ell_1 = \ell_2) \mathbb{1}(i=j) \\
&\qquad + \mathbb{1}(\ell_3 = \ell_4) \mathbb{1}(\ell_1 = j) \, \mathbb{1}(\ell_2 = i) \Big) \Bigg] + \sigma^2 \mathbb{1}(i=j) \\
&= d^{-2} \left[\sum_{\ell_3,\ell_4} \mathbf{\Sigma}^\natural_{i,j} + \mathbb{1}(i=j) \left(\sum_{\ell_1} \sum_{\ell_2} \mathbf{\Sigma}^\natural_{\ell_1,\ell_1} \right) + \mathbf{\Sigma}^\natural_{j,i} \sum_{\ell_3,\ell_4} \mathbb{1}(\ell_3 = \ell_4) \right] + \sigma^2 \mathbb{1}(i=j) \\
&= d^{-2} \left[d^2 \mathbf{\Sigma}^\natural_{i,j} + \mathbb{1}(i=j) d \sum_{\ell} \mathbf{\Sigma}^\natural_{\ell,\ell} + d \mathbf{\Sigma}^\natural_{i,j} \right] + \sigma^2 \mathbb{1}(i=j) \\
&= \left(1 + d^{-1}\right) \mathbf{\Sigma}^\natural_{i,j} + d^{-1} \left(\sum_{\ell} \mathbf{\Sigma}^\natural_{\ell,\ell} \right) \mathbb{1}(i=j) + \sigma^2 \mathbb{1}(i=j),
\end{aligned}$$

which completes the proof with the fact $\mathbf{\Sigma}^\natural_{\ell,\ell} = 1$. In ①, we use the Wick's theorem, which is listed as Thm. 10 for the sake of self-containing. □

Then we study the variance $\mathrm{Var}_{\widetilde{\mathbf{A}}}\widehat{\mathbf{\Sigma}}_{n,1}$, which is listed as the following. Due to the complex formula of the variance, we only consider the noiseless case, namely, $\sigma^2 = 0$.

Lemma 14. *Consider the noiseless case where $\sigma^2 = 0$, we have* $\mathrm{Var}_{\widetilde{\mathbf{A}}}(\widehat{\mathbf{\Sigma}}_{n,1})_{i,j} = \Omega(d^{-1}) + \Omega(d^{-1})\mathbb{1}(i = j) + \Omega(d^{-2})$.

The detailed proof is given in the appendix. We use results of Lemma 13 to improve our naive covariance estimator in two perspectives: **bias correction** and **variance reduction**.

Proof. With the relation $\mathrm{Var}_{\widetilde{\mathbf{A}}}\widehat{\mathbf{\Sigma}}_{n,1} = \mathbb{E}_{\widetilde{\mathbf{A}}}(\widehat{\mathbf{\Sigma}}_{n,1})^2_{i,j} - \left(\mathbb{E}\widehat{\mathbf{\Sigma}}_{n,1}\right)^2_{i,j}$, we complete the proof by invoking Lemma 13 and Lemma 15. The following context focuses on proving Lemma 15. □

Lemma 15. *We have*

$$
\begin{aligned}
&\mathbb{E}_{\widetilde{\mathbf{A}}}(\widehat{\mathbf{\Sigma}}_{n,1})^2_{i,j} = (\mathbf{\Sigma}_n)^2_{i,j} + d^{-1}\left[\|(\mathbf{\Sigma}_n)_i\|_2^2 + \|(\mathbf{\Sigma}_n)_j\|_2^2 + 4(\mathbf{\Sigma}_n)^2_{i,j} + 2(\mathbf{\Sigma}_n)_{i,i}(\mathbf{\Sigma}_n)_{j,j}\right] \\
&+ d^{-2}\left[2\left((\mathbf{\Sigma}_n)_{i,i} + (\mathbf{\Sigma}_n)_{j,j}\right)\mathrm{Tr}(\mathbf{\Sigma}_n) + 4(\mathbf{\Sigma}_n)_{i,i}(\mathbf{\Sigma}_n)_{j,j} + 2\|(\mathbf{\Sigma}_n)_j\|_2^2 + 2\|(\mathbf{\Sigma}_n)_i\|_2^2 + \|\mathbf{\Sigma}_n\|_{\mathrm{F}}^2 + 7(\mathbf{\Sigma}_n)^2_{i,j}\right] \\
&+ d^{-3}\left[3\|(\mathbf{\Sigma}_n)_i\|_2^2 + 3\|(\mathbf{\Sigma}_n)_j\|_2^2 + 2\left((\mathbf{\Sigma}_n)_{i,i} + (\mathbf{\Sigma}_n)_{j,j}\right)\mathrm{Tr}(\mathbf{\Sigma}_n) + \mathrm{Tr}^2(\mathbf{\Sigma}_n) + 2(\mathbf{\Sigma}_n)_{i,i}(\mathbf{\Sigma}_n)_{j,j} + 4(\mathbf{\Sigma}_n)^2_{i,j}\right] \\
&+ d^{-1}\mathbb{1}(i=j)\left[\langle(\mathbf{\Sigma}_n)_i, (\mathbf{\Sigma}_n)_j\rangle + 2(\mathbf{\Sigma}_n)_{i,j}\,\mathrm{Tr}(\mathbf{\Sigma}_n)\right] \\
&+ d^{-2}\mathbb{1}(i=j)\left[11\,\langle(\mathbf{\Sigma}_n)_i, (\mathbf{\Sigma}_n)_j\rangle + \mathrm{Tr}^2(\mathbf{\Sigma}_n) + 6(\mathbf{\Sigma}_n)_{i,j}\,\mathrm{Tr}(\mathbf{\Sigma}_n)\right] \\
&+ d^{-3}\mathbb{1}(i=j)\left[12\,\langle(\mathbf{\Sigma}_n)_i, (\mathbf{\Sigma}_n)_j\rangle + 8(\mathbf{\Sigma}_n)_{i,j}\,\mathrm{Tr}(\mathbf{\Sigma}_n) + \mathrm{Tr}^2(\mathbf{\Sigma}_n) + 2\|\mathbf{\Sigma}_n\|_{\mathrm{F}}^2\right].
\end{aligned}
$$

Proof. The proof procedure is fundamentally the same as the steps in computing $\mathbb{E}_{\mathbf{X}}\widehat{\mathbf{\Sigma}}_n^{\mathrm{param}}$ but is more involved (requires computing over 100 terms).

We begin the proof with the following expansion

$$
\begin{aligned}
&\mathbb{E}\left[\left(\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}(\mathbf{\Sigma}_n)\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\right)_{ij}\right]^2 \\
=&\sum_{\ell_1,\ell_2,\ell_3,\ell_4,\ell_5,\ell_6,\ell_7,\ell_8}(\mathbf{\Sigma}_n)_{\ell_1,\ell_2}(\mathbf{\Sigma}_n)_{\ell_5,\ell_6}\,\mathbb{E}\left(\widetilde{\mathbf{A}}_{\ell_3,\ell_1}\widetilde{\mathbf{A}}_{\ell_3,i}\,\widetilde{\mathbf{A}}_{\ell_4,\ell_2}\widetilde{\mathbf{A}}_{\ell_4,j}\,\widetilde{\mathbf{A}}_{\ell_7,\ell_5}\widetilde{\mathbf{A}}_{\ell_7,i}\,\widetilde{\mathbf{A}}_{\ell_8,\ell_6}\widetilde{\mathbf{A}}_{\ell_8,j}\right).
\end{aligned}
$$

Then we expand the term $\mathbb{E}\left(\widetilde{\mathbf{A}}_{\ell_3,\ell_1}\widetilde{\mathbf{A}}_{\ell_3,i}\,\widetilde{\mathbf{A}}_{\ell_4,\ell_2}\widetilde{\mathbf{A}}_{\ell_4,j}\,\widetilde{\mathbf{A}}_{\ell_7,\ell_5}\widetilde{\mathbf{A}}_{\ell_7,i}\,\widetilde{\mathbf{A}}_{\ell_8,\ell_6}\widetilde{\mathbf{A}}_{\ell_8,j}\right)$ via the Wick's Theorem, which is also listed as Thm. 10 for the sake of self-containing. Additionally, we need to divide d^4 for the following result.

□

A.3.2 Estimator Refinement

Bias Correction. First we note that the naive estimator is biased $\mathbb{E}_{\widetilde{\mathbf{A}},\mathbf{X}}\widehat{\mathbf{\Sigma}}_{n,1} \neq \mathbf{\Sigma}^\natural$. Adopting ideas similar to the moment estimator, we correct the bias of the estimator via

$$
\widehat{\mathbf{\Sigma}}_{n,2} \triangleq \frac{d}{d+1}\widetilde{\mathbf{A}}^\top\left(\frac{1}{n}\sum_{i=1}^n \mathbf{Y}^{(i)}\mathbf{Y}^{(i)\top}\right)\widetilde{\mathbf{A}} - \frac{p + d\sigma^2}{d+1}\mathbf{I}.
$$

It can be easily verified that $\mathbb{E}_{\widetilde{\mathbf{A}},\mathbf{X}}[\widehat{\mathbf{\Sigma}}_{n,2}] = \mathbf{\Sigma}^\natural$.

Variance Reduction. To further improve the performance of the covariance estimator, we perform variance reduction. From Lemma 14, we observe that the diagonal entries $(\widehat{\mathbf{\Sigma}}_{n,2})_{i,i}$ have a higher variance than non-diagonal entries. On the other hand, there is no need for estimating $\mathrm{diag}(\mathbf{\Sigma}^{\natural})$ since by the assumptions, we know that the diagonal elements of the covariance matrix are 1. Therefore, we suggest refining the estimator $\widehat{\mathbf{\Sigma}}_{n,2}$ by fixing its diagonal entries to 1, i.e., the resulting refined covariance estimator is given as

$$\widehat{\mathbf{\Sigma}}_n^{\text{param}} = \mathbf{I} + \frac{d}{d+1}\left[\widetilde{\mathbf{A}}^\top\left(\frac{1}{n}\sum_{i=1}^n \mathbf{Y}^{(i)}\mathbf{Y}^{(i)\top}\right)\widetilde{\mathbf{A}}\right]_{\text{off}}, \tag{27}$$

where $(\cdot)_{\text{off}}$ denotes the operation of picking non-diagonal entries. It can be easily verified that the estimator $\widehat{\mathbf{\Sigma}}_n^{\text{param}}$ in (27) is unbiased.

B Proof of Thm. 1

The analysis is based on **primal-dual method**, which is adapted from Ravikumar et al. (2011). First we write the optimality condition for equation 4 as

$$\widehat{\mathbf{\Sigma}}_n^{\text{param}} - \widehat{\mathbf{\Theta}}_{\text{param}}^{-1} + \lambda_{\text{param}}\mathbf{G} = \mathbf{0}, \tag{28}$$

where $\mathbf{G}$ is the sub-gradient Rockafellar (1970) of $\|\widehat{\mathbf{\Theta}}_{\text{param}}\|_{1,\text{off}}$ and is defined as

$$G_{ij} \triangleq \begin{cases} \mathrm{sgn}(\widehat{\mathbf{\Theta}}_{\text{param}})_{i,j}, & \text{if } (\widehat{\mathbf{\Theta}}_{\text{param}})_{i,j} \neq 0; \\ \in [-1,1], & \text{otherwise.} \end{cases}$$

for $i \neq j$. However, because of the complexity of (28), directly bounding the deviation $\|\widehat{\mathbf{\Theta}}_{\text{param}} - \mathbf{\Theta}^{\natural}\|_\infty$ can be difficult. Instead, we construct a pair $\left(\widetilde{\mathbf{\Theta}}_{\text{param}}, \widetilde{\mathbf{G}}\right)$ which satisfies: (i) $\widetilde{\mathbf{G}}$ is the sub-differential of $\|\widetilde{\mathbf{\Theta}}_{\text{param}}\|_{1,\text{off}}$; and (ii) the pair $(\widetilde{\mathbf{\Theta}}_{\text{param}}, \widetilde{\mathbf{G}})$ satisfies the condition in equation 28. Then we show it coincides with the solution of (5). The basic rational is as follows. First we verify that $\widehat{\mathbf{\Theta}}_{\text{param}}$ is the unique solution of (28). Since our constructed pairs $(\widetilde{\mathbf{\Theta}}_{\text{param}}, \widetilde{\mathbf{G}})$ satisfies the condition in equation 28, which corresponds to the solution in (5) exclusively, we can hence show the constructed pair $(\widetilde{\mathbf{\Theta}}_{\text{param}}, \widetilde{\mathbf{G}})$ is the identical solution of (5), namely, $(\widehat{\mathbf{\Theta}}_{\text{param}}, \mathbf{G})$. Afterwards we can upper bound $\|\widehat{\mathbf{\Theta}}_{\text{param}} - \mathbf{\Theta}^{\natural}\|_\infty$ by investigating $\|\widetilde{\mathbf{\Theta}}_{\text{param}} - \mathbf{\Theta}^{\natural}\|_\infty$, which is more amenable for the analysis as following.

Stage I: construct $\widetilde{\mathbf{\Theta}}_{\text{param}}$. We construct the primal-dual witness solution $\left(\widetilde{\mathbf{\Theta}}_{\text{param}}, \widetilde{\mathbf{G}}\right)$ assuming the support set S is given as a prior. This step can further be divided into three stages, as illustrated in the main context. First we construct the matrix $\widetilde{\mathbf{\Theta}}_{\text{param}}$. For the entries (i,j) restricted to the set S, we set them as

$$(\widetilde{\mathbf{\Theta}}_{\text{param}})_S = \text{argmin}_{\substack{\mathbf{\Theta}\succ\mathbf{0} \\ \mathbf{\Theta}=\mathbf{\Theta}^\top, \mathbf{\Theta}_{S^c}=\mathbf{0}}} -\log\det\mathbf{\Theta} + \left\langle \widehat{\mathbf{\Sigma}}_n^{\text{param}}, \mathbf{\Theta}\right\rangle + \lambda_{\text{param}}\|\mathbf{\Theta}\|_{1,\text{off}}.$$

For the rest of entries $(\widetilde{\mathbf{\Theta}}_{\text{param}})_{S^c}$, we set them to be zero values. Then we construct the sub-differential $\widetilde{\mathbf{G}}$ corresponding to the matrix $\widetilde{\mathbf{\Theta}}_{\text{param}}$. For the entry $(i,j) \in S$, we set $\widetilde{G}_{i,j} = \mathrm{sign}(\widetilde{\mathbf{\Theta}}_{\text{param}})_{i,j}$. For the entry (i,j) that is outside of the support set S, we set $\widetilde{G}_{i,j}$ as

$$\widetilde{G}_{i,j} = \lambda_{\text{param}}^{-1}\left[\left(\widetilde{\mathbf{\Theta}}_{\text{param}}\right)_{i,j}^{-1} - (\widehat{\mathbf{\Sigma}}_n^{\text{param}})_{i,j}\right].$$

The goal of this step is to ensure that $\left(\widetilde{\mathbf{\Theta}}_{\text{param}}, \widetilde{\mathbf{G}}\right)$ satisfies (28). As illustrated in the main context, we have the pair $\left(\widetilde{\mathbf{\Theta}}_{\text{param}}, \widetilde{\mathbf{G}}\right)$ coincides with the solution of equation 28 once $\left|\widetilde{G}_{i,j}\right| < 1$ for the entry $(i,j) \in S^c$. The following step focuses on showing $\left|\widetilde{G}_{i,j}\right| < 1$.

Stage II: construct $\widetilde{\mathbf{G}}$. For the entry $(i,j) \in S$, we set $\widetilde{G}_{i,j} = \mathrm{sign}(\widetilde{\mathbf{\Theta}}_{\mathrm{param}})_{i,j}$. For the entry (i,j) that is outside of the support set S, we set $\widetilde{G}_{i,j}$ as

$$\widetilde{G}_{i,j} = \lambda_{\mathrm{param}}^{-1} \left[\left(\widetilde{\mathbf{\Theta}}_{\mathrm{param}} \right)_{i,j}^{-1} - (\widehat{\mathbf{\Sigma}}_n^{\mathrm{param}})_{i,j} \right].$$

The goal of this step is to ensure that $\left(\widetilde{\mathbf{\Theta}}_{\mathrm{param}}, \widetilde{\mathbf{G}} \right)$ satisfies (28).

Stage III: verify $\widetilde{\mathbf{G}}$ to be the sub-differential of $\|\widetilde{\mathbf{\Theta}}_{\mathbf{param}}\|_{1,\mathbf{off}}$. In the following analysis, we verify that $|\widetilde{G}_{i,j}| < 1$, which yields the upper-bound on $\|\widetilde{\mathbf{\Theta}}_{\mathrm{param}} - \mathbf{\Theta}^{\natural}\|_\infty$ as a byproduct. We first need the necessary lemmas from Ravikumar et al. (2011).

Lemma 16 (Lemma 6 in Ravikumar et al. (2011))**.** *Suppose that* $r \triangleq 2\kappa_{\mathbf{\Gamma}} \left(\left\| \widehat{\mathbf{\Sigma}}_n^{\mathrm{param}} - \mathbf{\Sigma}^{\natural} \right\|_\infty + \lambda_{\mathrm{param}} \right) \leq \frac{1 \wedge \left(\kappa_{\mathbf{\Sigma}}^2 \kappa_{\mathbf{\Gamma}} \right)^{-1}}{3\kappa_{\mathbf{\Sigma}} \mathrm{deg}}$, *then* $\left\| \widetilde{\mathbf{\Theta}}_{\mathrm{param}} - \mathbf{\Theta}^{\natural} \right\|_\infty \leq r$.

Lemma 17 (Lemma 5 in Ravikumar et al. (2011))**.** *Provided that we have* $\left\| \widetilde{\mathbf{\Theta}}_{\mathrm{param}} - \mathbf{\Theta}^{\natural} \right\|_\infty \leq (3\kappa_{\mathbf{\Sigma}} \mathrm{deg})^{-1}$, *then*

$$\left\| \widetilde{\mathbf{\Theta}}_{\mathrm{param}}^{-1} - \mathbf{\Theta}^{\natural -1} + \mathbf{\Theta}^{\natural -1} \left(\widetilde{\mathbf{\Theta}}_{\mathrm{param}} - \mathbf{\Theta}^{\natural} \right) \mathbf{\Theta}^{\natural -1} \right\|_\infty \leq \frac{3}{2} \mathrm{deg} \cdot \ \kappa_{\mathbf{\Sigma}}^3 \left\| \widetilde{\mathbf{\Theta}}_{\mathrm{param}} - \mathbf{\Theta}^{\natural} \right\|_\infty^2.$$

Lemma 18 (Lemma 4 in Ravikumar et al. (2011))**.** *If we have*

$$\left\| \widehat{\mathbf{\Sigma}}_n^{\mathrm{param}} - \mathbf{\Sigma}^{\natural} \right\|_\infty \vee \left\| \widetilde{\mathbf{\Theta}}_{\mathrm{param}}^{-1} - \mathbf{\Theta}^{\natural -1} + \mathbf{\Theta}^{\natural -1} \left(\widetilde{\mathbf{\Theta}}_{\mathrm{param}} - \mathbf{\Theta}^{\natural} \right) \mathbf{\Theta}^{\natural -1} \right\|_\infty \leq \theta \lambda_{\mathrm{param}}/8,$$

we conclude that $|\widetilde{G}_{i,j}| < 1$.

Now, setting $\lambda_{\mathrm{param}} = 8\tau_\infty/\theta$, first we verify the conditions in Lemma 16. We have

$$r \overset{①}{\leq} 2\left(1 + 8\theta^{-1}\right) \kappa_{\mathbf{\Gamma}} \tau_\infty \overset{②}{\leq} (3\kappa_{\mathbf{\Sigma}} \mathrm{deg})^{-1} \left(1 \wedge \left(\kappa_{\mathbf{\Sigma}}^2 \kappa_{\mathbf{\Gamma}} \right)^{-1} \right),$$

where in ① we use $\left\| \widehat{\mathbf{\Sigma}}_n^{\mathrm{param}} - \mathbf{\Sigma}^{\natural} \right\|_\infty \leq \tau_\infty$ from Lemma 3, and in ② we use the assumptions of τ_∞ in Theorem 1. Then we conclude that

$$\left\| \widetilde{\mathbf{\Theta}}_{\mathrm{param}} - \mathbf{\Theta}^{\natural} \right\|_\infty \leq 2\kappa_{\mathbf{\Gamma}} \ \left(\left\| \widehat{\mathbf{\Theta}}_n - \mathbf{\Theta}^{\natural} \right\|_\infty + \frac{8\tau_\infty}{\theta} \right) \leq (3\kappa_{\mathbf{\Sigma}} \mathrm{deg})^{-1}.$$

Invoking Lemma 17, we have

$$\left\| \widetilde{\mathbf{\Theta}}_{\mathrm{param}}^{-1} - \mathbf{\Theta}^{\natural -1} + \mathbf{\Theta}^{\natural -1} \left(\widetilde{\mathbf{\Theta}}_{\mathrm{param}} - \mathbf{\Theta}^{\natural} \right) \mathbf{\Theta}^{\natural -1} \right\|_\infty \leq \frac{3}{2} \mathrm{deg} \times \ \kappa_{\mathbf{\Sigma}}^3 \tau_\infty^2 \overset{③}{\leq} \tau_\infty,$$

where ③ is due to the requirement of τ_∞ in Theorem 1. In the end, we verify the condition in Lemma 18,

$$\theta \lambda_{\mathrm{param}}/8 = \tau_\infty \geq \left\| \widehat{\mathbf{\Sigma}}_n^{\mathrm{param}} - \mathbf{\Sigma}^{\natural} \right\|_\infty \vee \left\| \widetilde{\mathbf{\Theta}}_{\mathrm{param}}^{-1} - \mathbf{\Theta}^{\natural -1} + \mathbf{\Theta}^{\natural -1} \left(\widetilde{\mathbf{\Theta}}_{\mathrm{param}} - \mathbf{\Theta}^{\natural} \right) \mathbf{\Theta}^{\natural -1} \right\|_\infty.$$

which concludes that $\widetilde{\mathbf{\Theta}}_{\mathrm{param}}$ is identical to the solution $\widehat{\mathbf{\Theta}}_{\mathrm{param}}$.

Step IV: bound the error. For the entries $(i,j) \in S^c$ outside the support set S, we have $(\widehat{\mathbf{\Theta}}_{\mathrm{param}})_{i,j}$ to be zero due to the construction method of $\widetilde{\mathbf{\Theta}}_{\mathrm{param}}$. For the entries $(i,j) \in S$ inside the support set, we first upper-bound the element-wise deviation $\left\| \widehat{\mathbf{\Theta}}_{\mathrm{param}} - \mathbf{\Theta}^{\natural} \right\|_\infty$ as

$$\left\| \widehat{\mathbf{\Theta}}_{\mathrm{param}} - \mathbf{\Theta}^{\natural} \right\|_\infty = \left\| \widetilde{\mathbf{\Theta}}_{\mathrm{param}} - \mathbf{\Theta}^{\natural} \right\|_\infty \overset{①}{\leq} 2\kappa_{\mathbf{\Gamma}} (1 + 8\theta^{-1}) \tau_\infty, \tag{29}$$

where ① is due to Lemma 16 and has been verified in Step II. Then we prove $\operatorname{sign}(\widehat{\mathbf{\Theta}}_{\text{param}})$ reaches the ground truth once $\min_{(i,j)\in S^c}\left|\mathbf{\Theta}^{\natural}_{i,j}\right| \geq 2\kappa_{\mathbf{\Gamma}}(1+8\theta^{-1})\tau_\infty$. W.l.o.g. we assume $\mathbf{\Theta}^{\natural}_{i,j} > 0$ for $(i,j)\in S$. Then we have

$$(\widehat{\mathbf{\Theta}}_{\text{param}})_{i,j} \geq \left|\mathbf{\Theta}^{\natural}_{i,j}\right| - \left|\widehat{\mathbf{\Theta}}_{\text{param}} - \mathbf{\Theta}^{\natural}\right| \geq \left|\mathbf{\Theta}^{\natural}_{i,j}\right| - \left\|\widehat{\mathbf{\Theta}}_{\text{param}} - \mathbf{\Theta}^{\natural}\right\|_\infty \overset{②}{>} 0,$$

and complete the proof, where ② is due to the assumption on $\min_{(i,j)\in S^c}\left|\mathbf{\Theta}^{\natural}_{i,j}\right|$ and equation 29.

C Analysis of the CDF Estimation

C.1 Notations

For the conciseness of notation, we drop the subscript i in the marginal CDFs $F_i(\cdot)$, $\widehat{F}_i(\cdot)$, and entry $\widehat{X}_i^{(s)}$, and denote them as $F(\cdot)$, $\widehat{F}(\cdot)$, and $\widehat{X}^{(s)}$, respectively. We define the approximate characteristics function $\widehat{\phi}_{\widehat{w}}(t)$ as

$$\widehat{\phi}_{\widehat{w}}(t) \triangleq \exp\left(-\frac{\sigma^2 t^2}{2(d-p)}\right).$$

Additionally, we construct the function $\widetilde{F}(\cdot)$ as

$$\widetilde{F}(x) = \frac{1}{2} - \frac{1}{\pi}\int_0^\infty \frac{1}{t}\Im\left[\frac{\phi_{\widehat{w}}(-t)\widehat{\phi}_{\widehat{X}}(t)}{\left|\phi_{\widehat{w}}(t)\right|^2 \vee \gamma t^a} e^{-\jmath t x}\right] dt,$$

where the characteristic function $\phi_{\widehat{w}}(\cdot)$ denotes the ground-truth characteristics function of the noise $\widehat{w}$, and the function $\widehat{\phi}_{\widehat{X}}$ is defined as $n^{-1}\sum_{s=1}^n e^{-\jmath t \widehat{X}^{(s)}}$.

Let the term $\Delta_x^{(s)}$ be

$$\Delta_x^{(s)} = \frac{1}{\pi}\int_0^\infty \frac{1}{t}\Im\left(\frac{\widehat{\phi}_{\widehat{w}}(-t)}{\left|\widehat{\phi}_{\widehat{w}}(t)\right|^2 \vee \gamma t^a}\exp\left(\jmath t \widehat{X}^{(s)} - \jmath t x\right)\right) dt - \mathbb{E}\left[\frac{1}{\pi}\int_0^\infty \frac{1}{t}\Im\left[\frac{\phi_{\widehat{X}}(t)}{\phi_{\widehat{w}}(t)} e^{-\jmath t x}\right] dt\right]. \tag{30}$$

Our goal is to obtain the uniform convergence rate of the CDF estimator $\left\|\widehat{F} - F\right\|_\infty$, which is written as

$$\left\|\widehat{F} - F\right\|_\infty = \sup_{x\in[0,1]}\left|\widehat{F}(x) - F(x)\right| = \sup_{x\in[0,1]}\frac{1}{n}\left|\sum_{s=1}^n \Delta_x^{(s)}\right| \triangleq \Delta_x.$$

Before proceed, we define the event $\mathcal{E}_w$ as

$$\mathcal{E}_w \triangleq \left\{\left|\mathbb{E}(\widehat{w}_i)^2 - \frac{\sigma^2}{d-p}\right| \leq \frac{\sigma^2}{5(d-p)}, \quad \forall\, 1 \leq i \leq p\right\}, \tag{31}$$

where $\widehat{w}_i$ denotes the ith entry of the noise $\widehat{\mathbf{W}}$.

The parameter γ is set as $c_0 \log(np)\left(\frac{\sigma^2}{d-p}\right)^{a/4}$ and the number $a > 1$ is some fixed positive constant.

C.2 Proof of Thm. 4

Proof. We decompose the probability $\sup_x \Delta_x \geq \mathbb{E}\sup_x \Delta_x + t$ as

$$\mathbb{P}\left(\left|\sup_x \Delta_x - \mathbb{E}\sup_x \Delta_x\right| \geq t\right) \leq \mathbb{E}\mathbb{1}(\overline{\mathcal{E}}_w) + \mathbb{E}\mathbb{1}\left[\left(\left|\sup_x \Delta_x - \mathbb{E}\sup_x \Delta_x\right| \geq t\right) \bigcap \mathcal{E}_w\right].$$

The first term $\mathbb{E}\mathbb{1}(\overline{\mathcal{E}}_w)$ is investigated in Lemma 19; while the second term is bounded with the Talagrand inequality (cf. Thm. 2.6 in Koltchinskii (2011)), which is stated as

$$\mathbb{E}\mathbb{1}\left[\left(\left|\sup_x \Delta_x - \mathbb{E}\sup_x \Delta_x\right| \geq t\right) \bigcap \mathcal{E}_w\right] \lesssim \mathbb{E}\left[\exp\left(-\frac{nt}{KC_U}\log\left(1+\frac{ntC_U}{V}\right)\right)\mathbb{1}(\mathcal{E}_w)\right],$$

where C_U is the uniform bound for $\Delta_x^{(s)}$, i.e., $|\Delta_x^{(s)}| \leq C_U$ for all x and s, and the variance V satisfies

$$V \geq n\sup_{x\in\mathbb{R}} \mathbb{E}\left|\Delta_x\right|^2 + 16C_U \cdot \left(\mathbb{E}\sup_x |\Delta_x|\right).$$

Setting t as $V\log n/n$, we conclude

$$\sup_x \Delta_x \leq \mathbb{E}\sup_x \Delta_x + \frac{cV\log n}{n}$$

holds with probability exceeding $1 - O(n^{-c})$. The following context focuses on computing the values of C_U, $\mathbb{E}\sup_x \Delta_x$, and $\sup_x \mathbb{E}\left|\Delta_x\right|^2$ conditional on event $\mathcal{E}_w$ in equation 31.

Step I. We show there exists some positive constant C_U such that $|\Delta_x^{(s)}| \leq C_U$ for all x and s, $1 \leq s \leq n$ (cf. Lemma 21);

Step II. We prove that $\mathbb{E}\sup_x \Delta_x \leq \mathbb{E}\Delta_x + c_0/\sqrt{n}$, where c_0 is some positive constant. With the symmetrization inequalities (cf. Lemma 11.4 in Boucheron et al. (2013)), we obtain

$$\mathbb{E}\sup_x (\Delta_x - \mathbb{E}\Delta_x) \leq \mathbb{E}\sup_x \frac{2}{n}\sum_{s=1}^n \varepsilon_s \Delta_x^{(s)}, \tag{32}$$

where $\{\varepsilon_s\}$ are the Rademacher RVs, i.e., $\mathbb{P}(\varepsilon_s = \pm 1) = 1/2$, $1 \leq s \leq n$. With regarding the empirical process $n^{-1}\left(\sum_s \varepsilon_s \Delta_x^{(s)}\right)$, we have

$$\mathbb{E}e^{\lambda n^{-1}\left(\sum_s \varepsilon_s\left(\Delta_{x_1}^{(s)} - \Delta_{x_2}^{(s)}\right)\right)} \overset{①}{=} \prod_{s=1}^n \mathbb{E}\exp\left[\frac{\lambda\varepsilon_s}{n}\left(\Delta_{x_1}^{(s)} - \Delta_{x_2}^{(s)}\right)\right] \overset{②}{\leq} \prod_{s=1}^n \mathbb{E}\exp\left[\frac{\lambda^2}{2n^2}\left(\Delta_{x_1}^{(s)} - \Delta_{x_2}^{(s)}\right)^2\right],$$

where in ① we exploit the independence across the samples, in ② we use the fact that ε_s is a Rademacher RV and the Hoeffding's lemma Boucheron et al. (2013) (Lemma 2.2).

Invoking the Dudley's entropy integral (cf. Corol 13.2 in Boucheron et al. (2013)), which is also listed as Thm. 9 for the sake of self-containing, we conclude

$$\mathbb{E}\sup_{x,\varepsilon_s} \frac{1}{n}\sum_{s=1}^n \varepsilon_s \Delta_x^{(s)} \lesssim \frac{1}{\sqrt{n}}\int_0^{2c_0} \sqrt{\mathcal{H}([0,1],\delta)}d\delta \overset{③}{\lesssim} \frac{1}{\sqrt{n}}\int_0^{2c_0}\left(1+\sqrt{\log\frac{c_0}{t}}\right)d\delta \asymp \frac{1}{\sqrt{n}}, \tag{33}$$

where $\mathcal{H}([0,1],\delta)$ denotes the δ-entropy number Boucheron et al. (2013), in ③ we upper-bound the δ-entropy number $\mathcal{H}([0,1],\delta)$ as $c_0 + c_1\log(b/\delta)$ (cf. Example 5.24 in Wainwright (2019)). The proof is then completed by combining equation 32 and equation 33.

Step III. We set the variance V as

$$\begin{aligned} V &= n\sup_{x\in\mathbb{R}} \mathbb{E}\left|\Delta_x\right|^2 + \frac{c_0}{\sqrt{n}} + c_1\sqrt{\mathbb{E}\Delta_x^2} \overset{④}{\geq} n\sup_{x\in\mathbb{R}} \mathbb{E}\left|\Delta_x\right|^2 + \frac{c_0}{\sqrt{n}} + c_1\mathbb{E}\Delta_x \\ &\overset{⑤}{\geq} n\sup_{x\in\mathbb{R}} \mathbb{E}\left|\Delta_x\right|^2 + c_0\mathbb{E}\sup_x \Delta_x, \end{aligned}$$

where in ④ we use the fact $\mathbb{E}\Delta_x \leq \sqrt{\mathbb{E}\Delta_x^2}$, and in ⑤ we use the results in Step II such that $\mathbb{E}\sup_x \Delta_x \leq \mathbb{E}\Delta_x + c_0/\sqrt{n}$. Invoking Lemma 19 and Lemma 22 will complete the proof.

□

C.3 Supporting Lemmas

Lemma 19. *The event $\mathcal{E}_w$ in equation 31 holds with probability exceeding $1-2pe^{-c_0 p}$, where c_0 is some fixed positive constants.*

Proof. Perform SVD for $\mathbf{A}$ as $\mathbf{A}=\mathbf{U}\mathbf{S}\mathbf{V}^\top$, we can rewrite the product $\left(\mathbf{A}^\top\mathbf{A}\right)^{-1}\mathbf{A}^\top$ as

$$\left(\mathbf{A}^\top\mathbf{A}\right)^{-1}\mathbf{A}^\top=\mathbf{V}\mathbf{S}^{-2}\mathbf{V}^\top\mathbf{V}\mathbf{S}^\top\mathbf{U}^\top=\mathbf{V}\mathbf{S}^{-1}\mathbf{U}^\top.$$

Then the ith entry of $\widehat{\mathbf{W}}$ can be written as $\boldsymbol{a}_i^\top\mathbf{U}^\top\mathbf{W}$, where $\boldsymbol{a}_i$ as

$$\boldsymbol{a}_i=\left[\lambda_1^{-1}V_{i1}\ \ldots\ \lambda_p^{-1}V_{ip}\right],$$

and λ_i is the ith singular value of $\mathbf{S}$, and V_{ij} is the (i,j)th entry of the matrix $\mathbf{V}$. Since each entry A_{ij} is iid standard normal RV, i.e., $A_{ij}\sim\mathsf{N}(0,1)$, we conclude that its eigenvalues $\{\lambda_i\}_{1\le i\le p}$ are independent from its eigenvectors $\mathbf{V}$, which is uniformly distributed on a Haar measure Tulino et al. (2004). Hence we can rewrite $\boldsymbol{a}_i$ as a product as $\mathbf{S}^{-1}\boldsymbol{g}/\|\boldsymbol{g}_i\|_2$, where $\boldsymbol{g}_i$ is a Gaussian distributed RV with zero mean and unit variance, namely, $\boldsymbol{g}_i\sim\mathsf{N}(0,\mathbf{I}_{p\times p})$.

First, we define the event $\mathcal{E}_g$ as

$$\mathcal{E}_g\triangleq\left\{\exists\ 1\le i\le p,\ \ \text{s.t.}\ \ \left|\|\boldsymbol{g}_i\|_2^2-p\right|\ge p/4\right\}.$$

Then we bound $\mathbb{E}\mathbb{1}(\overline{\mathcal{E}}_w)$ as

$$\mathbb{E}\mathbb{1}\left(\overline{\mathcal{E}}_w\right)\le\mathbb{E}\mathbb{1}\left(\overline{\mathcal{E}}_g\right)+\mathbb{E}\mathbb{1}\left(\overline{\mathcal{E}}_w\bigcap\mathcal{E}_g\right),$$

where $\overline{(\cdot)}$ denotes the complement of the event. Due to the fact that $\boldsymbol{g}$ is a Gaussian RV such that $\boldsymbol{g}\sim\mathsf{N}(0,\mathbf{I})$, we invoke Lemma 30 and bound the first term as $\mathbb{E}\mathbb{1}(\mathcal{E}_g)\le 2pe^{-c_0 p}$. While for the second term, we first compute the expectation $\mathbb{E}\left\|\mathbf{S}^{-1}\boldsymbol{g}_i\mathbf{U}^\top\mathbf{W}\right\|_2^2$ as

$$\mathbb{E}\left\|\mathbf{S}^{-1}\boldsymbol{g}_i\mathbf{U}^\top\mathbf{W}\right\|_2^2=\sigma^2\mathbb{E}\left\|\mathbf{S}^{-1}\boldsymbol{g}_i\right\|_2^2=\sigma^2\|\mathbf{S}^{-1}\|_{\mathrm{F}}^2.$$

Invoking the Marcenko-Pastur law in Tulino et al. (2004) (cf. Thm. 2.35, which is also listed as Lemma 8 for the self-containing of paper), we have

$$\begin{aligned}\|\mathbf{S}^{-1}\|_{\mathrm{F}}^2&=\sum_{i=1}^{p}\lambda_i^{-2}\to\frac{p}{d}\int_{L(\tau)}^{U(\tau)}t^{-1}\left[\left(1-\tau^{-1}\right)_+\mathbb{1}(t)+\frac{\sqrt{[t-L(\tau)][U(\tau)-t]}}{2\pi\tau t}\right]dt\\&=\frac{p}{d}\times\frac{1}{1-\tau}=\frac{p}{d-p},\end{aligned}$$

where τ is defined as p/d, function $L(\cdot)$ is defined as $\left(1-\sqrt{\cdot}\right)^2$, and function $U(\cdot)$ is defined as $\left(1+\sqrt{\cdot}\right)^2$. Hence we have $\mathbb{E}\left\|\mathbf{S}^{-1}\boldsymbol{g}\mathbf{U}^\top\boldsymbol{w}\right\|_2^2=\mathbb{E}\|\mathbf{S}^{-1}\|_{\mathrm{F}}^2\sigma^2=p\sigma^2/(d-p)$ and will show $\mathbb{E}\mathbb{1}\left(\overline{\mathcal{E}}_w\bigcap\mathcal{E}_g\right)$ to be zero. This is because

$$\mathbb{E}\left[\frac{\left\|\mathbf{S}^{-1}\boldsymbol{g}_i\mathbf{U}^\top\mathbf{W}\right\|_2^2}{\|\boldsymbol{g}_i\|_2^2}\mathbb{1}(\mathcal{E}_g)\right]\overset{①}{\ge}\frac{4}{3p}\mathbb{E}\left\|\mathbf{S}^{-1}\boldsymbol{g}_i\mathbf{U}^\top\mathbf{W}\right\|_2^2=\frac{4\sigma^2}{3(d-p)},$$

where ① is due to the definition of $\mathcal{E}_g$. Similarly we can show

$$\mathbb{E}\left[\frac{\left\|\mathbf{S}^{-1}\boldsymbol{g}_i\mathbf{U}^\top\mathbf{W}\right\|_2^2}{\|\boldsymbol{g}_i\|_2^2}\mathbb{1}(\mathcal{E}_g)\right]\le\frac{4}{5p}\mathbb{E}\left\|\mathbf{S}^{-1}\boldsymbol{g}_i\mathbf{U}^\top\mathbf{W}\right\|_2^2=\frac{4\sigma^2}{5(d-p)},$$

which suggests that $\mathcal{E}_w$ will always hold, and complete the proof. □

Lemma 20. *Conditional on the event $\mathcal{E}_w$ in equation 31, we have*

$$\left|\widehat{\phi}_{\widehat{w}}(t) - \phi_{\widehat{w}}(t)\right| \lesssim \frac{\log^2(np)\sigma^2 t^2}{d-p} e^{-\frac{\sigma^2 t^2}{2(d-p)}},$$

where $\widehat{\phi}_{\widehat{w}}(t)$ and $\phi_{\widehat{w}}(t)$ denotes the estimated characteristic function and the ground-truth characteristic function of $\widehat{w}_i$, respectively.

Proof. Notice that the characteristic function of the Gaussian $\mathsf{N}(0, \mathrm{Var})$ is written as

$$\phi_{\mathrm{Var}}(t) = \exp\left(-\frac{(\mathrm{Var})t^2}{2}\right).$$

Then we conclude that

$$\begin{aligned}
\left|\widehat{\phi}_{\widehat{w}}(t) - \phi_{\widehat{w}}(t)\right| &= \left|\exp\left(-\frac{\sigma^2 t^2}{2(d-p)}\right) - \exp\left(-\frac{\widehat{\sigma}^2 t^2}{2}\right)\right| \\
&\overset{①}{\leq} \frac{\left|\sigma^2/(d-p) - \widehat{\sigma}^2\right| t^2}{2} \cdot \left(\exp\left(-\frac{\widehat{\sigma}^2 t^2}{2}\right) + \exp\left(-\frac{\sigma^2 t^2}{2(d-p)}\right)\right) \\
&\overset{②}{\lesssim} \frac{\log^2(np)\sigma^2 t^2}{d-p} e^{-\frac{\sigma^2 t^2}{2(d-p)}},
\end{aligned}$$

where $\widehat{\sigma}^2$ is defined as the ground-truth variance of the noise $\widehat{w}$, ① is because of the relation $|e^{-z_1} - e^{-z_2}| \leq |z_1 - z_2|\,|e^{-z_1} + e^{-z_2}|/2$ for arbitrary z_1 and z_2, ② is because of event $\mathcal{E}_w$. □

Lemma 21. *Setting $\gamma \asymp \log(np)\left(\frac{\sigma^2}{d-p}\right)^{a/4}$, we conclude $|\Delta_x^{(s)}| \lesssim 1$ for all s, $1 \leq s \leq n$, where $\Delta_x^{(s)}$ is defined in equation 30.*

Proof. We verify that $\Delta_x^{(s)}$ to be bounded by some constant, which is written as

$$\Delta_x^{(s)} \leq \left|\frac{1}{\pi}\int_0^\infty \frac{1}{t}\Im\left(\frac{\widehat{\phi}_{\widehat{w}}(-t)}{\left|\widehat{\phi}_{\widehat{w}}(t)\right|^2 \vee \gamma t^a} \exp\left(\mathbb{j}t\left(\widehat{X}^{(s)} - x\right)\right)\right) dt\right| + \underbrace{F(x)}_{\leq 1}. \tag{34}$$

Defining the term $t_\perp$ as $(2\gamma)^{-1/a}$, we split the whole region $(0, \infty)$ as three disjoint sub-regions $\mathcal{R}_1 = (0, t_\perp)$ and $\mathcal{R}_2 = (t_\perp, \infty)$. Then we perform the decomposition as

$$\begin{aligned}
&\int_0^\infty \frac{1}{t}\Im\left(\frac{\widehat{\phi}_{\widehat{w}}(-t)}{\left|\widehat{\phi}_{\widehat{w}}(t)\right|^2 \vee \gamma t^a} \exp\left(\mathbb{j}t\widehat{X}^{(s)} - \mathbb{j}tx\right)\right) dt \\
&= \underbrace{\int_{\mathcal{R}_1} \frac{1}{t}\Im\left(\frac{\widehat{\phi}_{\widehat{w}}(-t)}{\left|\widehat{\phi}_{\widehat{w}}(t)\right|^2 \vee \gamma t^a} \exp\left(\mathbb{j}t\widehat{X}^{(s)} - \mathbb{j}tx\right)\right) dt}_{T_1} \\
&+ \underbrace{\int_{\mathcal{R}_2} \frac{1}{t}\Im\left(\frac{\widehat{\phi}_{\widehat{w}}(-t)}{\left|\widehat{\phi}_{\widehat{w}}(t)\right|^2 \vee \gamma t^a} \exp\left(\mathbb{j}t\widehat{X}^{(s)} - \mathbb{j}tx\right)\right) dt}_{T_2},
\end{aligned} \tag{35}$$

Then we separately bound the terms T_1 and T_2.

Step I. We can bound term T_1 as

$$T_1 = \int_{\mathcal{R}_1} \frac{1}{t}\Im\left(\widehat{\phi}_{\widehat{w}}(-t)\Im\left[\left(\exp\left(\mathbb{j}t\widehat{X}^{(s)} - \mathbb{j}tx\right)\right)\left[1+\sum_{k=1}^{\infty}\left(1-\left(\left|\widehat{\phi}_{\widehat{w}}(t)\right|^2 \vee \gamma t^a\right)\right)^k\right]\right]\right)dt$$
$$\overset{①}{\leq} \int_{\mathcal{R}_1} \frac{\sin\left(t\left(\widehat{X}^{(s)}-x\right)\right)}{t}dt + \int_{\mathcal{R}_1}\frac{1}{t}\sum_{k=1}^{\infty}\left(\frac{\sigma^2t^2}{d-p}\right)^k dt \overset{②}{\lesssim} 1+\sum_{k=1}^{\infty}\frac{1}{2k}\left(\frac{\sigma^2t_\perp^2}{d-p}\right)^k \overset{③}{\lesssim} 1, \tag{36}$$

where in ① we use the fact $\left|\phi_{\widehat{w}}(t)\right|^2 \vee \delta t^a \geq 1 - \frac{c_1\sigma^2t^2}{d-p}$ for $t \in \mathcal{R}_1$, in ② we use the fact $\sup_{\tau>0}\left|\int_0^\tau \sin(u)/u du\right| \leq 3$, and ③ is because $\sigma^2t_\perp^2 \asymp \sqrt{d-p}(\log(np))^{-\frac{2}{a}} \leq \frac{d-p}{2}$ and hence $\sum_{k=1}^{\infty}\frac{1}{2k}\left(\frac{\sigma^2t_\perp^2}{d-p}\right)^k \leq \sum_{k=1}^{\infty}\left(\frac{\sigma^2t_\perp^2}{d-p}\right)^k \lesssim 1$.

Step II. For term T_3, we have

$$T_2 \leq \int_{\mathcal{R}_2}\frac{\widehat{\phi}_{\widehat{w}}(-t)}{t\sqrt{\left|\widehat{\phi}_{\widehat{w}}(-t)\right|^2 \vee \delta t^a}}dt \overset{④}{\leq} \int_{\mathcal{R}_2}\frac{1}{\sqrt{\gamma}t^{1+\frac{a}{2}}}dt = \frac{1}{a\sqrt{\gamma}t_\perp^{a/2}} = \frac{1}{\sqrt{2}a} \asymp 1, \tag{37}$$

where ④ is because $\widehat{\phi}_{\widehat{w}}(\cdot) \leq 1$, and complete the proof by summaring equation 34, equation 35 and equation 37. □

Lemma 22. *Under the setting of Thm. 4, we have*

$$\mathbb{E}\left(\Delta_x\right)^2 \lesssim \frac{\log^{2/a}(np)\sigma}{\sqrt{d-p}} + \frac{(\log(np))^2}{(d-p)^{\frac{a}{4}}} + \left(\frac{\sigma^2}{d-p}\right)^{\frac{2\alpha+1}{4}} + \frac{1}{n},$$

when setting $\gamma \asymp \log(np)\left(\frac{\sigma^2}{d-p}\right)^{a/4}$

Proof. The proof largely follows Phuong (2020). However, extra measurements are required to estimate the characteristic function $\widehat{\phi}_{\widehat{w}}(\cdot)$ in Phuong (2020), which leads to a simple form of the error $|\widehat{\phi}_{\widehat{w}}(t) - \phi_{\widehat{w}}(t)|$ only depending on the number of extra measurements. While our setting does not need these additional measurements and the error $|\widehat{\phi}_{\widehat{w}}(t) - \phi_{\widehat{w}}(t)|$ varies with t.

With the decomposition

$$\mathbb{E}\left|\widehat{F}(x) - F(x)\right|^2 \leq 2\mathbb{E}\left|\widehat{F}(x) - \widetilde{F}(x)\right|^2 + 2\mathbb{E}\left|\widetilde{F}(x) - F(x)\right|^2,$$

we complete the proof by invoking Lemma 23 and Lemma 24. □

Lemma 23. *Under the setting of Thm. 4, we can upper-bound the deviation* $\mathbb{E}\left|\widehat{F}(x) - \widetilde{F}(x)\right|$ *as*

$$\mathbb{E}\left|\widehat{F}(x) - \widetilde{F}(x)\right|^2 \lesssim \frac{\log^{2/a}(np)\sigma}{\sqrt{d-p}},$$

when setting $\gamma \asymp \log(np)\left(\frac{\sigma^2}{d-p}\right)^{a/4}$.

Proof. First we expand the term $\widehat{F}(x) - \widetilde{F}(x)$ as

$$\widehat{F}(x) - \widetilde{F}(x) = \frac{1}{\pi}\int_0^\infty \frac{1}{t}\Im\left[\left(\frac{\widehat{\phi}_{\widehat{w}}(-t)}{\left|\widehat{\phi}_{\widehat{w}}(t)\right|^2 \vee \gamma t^a} - \frac{\phi_{\widehat{w}}(-t)}{\left|\phi_{\widehat{w}}(t)\right|^2 \vee \gamma t^a}\right)\widehat{\phi}_{\widehat{X}}(t)e^{-\mathbb{j}tx}\right]dt$$
$$\overset{①}{=} \frac{1}{n\pi}\int_0^\infty \frac{D(t)}{t}\sum_{s=1}^{n}\sin\left(t\left(\widehat{X}^{(s)}-x\right)\right)dt$$

where in ① we define $D(t)$ as

$$D(t) \triangleq \frac{\widehat{\phi}_{\widehat{w}}(-t)}{|\widehat{\phi}_{\widehat{w}}(t)|^2 \vee \gamma t^a} - \frac{\phi_{\widehat{w}}(-t)}{\left|\phi_{\widehat{w}}(t)\right|^2 \vee \gamma t^a}.$$

According to Phuong (2020), it satisfies the relation

$$|D(t)| \leq \frac{2\varepsilon_E(t)}{\sqrt{|\widehat{\phi}_{\widehat{w}}(t)|^2 \vee \gamma t^a} \cdot \sqrt{\left|\phi_{\widehat{w}}(t)\right|^2 \vee \gamma t^a}} + \frac{\varepsilon_E(t)}{\left|\phi_{\widehat{w}}(t)\right|^2 \vee \gamma t^a}, \tag{38}$$

where $\varepsilon(t)$ is defined as $\left|\widehat{\phi}_{\widehat{w}}(t) - \phi_{\widehat{w}}(t)\right|$, which is upper bounded by Lemma 20.

Define the terms I_1 and I_2 as

$$I_1 \triangleq \mathbb{E}\left(\frac{1}{n\pi}\int_0^{t_\perp}\left|\frac{D(t)}{t}\sum_{s=1}^n \sin\left(t\left(\widehat{X}^{(s)} - x\right)\right)\right| dt\right)^2,$$
$$I_2 \triangleq \mathbb{E}\left(\frac{1}{n\pi}\int_{t_\perp}^{\infty}\left|\frac{D(t)}{t}\sum_{s=1}^n \sin\left(t\left(\widehat{X}^{(s)} - x\right)\right)\right| dt\right)^2,$$

respectively, where $t_\perp$ is defined as $((d-p)/\sigma^2)^{1/4}\log^{1/a}(np)$. We upper-bound the term $\mathbb{E}\left|\widehat{F}(x) - \widetilde{F}(x)\right|^2$ as

$$\mathbb{E}\left|\widehat{F}(x) - \widetilde{F}(x)\right|^2 \leq 2I_1 + 2I_2. \tag{39}$$

Stage I. We bound the term I_1 as

$$I_1 \overset{②}{\leq} \left(\frac{1}{\pi}\int_0^{t_\perp}\left[\left|\frac{D(t)}{t}\right| \times n^{-1}\left(\sum_{s=1}^n \left|t\left(\widehat{X}^{(s)} - x\right)\right|\right)\right] dt\right)^2 \overset{③}{\lesssim} \left(\int_0^{t_\perp}|D(t)|\, dt\right)^2,$$

where in ② we use the fact $\sin(\cdot) \leq |\cdot|$, in ③ we use the fact $|\widehat{X}^{(s)} - x| \leq 1$ for all s, $1 \leq s \leq n$.

Notice in the region $\mathcal{R}_1$, we can lower bound the function $\phi_{\widehat{w}}(\cdot) \geq c_1$ as

$$\left|\phi_{\widehat{w}}(t)\right| \overset{④}{\geq} \left|\phi_{\widehat{w}}(t_\perp)\right| = O\left[\exp\left(-\frac{c_0\sigma^2}{d-p}\cdot\frac{\sqrt{d-p}}{\sigma}\right)\right] = O(1),$$

where in ④ we use the fact such that $\left|\phi(\cdot)_{\widehat{w}}\right|$ is non-increasing. Then we invoke equation 38 and bound $D(t)$ as $|D(t)| \lesssim \varepsilon_E(t)$, since $\left|\phi_{\widehat{w}}(t)\right|^2 \vee \gamma t^a \geq \left|\phi_{\widehat{w}}(t)\right|^2 \gtrsim 1$. Hence term I_1 is upper-bounded as

$$\begin{aligned} I_1 &\leq \left[\int_0^{t_\perp}\frac{\sigma^2\log^2(np)t^2}{d-p}e^{-\frac{\sigma^2 t^2}{2(d-p)}}dt\right]^2 \overset{⑤}{\leq} \log^4(np)\left(\int_0^{\log^{1/a}(np)(\frac{\sigma^2}{d-p})^{1/4}}\xi^2 e^{-\xi^2/2}d\xi\right)^2, \\ &\leq \frac{\sigma^2\log^{4+\frac{2}{a}}(np)}{d-p}\left(\int_0^\infty e^{-\xi^2/2}d\xi\right)^2 \asymp \frac{\sigma^2\log^{4+\frac{2}{a}}(np)}{d-p}, \end{aligned} \tag{40}$$

where in ⑤ we use the substitution $\xi = \sigma t/\sqrt{d-p}$.

Stage II. We define the function $\Lambda(t)$ as

$$\Lambda(t) = \frac{1}{n}\sum_{s=1}^n \sin\left[t\left(\widehat{X}^{(s)} - x\right)\right] - \Im\left(\phi_{\widehat{X}}(t)e^{-\mathbb{j}tx}\right),$$

and bound the term I_2 as

$$I_2 = \mathbb{E}\left|\int_{t_\perp}^{\infty} \frac{D(t)}{t}\left[\Lambda(t) + \Im\left(\phi_{\widehat{X}}(t)e^{-\mathrm{j}tx}\right)\right]\right|^2 \lesssim I_{2,1} + I_{2,2},$$

where $I_{2,1}$ and $I_{2,2}$ are defined as

$$I_{2,1} \triangleq \mathbb{E}\left|\int_{t_\perp}^{\infty} \frac{D(t)}{t}\Im\left(\phi_{\widehat{X}}(t)e^{-\mathrm{j}tx}\right)\right|^2,$$
$$I_{2,2} \triangleq \mathbb{E}\left|\int_{t_\perp}^{\infty} \frac{D(t)\Lambda(t)}{t}\right|^2.$$

Notice that within region $\mathcal{R}_2$, we can upper-bound $|D(t)|$ as $|D(t)| \lesssim \frac{\varepsilon_E(t)}{\gamma t^a}$ and hence

$$\begin{aligned}
I_{2,1} &\leq \left|\int_{t_\perp}^{\infty} \frac{\left|\phi_{\widehat{X}}(t)\right|\sqrt{\mathbb{E}|D(t)|^2}}{t}dt\right|^2 \overset{⑥}{\leq} \left|\int_{t_\perp}^{\infty} \frac{\sqrt{\mathbb{E}|D(t)|^2}}{t}dt\right|^2 \overset{⑦}{\leq} \left(\int_{t_\perp}^{\infty} \frac{\varepsilon_E(t)}{\gamma t^{1+a}}dt\right)^2 \\
&= \frac{\log^4(np)\sigma^4}{(d-p)^2\gamma^2}\left(\int_{t_\perp}^{\infty} \exp\left(-\frac{\sigma^2 t^2}{2(d-p)}\right)t^{1-a}dt\right)^2 \\
&= \frac{\log^4(np)\sigma^{2a}}{(d-p)^a\gamma^2}\left(\int_{\log^{1/a}(np)(\frac{\sigma^2}{d-p})^{1/4}}^{\infty} \xi^{1-a}e^{-\xi^2/2}d\xi\right)^2 \\
&\overset{⑧}{\leq} \frac{\log^4(np)\sigma^{2a}}{(d-p)^a\gamma^2}\log^{\frac{2}{a}-2}(np)\left(\frac{\sigma^2}{d-p}\right)^{\frac{1-a}{2}}\left(\int_0^{\infty} e^{-\xi^2/2}d\xi\right)^2 \\
&\asymp \frac{\log^{2/a}(np)\sigma}{\sqrt{d-p}},
\end{aligned}$$

where in ⑥ we use the fact $|\phi_{\widehat{X}}(\cdot)| \leq 1$, in ⑦ we use the bound $|D(t)| \leq \varepsilon_E/(\gamma t^a)$, and in ⑧ we use the assumption $a > 1$.

Afterwards, we bound term $I_{2,2}$ as

$$\begin{aligned}
I_{2,2} &\leq 2\mathbb{E}\left|\int_{t_\perp}^{\infty}\int_{t_\perp}^{\infty} \frac{D(u)D(v)\Lambda(u)\Lambda(v)}{uv}dudv\right| \\
&\leq 2\int_{t_\perp}^{\infty}\int_{t_\perp}^{\infty} \frac{\sqrt{\mathbb{E}|D(u)|^2}\sqrt{\mathbb{E}|D(v)|^2}\cdot\mathbb{E}\left(\Lambda(u)\Lambda(v)\right)}{uv}dudv.
\end{aligned}$$

According to Lemma 5.1 in Phuong (2020), we can bound the term $\mathbb{E}\left(\Lambda(u)\Lambda(v)\right)$ as

$$\begin{aligned}
\mathbb{E}\left(\Lambda(u)\Lambda(v)\right) &= \frac{1}{2n}\left[\Re\left[e^{\mathrm{j}(v-u)t}\phi_{\widehat{X}}(u-v)\right] - \Re\left[e^{-\mathrm{j}(u+v)x}\phi_{\widehat{X}}(u+v)\right] - 2\Im\left[e^{-\mathrm{j}ux}\phi_{\widehat{X}}(u)\right]\cdot\Im\left[e^{-\mathrm{j}vx}\phi_{\widehat{X}}(s)\right]\right] \\
&\leq \frac{1}{2n}\left[\left|\phi_{\widehat{X}}(u+v)\right| + \left|\phi_{\widehat{X}}(u-v)\right| + 2\left|\phi_{\widehat{X}}(u)\right|\left|\phi_{\widehat{X}}(v)\right|\right] \overset{⑨}{\leq} \frac{2}{n},
\end{aligned}$$

where in ⑨ we use the fact $|\phi_{\widehat{X}}(\cdot)| \leq 1$.

Following the same strategy as above, we can upper-bound $|D(t)|$ as $|D(t)| \lesssim \frac{\varepsilon_E(t)}{\gamma t^a}$ and hence

$$\sqrt{\mathbb{E}|D(u)|^2}\sqrt{\mathbb{E}|D(v)|^2} \leq \frac{\varepsilon_E(u)\varepsilon_E(v)}{\gamma^2 u^a v^a}.$$

Combing the above then yields the bound

$$I_{2,2} \lesssim \frac{1}{n\gamma^2}\int_{t_\perp}^{\infty}\int_{t_\perp}^{\infty}\frac{\varepsilon_E(u)\varepsilon_E(v)}{u^{1+a}v^{1+a}}dudv = \frac{\log^{4/a}(np)\sigma^{2+a}}{n(d-p)}.$$

To sum up, we have

$$I_2 = I_{2,1} + I_{2,2} \lesssim \frac{\log^{2/a}(np)\sigma^{1+\frac{a}{2}}}{\sqrt{d-p}} + \frac{\log^{4/a}(np)\sigma^2}{n(d-p)} \asymp \frac{\log^{2/a}(np)\sigma}{\sqrt{d-p}},$$

and complete the proof by combining it with equation 39 and equation 40. □

Lemma 24. *Under the setting of Thm. 4, we can upper-bound the deviation* $\left|\widetilde{F}(x) - F(x)\right|$ *as*

$$\mathbb{E}\left|\widetilde{F}(x) - F(x)\right|^2 \lesssim \frac{(\log(np))^2}{(d-p)^{\frac{a}{4}}} + \left(\frac{\sigma^2}{d-p}\right)^{\frac{2\alpha+1}{4}} + \frac{1}{n},$$

when setting $\gamma \asymp \log(np)\left(\frac{\sigma^2}{d-p}\right)^{a/4}$.

Proof. We decompose the deviation $\mathbb{E}|\widetilde{F}(x) - F(x)|^2$ as the the bias and variance, which are defined respectively as

$$\text{Bias} \triangleq \left|\mathbb{E}\widetilde{F}(x) - F(x)\right|^2,$$
$$\text{Variance} \triangleq \mathbb{E}\left|\widetilde{F}(x) - \mathbb{E}\widetilde{F}(x)\right|^2.$$

The following context separately bound the bias and variance.

Bounding bias. We rewrite the difference $\mathbb{E}\widetilde{F}(x) - F(x)$ as

$$\mathbb{E}\widetilde{F}(x) - F(x) = \frac{1}{\pi}\int_0^\infty \frac{1}{t}\Im\left[\left(\frac{\left|\phi_{\widehat{w}}(t)\right|^2\phi_X(t)}{\left|\phi_{\widehat{w}}(t)\right|^2 \vee \gamma t^a} - \phi_X(t)\right)e^{-\mathbb{j}tx}\right]dt,$$

which yields

$$\begin{aligned}
\text{Bias} &\le \frac{1}{\pi^2}\left[\int_0^\infty \frac{|\phi_X(t)|}{t}\Im\left[\left(\frac{\left|\phi_{\widehat{w}}(t)\right|^2}{\left|\phi_{\widehat{w}}(t)\right|^2 \vee \gamma t^a} - 1\right)e^{-\mathbb{j}tx}\right]dt\right]^2 \\
&\le \frac{1}{\pi^2}\left[\int_0^\infty \frac{|\phi_X(t)|}{t}\left|1 - \frac{\left|\phi_{\widehat{w}}(t)\right|^2}{\left|\phi_{\widehat{w}}(t)\right|^2 \vee \gamma t^a}\right|dt\right]^2 \\
&\lesssim \left[\int_0^{t_\square} \frac{|\phi_X(t)|}{t}\left|1 - \frac{\left|\phi_{\widehat{w}}(t)\right|^2}{\left|\phi_{\widehat{w}}(t)\right|^2 \vee \gamma t^a}\right|dt + \int_{t_\square}^\infty \frac{|\phi_X(t)|}{t}\left|1 - \frac{\left|\phi_{\widehat{w}}(t)\right|^2}{\left|\phi_{\widehat{w}}(t)\right|^2 \vee \gamma t^a}\right|dt\right]^2 \le J_1^2 + J_2^2,
\end{aligned}$$

where $t_\square$ is defined as $c_0\left(\frac{d-p}{\sigma^2}\right)^{1/8}$, and terms J_1 and J_2 are defined as

$$J_1 \triangleq \int_0^{t_\square} \frac{|\phi_X(t)|}{t}\left|1 - \frac{\left|\phi_{\widehat{w}}(t)\right|^2}{\left|\phi_{\widehat{w}}(t)\right|^2 \vee \gamma t^a}\right|dt,$$
$$J_2 \triangleq \int_{t_\square}^\infty \frac{|\phi_X(t)|}{t}\left|1 - \frac{\left|\phi_{\widehat{w}}(t)\right|^2}{\left|\phi_{\widehat{w}}(t)\right|^2 \vee \gamma t^a}\right|dt,$$

respectively. For the term J_1, we have

$$1-\frac{\left|\phi_{\widehat{w}}(t)\right|^2}{\left|\phi_{\widehat{w}}(t)\right|^2\vee\gamma t^a}\leq\frac{\left(\left|\phi_{\widehat{w}}(t)\right|^2\vee\gamma t^a\right)-\left|\phi_{\widehat{w}}(t)\right|^2}{\left|\phi_{\widehat{w}}(t)\right|^2\vee\gamma t^a}\overset{①}{\leq}\frac{\gamma t^a}{\left|\phi_{\widehat{w}}(t)\right|^2\vee\gamma t^a}\overset{②}{\leq}\frac{\gamma t^a}{c_R},$$

where in ① we use the relation $\left|\phi_{\widehat{w}}(t)\right|^2\vee\gamma t^a-\left|\phi_{\widehat{w}}(t)\right|^2\leq\gamma t^a$; in ② we use the fact $\left|\phi_{\widehat{w}}(t)\right|^2\vee\gamma t^a\geq c_R$ in the regime $[0,t_\square)$, which can be easily verified. Then we obtain

$$J_1\leq\int_0^{t_\square}\frac{|\phi_X(t)|}{t}\frac{\gamma t^a}{c_R}dt=\frac{\gamma}{c_R}\int_0^{t_\square}t^{a-1}dt=\frac{\gamma}{c_R}\frac{t_\square^a}{a}\asymp\frac{\sigma^{\frac{a}{4}}\log(np)}{(d-p)^{\frac{a}{8}}a}. \tag{41}$$

Afterwards, we bound term J_2 as

$$J_2\leq\int_{t_\square}^{\infty}\frac{|\phi_X(t)|}{t}dt\overset{③}{\lesssim}t_\square^{-(2\alpha+1)/16}\asymp\left(\frac{\sigma^2}{d-p}\right)^{\frac{2\alpha+1}{8}}, \tag{42}$$

where in ③ we use the Lemma 6 from Phuong (2020) since $\phi_X(\cdot)$ satisfies the Assumption 2. Combining equation 41 and equation 42 then yields

$$\text{Bias}\lesssim\frac{(\log(np))^2}{(d-p)^{\frac{a}{4}}}+\left(\frac{\sigma^2}{d-p}\right)^{\frac{2\alpha+1}{4}}. \tag{43}$$

Bounding variance. We bound the $\text{Var}\widetilde{F}(x)$ as

$$\begin{aligned}
&\text{Var}\widetilde{F}(x)\\
&=\text{Var}\left[\frac{1}{\pi}\int_0^\infty\frac{1}{t}\Im\left(\frac{\phi_{\widehat{w}}(-t)}{\left|\phi_{\widehat{w}}(t)\right|^2\vee\gamma t^a}\widehat{\phi}_{\widehat{X}}(t)e^{-\jmath tx}\right)dt\right]\\
&\overset{④}{\leq}\frac{1}{n\pi^2}\mathbb{E}\left[\int_0^\infty\frac{1}{t}\Im\left(\frac{\phi_{\widehat{w}}(-t)}{\left|\phi_{\widehat{w}}(t)\right|^2\vee\gamma t^a}e^{\jmath t(\widehat{X}-x)}\right)dt\right]^2\leq\frac{2}{n\pi^2}\left(K_1+K_2\right),
\end{aligned}$$

where in ④ we use the bound $\text{Var}\left(\cdot\right)\leq\mathbb{E}(\cdot)^2$, and the terms K_1 and K_2 are defined as

$$\begin{aligned}
K_1&\triangleq\mathbb{E}\left[\int_0^{t_\perp}\frac{1}{t}\Im\left(\frac{\phi_{\widehat{w}}(-t)}{\left|\phi_{\widehat{w}}(t)\right|^2\vee\gamma t^a}e^{\jmath t(\widehat{X}-x)}\right)dt\right]^2,\\
K_2&\triangleq\mathbb{E}\left[\int_{t_\perp}^{\infty}\frac{1}{t}\Im\left(\frac{\phi_{\widehat{w}}(-t)}{\left|\phi_{\widehat{w}}(t)\right|^2\vee\gamma t^a}e^{\jmath t(\widehat{X}-x)}\right)dt\right]^2,
\end{aligned}$$

and $t_\perp$ is defined as $(2\gamma)^{-1/a}$. First, we bound K_1 as

$$\begin{aligned}
K_1&=\mathbb{E}\left|\int_0^{t_\perp}\frac{1}{t}\Im\left(\phi_{\widehat{w}}(-t)e^{\jmath t(\widehat{X}-x)}\left(1+\sum_{k=1}^{\infty}\left(1-\left(\left|\phi_{\widehat{w}}(t)\right|^2\vee\gamma t^a\right)\right)^k\right)\right)dt\right|^2\\
&\lesssim\mathbb{E}\left|\int_0^{t_\perp}\frac{1}{t}\Im\left(\phi_{\widehat{w}}(-t)e^{\jmath t(\widehat{X}-x)}\right)dt\right|^2+\mathbb{E}\left|\int_0^{t_\perp}\frac{1}{t}\sum_{k=1}^{\infty}\left(1-\left(\left|\phi_{\widehat{w}}(t)\right|^2\vee\gamma t^a\right)\right)^k dt\right|^2\\
&\overset{⑤}{\leq}\mathbb{E}\sup_{\tau>0}\left|\int_0^{\tau}\frac{\sin t}{t}dt\right|^2+\mathbb{E}\left|\int_0^{t_\perp}\frac{1}{t}\sum_{k=1}^{\infty}\left(1-\left|\phi_{\widehat{w}}(t)\right|^2\right)^k dt\right|^2\\
&\leq 3+\mathbb{E}\left[\int_0^{t_\perp}\frac{1}{t}\sum_{k=1}^{\infty}\left(\frac{\sigma^2t^2}{d-p}\right)^k\right]^2\overset{⑥}{\lesssim}1,
\end{aligned} \tag{44}$$

where in ⑤ we use the fact $\left|\phi_{\widehat{w}}(t)\right|^2 \vee \delta t^a \geq 1 - \frac{c_1\sigma^2 t^2}{d-p}$ for $t \in (0, t_\perp)$, and in ⑥ we use $\sigma^2 t_\perp^2 \asymp \sqrt{d-p}(\log(np))^{-\frac{2}{a}} \leq \frac{d-p}{2}$.

Then we expand the term K_2 as a product of two terms reading as

$$\begin{aligned} K_2 = & \ \mathbb{E}\left[\int_{t_\perp}^{\infty}\int_{t_\perp}^{\infty} \frac{1}{uv}\Im\left(\frac{\phi_{\widehat{w}}(-u)\phi_{\widehat{w}}(-v)e^{\mathbb{j}(u+v)(\widehat{X}-x)}}{\left[\left|\phi_{\widehat{w}}(u)\right|^2 \vee \gamma u^a\right]\left[\left|\phi_{\widehat{w}}(v)\right|^2 \vee \gamma v^a\right]}\right) dudv\right] \\ \leq & \ \mathbb{E}\left[\int_{t_\perp}^{\infty}\int_{t_\perp}^{\infty} \frac{\phi_{\widehat{w}}(-u)\phi_{\widehat{w}}(-v)}{uv} \cdot \frac{1}{\left|\phi_{\widehat{w}}(u)\right|^2 \vee \gamma u^a} \cdot \frac{1}{\left|\phi_{\widehat{w}}(v)\right|^2 \vee \gamma v^a} dudv\right] \\ \leq & \ \frac{1}{\gamma^2}\mathbb{E}\left[\int_{t_\perp}^{\infty} \frac{1}{t^{a+1}}dt\right]^2 = \frac{1}{a^2\gamma^2 t_\perp^{2a}} \asymp 1. \end{aligned} \tag{45}$$

Combining equation 44 and equation 45 generates

$$\mathrm{Var}\widetilde{F}(x) \lesssim n^{-1}. \tag{46}$$

And the proof is completed by combining equation 43 and equation 46. □

D Analysis of the Graphical Structure Model

Denote the empirical covariance matrix $\mathbf{\Sigma}_n$ as

$$\mathbf{\Sigma}_n^{\text{non-param}} \triangleq \frac{1}{n}\sum_{s=1}^{n} \boldsymbol{h}(\mathbf{X}^{(s)})\boldsymbol{h}(\mathbf{X}^{(s)})^\top - \left(\frac{1}{n}\sum_{s=1}^{n}\boldsymbol{h}(\mathbf{X}^{(s)})\right)\left(\frac{1}{n}\sum_{s=1}^{n}\boldsymbol{h}(\mathbf{X}^{(s)})\right)^\top,$$

where $h(\cdot)$ denotes the oracle estimator of the transform functions in Def. 1. We first analyze the estimation error of the covariance matrix $\mathbf{\Sigma}_n^{\text{non-param}}$ in terms of the infinity norm ℓ_∞.

Theorem 7. *Under the Assumption 2 and Assumption 3, we have*

$$\left\|\mathbf{\Sigma}_n^{\text{non-param}} - \widehat{\mathbf{\Sigma}}_n^{\text{non-param}}\right\|_\infty \lesssim \sqrt{\log n \vee \log(d-p)}\left(\frac{\sqrt{\log n}}{n^{1/4}} \vee \frac{\sqrt{\log(d-p)}}{(d-p)^{\beta/4}}\right),$$

with probability exceeding $1 - o(1)$, where $\delta_{n,d,p}$ is set as equation 15, β is defined as $\frac{1}{2} \wedge \frac{a}{4} \wedge \frac{2\alpha+1}{4}$, and $\widehat{\mathbf{\Sigma}}_n^{\text{param}}$ is defined in equation 10.

The proof largely follows Liu et al. (2009). However, we cannot directly access the samples $\left\{\mathbf{X}^{(s)}\right\}_{1\leq s\leq n}$ and have to use the perturbed samples $\{\widehat{\mathbf{X}}^{(s)}\}_{1\leq s\leq n}$ instead. This will lead to additional errors in estimating the covariance matrix and how to bound these errors constitutes the technical bottleneck.

D.1 Notations

We assume that the correct estimation of $m_i = 0$ and $v_i = 1$ w.l.o.g. Let $h_i(x) = \Phi^{-1}\left(F_i(x)\right)$, where $(\cdot)^{-1}$ denotes the inverse of the function. For the conciseness of the notation, we define $\widehat{\varphi}_i^{(s)}$, $\widetilde{\varphi}_i^{(s)}$, and $\varphi_i^{(s)}$ as

$$\begin{aligned} \widehat{\varphi}_i^{(s)} &= \widehat{h}_i(\widehat{X}_i^{(s)}); \\ \widetilde{\varphi}_i^{(s)} &= \widehat{h}_i(X_i^{(s)}); \\ \varphi_i^{(s)} &= h_i(X_i^{(s)}). \end{aligned}$$

The (i,j)th entries of the corresponding covariance matrices $\widehat{\mathbf{\Sigma}}_n^{\text{non-param}}$, $\widetilde{\mathbf{\Sigma}}_n^{\text{non-param}}$, $\mathbf{\Sigma}_n^{\text{non-param}}$ are written as

$$\left(\widehat{\mathbf{\Sigma}}_n^{\text{non-param}}\right)_{i,j} = \frac{1}{n}\sum_{s=1}^{n}\widehat{\varphi}_i^{(s)}\widehat{\varphi}_j^{(s)} - \widehat{\mu}_i\widehat{\mu}_j;$$
$$\left(\widetilde{\mathbf{\Sigma}}_n^{\text{non-param}}\right)_{i,j} = \frac{1}{n}\sum_{s=1}^{n}\widetilde{\varphi}_i^{(s)}\widetilde{\varphi}_j^{(s)} - \widetilde{\mu}_i\widetilde{\mu}_j;$$
$$(\mathbf{\Sigma}_n^{\text{non-param}})_{i,j} = \frac{1}{n}\sum_{s=1}^{n}\varphi_i^{(s)}\varphi_j^{(s)} - \mu_i\mu_j.$$

Moreover, we define two regions $\mathcal{R}_E$ and $\mathcal{R}_M$ as

$$\mathcal{R}_E \triangleq \left[-c_U\sqrt{\log(n\vee(d-p))}, -c_L\sqrt{\log n\vee\log(d-p)}\right)\bigcup\left(c_L\sqrt{\log(n\vee(d-p))}, c_U\sqrt{\log n\vee\log(d-p)}\right];$$
$$\mathcal{R}_M \triangleq \left[-c_L\sqrt{\log(n\vee(d-p))}, c_L\sqrt{\log(n\vee(d-p))}\right]. \tag{47}$$

D.2 Proof of Thm. 7

Proof. We bound the deviation between $\mathbf{\Sigma}_n^{\text{non-param}}$ and $\widehat{\mathbf{\Sigma}}_n^{\text{non-param}}$ as

$$\left\|\mathbf{\Sigma}_n^{\text{non-param}} - \widehat{\mathbf{\Sigma}}_n^{\text{non-param}}\right\|_\infty \le \underbrace{\left\|\mathbf{\Sigma}_n^{\text{non-param}} - \widetilde{\mathbf{\Sigma}}_n^{\text{non-param}}\right\|_\infty}_{\triangleq T_1} + \underbrace{\left\|\widetilde{\mathbf{\Sigma}}_n^{\text{non-param}} - \widehat{\mathbf{\Sigma}}_n^{\text{non-param}}\right\|_\infty}_{\triangleq T_2}.$$

Step I. For the first term T_1, we invoke the triangle inequality and have

$$\left\|\widetilde{\mathbf{\Sigma}}^{\text{non-param}} - \mathbf{\Sigma}^{\natural}\right\|_\infty \le \max_{i,j}\frac{1}{n}\left|\sum_{s=1}^{n}\widetilde{\varphi}_i^{(s)}\widetilde{\varphi}_j^{(s)} - \varphi_i^{(s)}\varphi_j^{(s)}\right| + \|\mu_i\mu_j - \widetilde{\mu}_i\widetilde{\mu}_j\|_\infty.$$

Following a similar strategy that is used in Liu et al. (2009), we focus on the first term as the second term is of higher order.

We bound the value of $\max_{i,j} n^{-1}\left|\sum_{s=1}^{n}\widetilde{\varphi}_i^{(s)}\widetilde{\varphi}_j^{(s)} - \varphi_i^{(s)}\varphi_j^{(s)}\right|$ as

$$\begin{aligned}
&\mathbb{P}\left(\max_{i,j} n^{-1}\left|\sum_{s=1}^{n}\widetilde{\varphi}_i^{(s)}\widetilde{\varphi}_j^{(s)} - \varphi_i^{(s)}\varphi_j^{(s)}\right| \ge \vartheta\right)\\
&\le p^2\mathbb{E}\left[\left(n^{-1}\left|\sum_{s=1}^{n}\widetilde{\varphi}_i^{(s)}\widetilde{\varphi}_j^{(s)} - \varphi_i^{(s)}\varphi_j^{(s)}\right| \ge \vartheta\right)\mathbb{1}\left(\varphi_i^{(s)}\in\mathcal{R}_E\bigcup\mathcal{R}_M, \forall\, 1\le s\le n, 1\le i\le p\right)\right]\\
&+ np\mathbb{E}\mathbb{1}\left(\varphi_i^{(s)}\notin\mathcal{R}_E\bigcap\mathcal{R}_M,\ \exists\, 1\le s\le n, 1\le i\le p\right).
\end{aligned}$$

Following a classical procedure as in Boucheron et al. (2013), we can show the second probability $\mathbb{E}\mathbb{1}\left(\varphi_i^{(s)}\notin\mathcal{R}_E\bigcap\mathcal{R}_M\right)$ is no greater than $e^{-c_0(n\wedge(d-p))}$. For the conciseness of notation, we define the deviation $\delta_{i,j}^{(s)}$ as

$$\delta_{i,j}^{(s)} = \widetilde{\varphi}_i^{(s)}\widetilde{\varphi}_j^{(s)} - \varphi_i^{(s)}\varphi_j^{(s)}.$$

Then the summary $n^{-1}\left(\sum_{s=1}^{n}\widetilde{\varphi}_i^{(s)}\widetilde{\varphi}_j^{(s)} - \varphi_i^{(s)}\varphi_j^{(s)}\right)\mathbb{1}\left(\varphi_i^{(s)}\in\mathcal{R}_E\bigcup\mathcal{R}_M, \forall\, 1\le s\le n, 1\le i\le p\right)$ can be decomposed as

$$\frac{1}{n}\sum_{s=1}^{n}\delta_{i,j}^{(s)} = \underbrace{\frac{1}{n}\sum_{s=1}^{n}\delta_{i,j}^{(s)}\mathbb{1}[\mathcal{C}_1^{(s)}(i,j)]}_{\triangleq T_{1,1}} + \underbrace{\frac{1}{n}\sum_{s=1}^{n}\delta_{i,j}^{(s)}\mathbb{1}[\mathcal{C}_2^{(s)}(i,j)]}_{\triangleq T_{1,2}} + \underbrace{\frac{2}{n}\sum_{s=1}^{n}\delta_{i,j}^{(s)}\mathbb{1}[\mathcal{C}_3^{(s)}(i,j)]}_{\triangleq T_{1,3}},$$

where the events $\mathcal{C}_1^{(s)}(i,j)$, $\mathcal{C}_2^{(s)}(i,j)$, and $\mathcal{C}_3^{(s)}(i,j)$ are defined as

$$\begin{aligned}
\mathcal{C}_1^{(s)}(i,j) &\triangleq \left\{\varphi_i^{(s)} \in \mathcal{R}_E, \varphi_j^{(s)} \in \mathcal{R}_E\right\};\\
\mathcal{C}_2^{(s)}(i,j) &\triangleq \left\{\varphi_i^{(s)} \in \mathcal{R}_M, \varphi_j^{(s)} \in \mathcal{R}_M\right\};\\
\mathcal{C}_3^{(s)}(i,j) &\triangleq \left\{\varphi_i^{(s)} \in \mathcal{R}_E, \varphi_j^{(s)} \in \mathcal{R}_M\right\},
\end{aligned}$$

respectively, where the definitions of $\mathcal{R}_E$ and $\mathcal{R}_M$ can be found in equation 47.

In the following, we will separately bound the three terms and show $n^{-1}\left|\sum_{s=1}^n \delta_{i,j}^{(s)}\right| \lesssim \vartheta \triangleq \vartheta_1 \vee \vartheta_2$, where the quantities ϑ_1 and ϑ_2 are defined in equation 48 and equation 50, respectively. The analysis of the first term $T_{1,1}$ and second term $T_{1,2}$ is deferred to Lemma 25 and Lemma 26, respectively; while that of the third term $T_{1,3}$ is omitted due to their similarities of Lemma 25 and Lemma 26.

Step II. The second term T_2 is upper-bounded in Lemma 27. The analysis is in the same spirit as the above procedure but requires some modifications. □

Lemma 25. *We have*

$$\frac{1}{n}\left|\sum_{s=1}^n \delta_{i,j}^{(s)} \mathbb{1}[\mathcal{C}_1^{(s)}(i,j)]\right| \le 2c_0\left[n^{-c_1} \vee (d-p)^{-c_1}\right](\log n \vee \log(d-p))^{\frac{3}{2}} \triangleq \vartheta_1, \tag{48}$$

with probability exceeding $1 - c_2 n^{-c_3} \wedge (d-p)^{-c_4} - c_5 n^{-c_6}$*, where the parameters* c_i *are some fixed constant,* $0 \le i \le 6$.

Proof. Invoking the union bound, we obtain

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{n}\left|\sum_{s=1}^n \delta_{i,j}^{(s)} \mathbb{1}[\mathcal{C}_1^{(s)}(i,j)]\right| \gtrsim 2c_0\left[n^{-c_1} \vee (d-p)^{-c_1}\right](\log n \vee \log(d-p))^{\frac{3}{2}}\right)\\
\overset{①}{\le}\ & \underbrace{\mathbb{E}\mathbb{1}\left(\max_s \left|\delta_{i,j}^{(s)}\right| \gtrsim (\log n \vee \log(d-p))\right)\mathbb{1}[\mathcal{C}_1^{(s)}(i,j)]}_{P_1} + \underbrace{\mathbb{P}\left(\frac{1}{n}\sum_{s=1}^n \mathbb{1}[\mathcal{C}_1^{(s)}(i,j)] \ge \vartheta\right)}_{P_2},
\end{aligned}$$

where ϑ is defined as

$$\vartheta = 2c_0\left[n^{-c_1} \vee (d-p)^{-c_1}\right]\sqrt{\log n \vee \log(d-p)}, \tag{49}$$

and in ① we use the union bound.

With the relation

$$\left|\delta_{i,j}^{(s)}\right| \le \left|\left(\widetilde{\varphi}_i^{(s)} - \varphi_i^{(s)}\right)\left(\widetilde{\varphi}_j^{(s)} - \varphi_j^{(s)}\right)\right| + \left|\varphi_i^{(s)}\left(\widetilde{\varphi}_j^{(s)} - \varphi_j^{(s)}\right)\right| + \left|\varphi_j^{(s)}\left(\widetilde{\varphi}_i^{(s)} - \varphi_i^{(s)}\right)\right|,$$

we can upper bound the probability P_1 as

$$P_1 \le np\mathbb{E}\mathbb{1}\left(\varphi_i^{(s)} \le c_U\sqrt{\log p \vee \log(d-p)}\mathbb{1}[\mathcal{C}_1^{(s)}(i,j)],\ \widetilde{\varphi}_i^{(s)} \lesssim c_U\sqrt{\log n \vee \log(d-p)}\right) \overset{②}{\lesssim} n^{-c} \wedge (d-p)^{-c},$$

where ② is due to Lemma 28.

While for probability P_2, we have

$$
\begin{aligned}
&\mathbb{P}\left(\sum_{s=1}^{n}\mathbb{1}[\mathcal{C}_1^{(s)}(i,j)]\geq n\vartheta\right)\leq\mathbb{P}\left(\frac{1}{n}\sum_{s=1}^{n}\mathbb{1}[\varphi_i^{(s)}\in\mathcal{R}_E]\geq\vartheta\right)\\
&=\mathbb{P}\left(\frac{1}{n}\sum_{s=1}^{n}\left[\mathbb{1}[\varphi_i^{(s)}\in\mathcal{R}_E]-\mathbb{E}\mathbb{1}\left(\varphi_i^{(s)}\in\mathcal{R}_E\right)\right]\geq\ \vartheta-\frac{1}{n}\left[\sum_{s=1}^{n}\mathbb{E}\mathbb{1}\left(\varphi_i^{(s)}\in\mathcal{R}_E\right)\right]\right)\\
&\overset{③}{\leq}\ \exp\left(-\frac{n}{2}\left[\vartheta-\mathbb{P}\left(\varphi_i^{(s)}\in\mathcal{R}_E\right)\right]^2\right),
\end{aligned}
$$

where in ③ we use the Hoeffding's inequality (cf. Thm. 2.8 in Boucheron et al. (2013)). Notice that the probability $\mathbb{P}\left(\varphi_i^{(s)}\in\mathcal{R}_E\right)$ can be bounded as

$$
\begin{aligned}
\mathbb{P}\left(\varphi_i^{(s)}\in\mathcal{R}_E\right)&=\frac{2}{\sqrt{2\pi}}\int_{c_L\sqrt{\log n\vee\log(d-p)}}^{c_U\sqrt{\log n\vee\log(d-p)}}e^{-t^2/2}dt\lesssim\sqrt{\frac{2}{\pi}}e^{-\frac{c_L}{2}(\log n\vee\log(d-p))}\sqrt{\log n\vee\log(d-p)}\\
&\leq c_0\left[n^{-c_1}\vee(d-p)^{-c_1}\right]\sqrt{\log n\vee\log(d-p)},
\end{aligned}
$$

where $0<c_1<1/2$ is some fixed positive constant. Recalling the value of ϑ in equation 49, we have

$$
\mathbb{P}\left(\sum_{s=1}^{n}\mathbb{1}[\mathcal{C}_1^{(s)}(i,j)]\geq n\vartheta\right)\leq\exp\left(-c_0n^{1-2c_1}\log n\right)\ll n^{-c},
$$

and complete the proof.

□

Lemma 26. *We have*

$$
\frac{1}{n}\left|\sum_{s=1}^{n}\delta_{i,j}^{(s)}\mathbb{1}\left(C_2^{(s)}(i,j)\right)\right|\lesssim\frac{\sqrt{\log n}}{n^{1/4}}\vee\frac{\sqrt{\log(d-p)}}{(d-p)^{\beta/4}}\triangleq\vartheta_2, \tag{50}
$$

with probability exceeding the probability $1-p^2\exp\left(-\frac{c_0\sqrt{n}}{\log^2 n}-\frac{c_1n\log^4(np)}{\log(d-p)(d-p)^{\beta/2}}\right)-p^2n^{-c_3}-2p^3e^{-c_4p}-4n^{-c_5}p^{-c_6}$, *where* β *is defined as* $\frac{1}{2}\wedge\frac{a}{4}\wedge\frac{2\alpha+1}{4}$.

Proof. For each term $\delta_{i,j}^{(s)}$, we can decompose it as

$$
\delta_{i,j}^{(s)}=\left(\widetilde{\varphi}_i^{(s)}-\varphi_i^{(s)}\right)\left(\widetilde{\varphi}_j^{(s)}-\varphi_j^{(s)}\right)+\varphi_i^{(s)}\left(\widetilde{\varphi}_j^{(s)}-\varphi_j^{(s)}\right)+\varphi_j^{(s)}\left(\widetilde{\varphi}_i^{(s)}-\varphi_i^{(s)}\right).
$$

Then we can decompose the summary as

$$
\begin{aligned}
\frac{1}{n}\left|\sum_{s=1}^{n}\delta_{i,j}^{(s)}\mathbb{1}\left(C_2^{(s)}(i,j)\right)\right|\ &\leq\frac{1}{n}\sum_{s=1}^{n}\left|\left(\widehat{\varphi}_i^{(s)}-\widetilde{\varphi}_i^{(s)}\right)\left(\widehat{\varphi}_j^{(s)}-\widetilde{\varphi}_j^{(s)}\right)\mathbb{1}\left(C_2^{(s)}(i,j)\right)\right|\\
&+\frac{1}{n}\sum_{s=1}^{n}\left|\widetilde{\varphi}_i^{(s)}\left(\widehat{\varphi}_j^{(s)}-\widetilde{\varphi}_j^{(s)}\right)\mathbb{1}\left(C_2^{(s)}(i,j)\right)\right|\\
&+\frac{1}{n}\sum_{s=1}^{n}\left|\widetilde{\varphi}_j^{(s)}\left(\widehat{\varphi}_i^{(s)}-\widetilde{\varphi}_i^{(s)}\right)\mathbb{1}\left(C_2^{(s)}(i,j)\right)\right|.
\end{aligned}
$$

Our next goal is to investigate the behavior of $\sup_i\left|\widetilde{\varphi}_i^{(s)}-\widehat{\varphi}_i^{(s)}\right|\mathbb{1}\left[C_2^{(s)}(i,j)\right]$. Define the an event $\mathcal{E}_F(\cdot)$ as

$$
\mathcal{E}_F(i)=\left\{\delta_{n,d,p}\leq\widehat{F}_i\leq1-\delta_{n,d,p}\right\},\quad 1\leq i\leq n.
$$

We have

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{s=1}^{n}\delta_{i,j}^{(s)}\mathbb{1}\left(C_2^{(s)}(i,j)\right)\right|\geq\vartheta_2\right)\leq\mathbb{P}\left(\sup_i\left|\widetilde{\varphi}_i^{(s)}-\widehat{\varphi}_i^{(s)}\right|\mathbb{1}\left[C_2^{(s)}(i,j)\right]\geq\vartheta_2\right)$$
$$\leq p^2\underbrace{\mathbb{E}\mathbb{1}\left(\mathbb{1}\left(C_2^{(s)}(i,j)\right)\overline{\mathcal{E}}_F(i)\right)}_{T_1}+p^2\underbrace{\mathbb{E}\mathbb{1}\left(\left(\left|\widetilde{\varphi}_i^{(s)}-\widehat{\varphi}_i^{(s)}\right|\mathbb{1}\left(C_2^{(s)}(i,j)\right)\geq\vartheta_2\right)\bigcup\mathcal{E}_F(i)\right)}_{T_2}.$$

Easily we can verify that $\delta_{n,d,p}$ satisfy the relation

$$2\delta_{n,d,p}\leq 1-\Phi\left(c_L\sqrt{\log n\vee\log(d-p)}\right)-\sqrt{\varepsilon_x},$$

where ε_x is defined in equation 13. Invoking Lemma 29, we can bound term T_1 as

$$T_1\leq 2\exp\left(-2n\left(1-\delta_{n,d,p}-\Phi\left(c_L\sqrt{\log n\vee\log(d-p)}\right)-\sqrt{\varepsilon_x}\right)^2\right)\leq 2\exp(-2n\delta_{n,d,p}^2)$$
$$\lesssim 2\exp\left(-\frac{c_0\sqrt{n}}{\log^2 n}-\frac{c_1 n\log^4(np)}{\log(d-p)(d-p)^{\beta/2}}\right).$$

Conditional on event $\mathcal{E}_F(i)$, we study the term T_1 by investigating the difference $\widetilde{\varphi}_i^{(s)}-\widehat{\varphi}_i^{(s)}$. We assume $\widehat{F}_i^{\text{tr}}(X_i^{(s)})\geq F_i(X_i^{(s)})$ w.l.o.g. According to the mean value theorem, we have

$$\widetilde{\varphi}_i^{(s)}-\varphi_i^{(s)}=\Phi^{-1}\left(\widehat{F}_i^{\text{tr}}(X_i^{(s)})\right)-\Phi^{-1}\left(F_i(X_i^{(s)})\right)=\left(\Phi^{-1}\right)'(\xi)\left|\widehat{F}_i^{\text{tr}}(X_i^{(s)})-F_i(X_i^{(s)})\right|,$$

where ξ is some point such that $F_i(X_i^{(s)})\leq\xi\leq\widehat{F}_i^{\text{tr}}(X_i^{(s)})$. Due to the fact that $X_i^{(s)}\in\mathcal{R}_M$ and conditional on event $\mathcal{E}_F(i)$, we conclude $\delta_n\leq\xi\leq 1-\delta_n$ and hence

$$\left|\left(\Phi^{-1}(\xi)\right)'\right|\leq\left|\left(\Phi^{-1}(1-\delta_{n,d,p})\right)'\right|\vee\left|\left(\Phi^{-1}(\delta_{n,d,p})\right)'\right|=\left|\left(\Phi^{-1}(1-\delta_{n,d,p})\right)'\right|$$
$$=\frac{1}{\phi\left(\Phi^{-1}(1-\delta_{n,d,p})\right)}\leq\exp\left[\frac{1}{2}\left(\Phi^{-1}(1-\delta_{n,d,p})\right)^2\right]\overset{①}{\leq}\frac{1}{\delta_{n,d,p}},$$

where ① is due to Lemma 31 and the fact that $\delta_{n,d,p}\to 0$. Set ϑ_2 such that

$$\vartheta_2\delta_{n,d,p}\geq(\log n)\varepsilon_x+\frac{c_0}{\sqrt{n}}+c_1\sqrt{\varepsilon_x},$$

where ε_x is defined in equation 13. We invoke Thm. 4 and conclude

$$T_1\leq\mathbb{P}\left(\left|\widehat{F}_i^{\text{tr}}(X_i^{(s)})-F_i(X_i^{(s)})\right|\geq\delta_{n,d,p}\vartheta_2\right)\leq n^{-c_3}+2pe^{-c_4 p}+4n^{-c_5}p^{-c_6}.$$

Recalling the definition of $\delta_{n,d,p}$ in equation 15, we conclude ϑ_2 to be approximately

$$\vartheta_2\asymp\frac{\sqrt{\log n}}{n^{1/4}}\vee\frac{\sqrt{\log(d-p)}}{(d-p)^{\beta/4}},$$

where β is defined as $\frac{1}{2}\wedge\frac{a}{4}\wedge\frac{2\alpha+1}{4}$.

□

Lemma 27. *Under the Assumption 3, we have*

$$\left\|\widehat{\mathbf{\Sigma}}_n^{\text{non-param}}-\widetilde{\mathbf{\Sigma}}_n^{\text{non-param}}\right\|_\infty\lesssim\sqrt{\log n\vee\log(d-p)}\left(\frac{\sqrt{\log n}}{n^{1/4}}\vee\frac{\sqrt{\log(d-p)}}{(d-p)^{\beta/4}}\right)+\frac{L\sigma\left(\log n\vee\log(d-p)\right)}{c_0\log^2(np)(d-p)^{1/4}}\triangleq\vartheta_3, \tag{51}$$

with probability exceeding $1-e^{-c_0(n\wedge(d-p))}-e^{-c_1 n}$.

Proof. The proof conditions on the event

$$\left|\widetilde{\varphi}_i^{(s)}\right| \lesssim c_U \sqrt{\log n \vee \log(d-p)},$$

holds for all s and i with probability exceeding $1 - e^{-c_0(n \wedge (d-p))}$.

Following the same argument as in Thm. 7, our analysis focus on the error $n^{-1}\sum_{s=1}^n \left(\widehat{\varphi}_i^{(s)}\widehat{\varphi}_j^{(s)} - \varphi_i^{(s)}\varphi_j^{(s)}\right)$. Adopting the decomposition such that

$$\widehat{\varphi}_i^{(s)}\widehat{\varphi}_j^{(s)} - \widetilde{\varphi}_i^{(s)}\widetilde{\varphi}_j^{(s)} = \widetilde{\varphi}_i^{(s)}\left(\widehat{\varphi}_j^{(s)} - \widetilde{\varphi}_j^{(s)}\right) + \widetilde{\varphi}_j^{(s)}\left(\widehat{\varphi}_i^{(s)} - \widetilde{\varphi}_i^{(s)}\right) + \left(\widehat{\varphi}_i^{(s)} - \widetilde{\varphi}_i^{(s)}\right)\left(\widehat{\varphi}_j^{(s)} - \widetilde{\varphi}_j^{(s)}\right),$$

we have

$$\begin{aligned}\frac{1}{n}\sum_{s=1}^n \left(\widehat{\varphi}_i^{(s)}\widehat{\varphi}_j^{(s)} - \varphi_i^{(s)}\varphi_j^{(s)}\right) = & \underbrace{\frac{1}{n}\sum_{s=1}^n \widetilde{\varphi}_i^{(s)}\left(\widehat{\varphi}_j^{(s)} - \widetilde{\varphi}_j^{(s)}\right)}_{T_1} + \underbrace{\frac{1}{n}\sum_{s=1}^n \widetilde{\varphi}_j^{(s)}\left(\widehat{\varphi}_i^{(s)} - \widetilde{\varphi}_i^{(s)}\right)}_{T_2} \\ & + \underbrace{\frac{1}{n}\sum_{s=1}^n \left(\widehat{\varphi}_i^{(s)} - \widetilde{\varphi}_i^{(s)}\right)\left(\widehat{\varphi}_j^{(s)} - \widetilde{\varphi}_j^{(s)}\right)}_{T_3}.\end{aligned}$$

We only need to analyze the behavior of the terms T_1 and T_2, since the term T_3 is of higher order. Conditional on the event discussed in Thm. 4, we have

$$\begin{aligned}\left|\widehat{\varphi}_i^{(s)} - \widetilde{\varphi}_i^{(s)}\right| &= \left|\left(\Phi^{-1}(\xi)\right)'\right|\left|\widehat{F}^{\mathrm{tr}}(\widehat{X}_i^{(s)}) - \widehat{F}^{\mathrm{tr}}(X_i^{(s)})\right| \leq \frac{1}{\delta_{n,d,p}}\left|\widehat{F}^{\mathrm{tr}}(\widehat{X}_i^{(s)}) - \widehat{F}^{\mathrm{tr}}(X_i^{(s)})\right| \\ &\leq \frac{1}{\delta_{n,d,p}}\left[\left|\widehat{F}^{\mathrm{tr}}(\widehat{X}_i^{(s)}) - F(\widehat{X}_i^{(s)})\right| + \left|F(\widehat{X}_i^{(s)}) - F(X_i^{(s)})\right| + \left|F(X_i^{(s)}) - \widehat{F}^{\mathrm{tr}}(X_i^{(s)})\right|\right] \\ &\overset{①}{\leq} \frac{2}{\delta_{n,d,p}}\left[(\log n)\varepsilon_x + \frac{c_0}{\sqrt{n}} + c_1\sqrt{\varepsilon_x}\right] + \frac{L\left|\widehat{w}_i^{(s)}\right|}{\delta_{n,d,p}},\end{aligned}$$

where ① is because of the Lipschitz property in Assumption 3. Then we obtain

$$\begin{aligned}T_1 &\lesssim \underbrace{\frac{2\sqrt{\log n \vee \log(d-p)}}{\delta_{n,d,p}}\left[(\log n)\varepsilon_x + \frac{c_0}{\sqrt{n}} + c_1\sqrt{\varepsilon_x}\right]}_{T_{1,1}} + \underbrace{\frac{L\sqrt{\log n \vee \log(d-p)}}{n\delta_{n,d,p}}\sum_{s=1}^n |\widehat{w}_i^{(s)}|}_{T_{1,2}} \\ &\overset{②}{\lesssim} 2\sqrt{\log n \vee \log(d-p)}\left(\frac{\sqrt{\log n}}{n^{1/4}} \vee \frac{\sqrt{\log(d-p)}}{(d-p)^{\beta/4}}\right) + \frac{L\sqrt{\log n \vee \log(d-p)}}{\delta_{n,d,p}}\sqrt{\frac{\sum_{s=1}^n |\widehat{w}_i^{(s)}|^2}{n}},\end{aligned}$$

where in ② we plug in the definition of ε_x in equation 13, that is,

$$T_{1,1} \asymp 2\sqrt{\log n \vee \log(d-p)}\left(\frac{\sqrt{\log n}}{n^{1/4}} \vee \frac{\sqrt{\log(d-p)}}{(d-p)^{\beta/4}}\right).$$

Since the $\widehat{w}_i^{(s)}$ is approximately Gaussian distribution with mean zero and $\sigma^2/(d-p)$ variance, we have $(d-p)/\sigma^2\sum_{s=1}^n |\widehat{w}_i^{(s)}|^2$ be ξ^2-RV with freedom n, which means $\sum_{s=1}^n |\widehat{w}_i^{(s)}|^2 \leq 2n\sigma^2/(d-p)$ holds with probability at least $1 - e^{-cn}$. To sum up, we obtain

$$\begin{aligned}T_1 &\lesssim \sqrt{\log n \vee \log(d-p)}\left(\frac{\sqrt{\log n}}{n^{1/4}} \vee \frac{\sqrt{\log(d-p)}}{(d-p)^{\beta/4}}\right) + \frac{L\sigma\sqrt{\log n \vee \log(d-p)}}{\delta_{n,d,p}\sqrt{d-p}} \\ &\lesssim \sqrt{\log n \vee \log(d-p)}\left(\frac{\sqrt{\log n}}{n^{1/4}} \vee \frac{\sqrt{\log(d-p)}}{(d-p)^{\beta/4}}\right) + \frac{L\sigma\left(\log n \vee \log(d-p)\right)}{c_0\log^2(np)(d-p)^{1/4}},\end{aligned}$$

and complete the proof.

□

D.3 Supporting Lemmas

Lemma 28. *For all possible i ($1 \le i \le n$), we conclude*

$$\sup_{t \in \mathcal{R}_E} \left| \Phi^{-1}\left(\widehat{F}_i^{\mathrm{tr}}(t)\right) - \Phi^{-1}\left(F_i(t)\right) \right| \le 2c_U \sqrt{\log n \vee \log(d-p)},$$

where $\mathcal{R}_E$ is defined in equation 47.

Proof. We conclude that

$$\left| \Phi^{-1}\left(\widehat{F}_i^{\mathrm{tr}}(t)\right) - \Phi^{-1}\left(F_i(t)\right) \right| \le \left| \Phi^{-1}\left(F_i(t)\right) \right| + \left| \Phi^{-1}\left(\widehat{F}_i^{\mathrm{tr}}(t)\right) \right|.$$

For the first term, we have

$$\left| \Phi^{-1}\left(F_i(t)\right) \right| \le c_U \sqrt{\log n \vee \log(d-p)},$$

according to the definitions of $\mathcal{R}_E$. Meanwhile for the second term $\left| \Phi^{-1}\left(\widehat{F}_i^{\mathrm{tr}}(t)\right) \right|$, we have

$$\Phi^{-1}\left(\widehat{F}_i^{\mathrm{tr}}(t)\right) \overset{①}{\le} \Phi^{-1}\left(1-\delta_{n,d,p}\right) \overset{②}{\le} \sqrt{2\log\frac{1}{1-\delta_{n,d,p}}} \overset{③}{\le} c_U \sqrt{\log n \vee \log(d-p)},$$

where in ① we exploit the fact $\widehat{F}_i^{\mathrm{tr}}(t) \le 1-\delta_{n,d,p}$, in ② we invoke Lemma 11 in Liu et al. (2009) (cf. Lemma 31), and in ③ we use the fact that $\delta_{n,d,p} \ge \frac{1}{n} \vee \frac{1}{d-p}$.

□

Lemma 29. *Provided that $\delta_{n,d,p} \le 1 - \Phi\left(c_L\sqrt{\log n \vee \log(d-p)}\right) - \sqrt{\varepsilon_x}$, we can bound the probability*

$$\mathbb{P}\left(\widehat{F}_i(g_i(c_L\sqrt{\log n \vee \log(d-p)})) \ge 1-\delta_{n,d,p}\right) \le \exp\left(-2n\left(1-\delta_{n,d,p}-\Phi\left(c_L\sqrt{\log n \vee \log(d-p)}\right)-\sqrt{\varepsilon_x}\right)^2\right);$$

$$\mathbb{P}\left(\widehat{F}_i(g_i(-c_L\sqrt{\log n \vee \log(d-p)})) \le \delta_{n,d,p}\right) \le \exp\left(-2n\left(1-\delta_{n,d,p}-\Phi\left(c_L\sqrt{\log n \vee \log(d-p)}\right)-\sqrt{\varepsilon_x}\right)^2\right),$$

where ε_x is defined in equation 13.

Proof. We have

$$\begin{aligned}
&\mathbb{P}\left(\widehat{F}_i(g_i(c_L\sqrt{\log n \vee \log(d-p)})) \ge 1-\delta_{n,d,p}\right) \\
&= \mathbb{P}\left(\widehat{F}_i(g_i(c_L\sqrt{\log n \vee \log(d-p)})) - \mathbb{E}\widehat{F}_i(g_i(c_L\sqrt{\log n \vee \log(d-p)})) \ge 1-\delta_{n,d,p}-\mathbb{E}\widehat{F}_i(g_i(c_L\sqrt{\log n \vee \log(d-p)}))\right) \\
&\le \exp\left[-2n\left(1-\delta_{n,d,p}-\mathbb{E}\widehat{F}_i(g_i(c_L\sqrt{\log n \vee \log(d-p)}))\right)^2\right].
\end{aligned}$$

The term is bounded as

$$\begin{aligned}
&1-\delta_{n,d,p}-\mathbb{E}\widehat{F}_i\left(g_i(c_L\sqrt{\log n \vee \log(d-p)})\right) \\
&\ge \left|1-\delta_{n,d,p}-F_i\left(g_i(c_L\sqrt{\log n \vee \log(d-p)})\right)\right| - \left|F_i\left(g_i(c_L\sqrt{\log n \vee \log(d-p)})\right) - \mathbb{E}\widehat{F}_i(g_i(c_L\sqrt{\log n \vee \log(d-p)}))\right| \\
&= \left|1-\delta_{n,d,p}-\Phi\left(c_L\sqrt{\log n \vee \log(d-p)}\right)\right| - \sqrt{\varepsilon_x}.
\end{aligned}$$

Similarly, we can prove

$$\mathbb{P}\left(\widehat{F}_i(g_i(-c_L\sqrt{\log n \vee \log(d-p)})) \leq \delta_{n,d,p}\right) \leq \exp\left(-2n\left(1-\delta_{n,d,p}-\Phi\left(c_L\sqrt{\log n \vee \log(d-p)}\right)-\sqrt{\varepsilon_x}\right)^2\right).$$

□

E Additional Simulations

This section presents additional simulation results with synthetic data. Here we adopt different underlying structures of $\mathbf{\Theta}^\natural$ and study the performance of our algorithm.

E.1 Synthetic data with parametric method

We adopt the same graph construction in the main context.

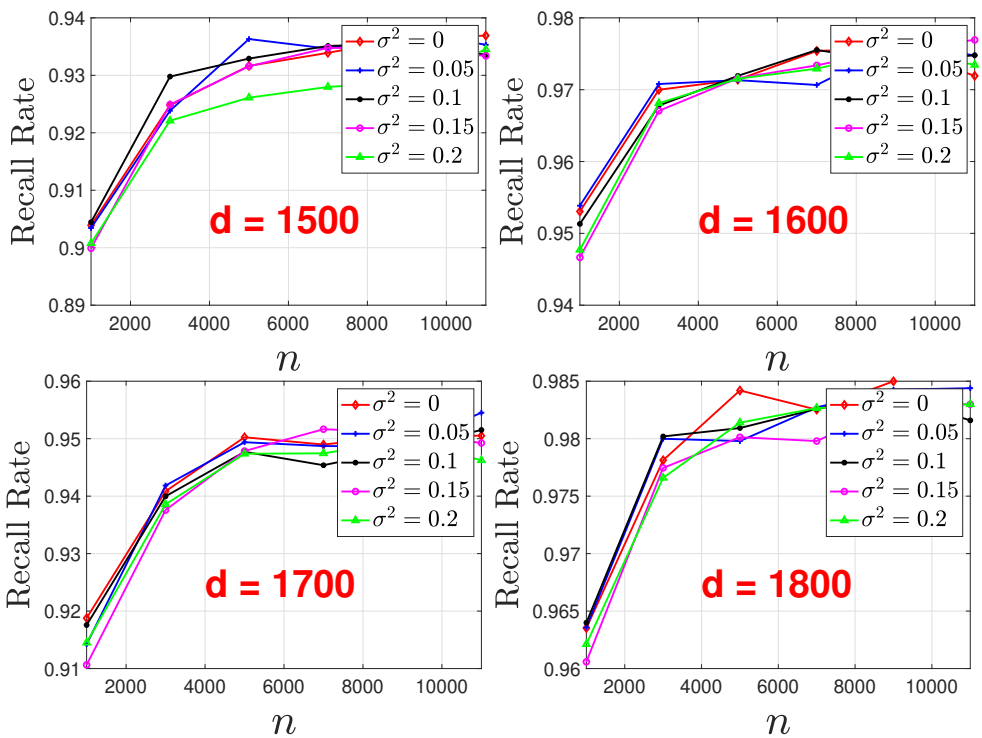

Figure 4: We study the impact of sample size n on the recall rate. The signal dimension p is fixed as 2000.

E.2 Synthetic data with non-parametric method

This subsection uses the same setting as that in the main context but evaluates the performance of our estimator via the recall rate and the precision rate of the edge selection, which is shown in Tab. 2.

We construct the sparse precision matrix $\mathbf{\Theta}$ as

$$\Theta_{i,j} = \begin{cases} \rho_1, & \text{if } i = j; \\ \rho_2, & \text{if } |i-j| = 1; \\ 0. & \text{otherwise,} \end{cases}$$

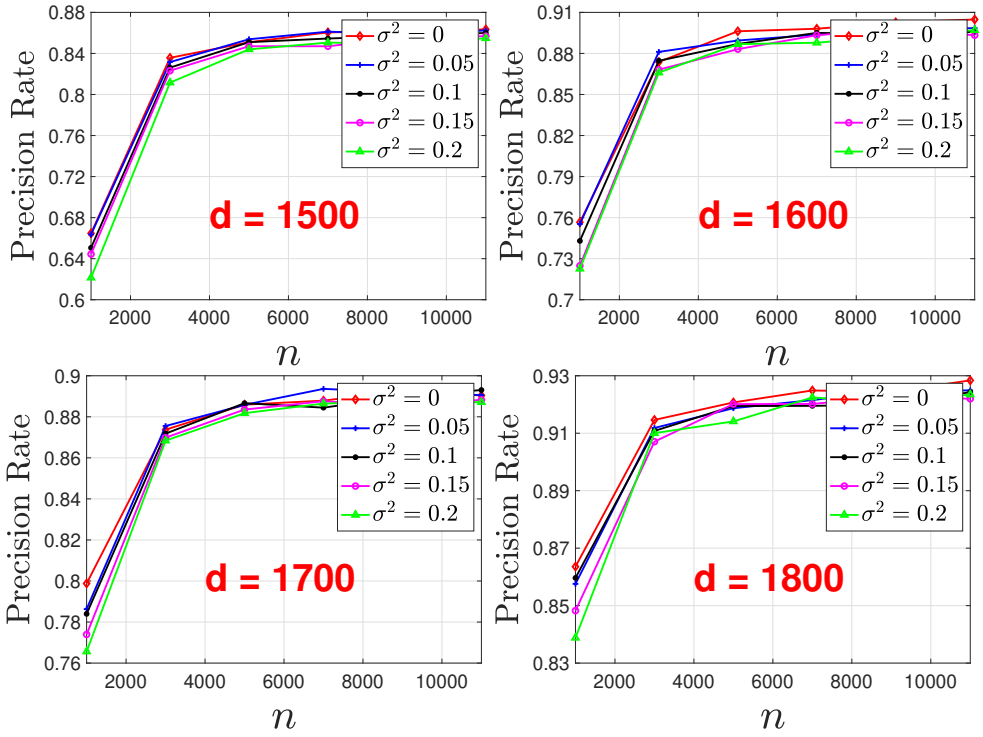

Figure 5: We study the impact of sample size n on the precision rate. The signal dimension p is fixed as 2000.

which is previously adopted in Ravikumar et al. (2011). The corresponding edge set of is denoted as E. We fix the signal length p as 100 and evaluate the performance with the following three types of marginal distribution for the random vector $\mathbf{X}$:

- uniform distribution within the region $[0, 1]$;
- exponential distribution, i.e., e^{-z} for $z \geq 0$;
- Gaussian mixture, i.e., $0.25 \sum_{i=1}^{4} \mathsf{N}(\mu_i, 10^{-2})$, where $\mu_i \in \{\pm 0.25, \pm 0.5\}$.

We set ρ_1 and ρ_2 as 1 and 0.4, respectively. For the baseline in 3, we assume the underlying distribution of $\mathbf{X}$ to be jointly Gaussian. With direct observations, we learn the graphical structure with graphical lasso. The recall rate and precision rate is shown in 3. For the uniform and exponential distribution, our algorithm has a significant improvement in terms of both the recall rate and precision rate. While for mixture Gaussian, the improvement is modest. This phenomenon may justify the approximation of mixture Gaussian with Gaussian distribution, which is widely used in the field of coding theory, machine learning, etc.

E.3 Real-World Data

We now consider the real-world databases, which consists of 5 databases: Carolina Breast Cancer (GSE148426) with 2497 samples (patients) Bhattacharya et al. (2020), Lung Cancer (GSE137140) with 3924 samples Asakura et al. (2020), Ovarian Cancer (GSE106817) with 4046 samples Yokoi et al. (2018), Colorectal Cancer (GSE115513) with 1513 samples Slattery et al. (2016), and Esophageal Squamous Cell Carcinoma (GSE122497) with 5531 samples Sudo et al. (2019). Each database is divided into two categories, i.e., **Healthy group** and **Unhealthy group**, where the measurements are given as the concentration of miRNAs. The miRNAs are known to have dependency among each other, i.e., a non-diagonal precision matrix, and hence there is an underlying graphical model describing these dependency structure based on the associated precision matrix.

Preprocess For each database, there are multiple types of data, namely, healthy vs non-healthy, benign cancer vs non-benign cancer, etc. Based on the labels, we first divide the whole dataset into two types of

	Uniform		**Exponential**		**Gauss Mixture**	
n	Recall Rate	Precision Rate	Recall Rate	Precision Rate	Recall Rate	Precision Rate
$d = 200$						
100	0.9315	0.9257	0.9839	0.9029	0.9732	0.9114
115	0.9342	0.9527	0.9866	0.9043	0.9758	0.9331
130	0.9369	0.9555	0.9906	0.9323	0.9826	0.9363
145	0.9450	0.9655	0.9946	0.9375	0.9866	0.9534
160	0.9490	0.9756	0.9946	0.9527	0.9866	0.9638
175	0.9490	0.9793	0.9946	0.9504	0.9879	0.9712
$d = 300$						
100	0.9584	0.9347	0.9852	0.9074	0.9852	0.9176
115	0.9651	0.9400	0.9799	0.9070	0.9906	0.9191
130	0.9812	0.9545	0.9919	0.9276	0.9933	0.9290
145	0.9812	0.9549	0.9933	0.9333	0.9933	0.9298
160	0.9852	0.9597	0.9973	0.9411	0.9946	0.9376
175	0.9839	0.9656	0.9973	0.9458	0.9960	0.9513
$d = 500$						
100	0.9745	0.9455	0.9812	0.9097	0.9879	0.9126
115	0.9758	0.9758	0.9866	0.9320	0.9893	0.9320
130	0.9758	0.9813	0.9919	0.9467	0.9906	0.9343
145	0.9826	0.9826	0.9946	0.9470	0.9933	0.9477
160	0.9879	0.9879	0.9946	0.9502	0.9946	0.9537
175	0.9893	0.9893	0.9973	0.9619	0.9960	0.9604
Baseline	0.3356	0.1834	0.3765	0.0497	0.8188	0.8538

Table 2: Recall rate and precision rate. The signal length p is fixed as 100 and σ^2 is fixed as 1.0. The sample number corresponding to the baseline is set as 175, the maximum sample number.

data and separately preprocess them. For each type, we split the data into the training and testing sets $\{\mathcal{D}^i_{\text{train}}, \mathcal{D}^i_{\text{test}}\}$, $i \in \{1, 2\}$. The sensing matrix $\mathbf{A}$ is assumed to be $A_{ij} \overset{\text{i.i.d}}{\sim} \mathsf{N}(0, 1)$ and the measurement noise $\boldsymbol{w}$ is set to be zero.

Training Using the training set $\mathcal{D}^i_{\text{train}}$, which are directly observed from the desired signal (miRNAs), we first select the penalty coefficients in graphical lasso via cross-validation. Then estimate the precision matrix via the graphical Lasso and use it as the baseline $\mathbf{\Theta}_i$, $i \in \{1, 2\}$ as they are obtained through direct observation. We denote $\mathbf{\Theta}_1$ for $\mathcal{D}^1$ while $\mathbf{\Theta}_2$ for $\mathcal{D}^2$. Then we mask the training set with the synthetic sensing matrix $\mathbf{A}$ and create indirect observation of data. Last we estimate the precision matrix $\widehat{\mathbf{\Theta}}_i$, from the indirect observations, i.e., $\mathbf{A}\mathcal{D}^i_{\text{train}}$ using M-gLasso, $i \in \{1, 2\}$.

Testing Notice that the ground-truth precision matrix $\mathbf{\Theta}^\natural_i$, $i \in \{1, 2\}$ cannot be obtained in the real-world applications, even with the direct observations. Hence a direct comparison between $\mathbf{\Theta}^\natural_i$ and $\widehat{\mathbf{\Theta}}_i$ cannot be performed. To evaluate the performance of the algorithm, we take an indirect way by using *quadratic discriminant analysis* Hastie et al. (2009) to perform classification in the testing set $\{\mathcal{D}^1_{\text{test}} \bigcup \mathcal{D}^2_{\text{test}}\}$, with the estimated matrix $\widehat{\mathbf{\Theta}}_i$, $i \in \{1, 2\}$. Then we compare with the classification accuracy when using $\mathbf{\Theta}_i$, $i \in \{1, 2\}$. The summary of the classification rates are shown in Tab. **??**.

Discussion From the tables we conclude that our estimated precision matrix $\widehat{\mathbf{\Theta}}_i$ achieves almost the same classification accuracy with the baseline $\mathbf{\Theta}_i$ when $d = p$, and is only a slightly worse when $d = 0.5p$, i.e., the dimension of the projection space under the indirect observation is half of the dimension of the signal space. In some special cases, we even see some improvements with indirect observations. One possible reason is that the features are mixed by sensing matrix $\mathbf{A}$, which lead to better quantities for the classification.

F Useful Facts about Probability Inequalities and Random Matrices

For the self-containing of this paper, we list some useful facts about probability inequalities and random matrices in this section.

Table 3: Recall rate and precision rate. The signal length p is fixed as 100 and σ^2 is fixed as 1.0. The sample number corresponding to the baseline is set as 175, the maximum sample number.

	Uniform		**Exponential**		**Gauss Mixture**	
n	Recall Rate	Precision Rate	Recall Rate	Precision Rate	Recall Rate	Precision Rate
$d = 200$						
100	0.9315	0.9257	0.9839	0.9029	0.9732	0.9114
115	0.9342	0.9527	0.9866	0.9043	0.9758	0.9331
130	0.9369	0.9555	0.9906	0.9323	0.9826	0.9363
145	0.9450	0.9655	0.9946	0.9375	0.9866	0.9534
160	0.9490	0.9756	0.9946	0.9527	0.9866	0.9638
175	0.9490	0.9793	0.9946	0.9504	0.9879	0.9712
$d = 300$						
100	0.9584	0.9347	0.9852	0.9074	0.9852	0.9176
115	0.9651	0.9400	0.9799	0.9070	0.9906	0.9191
130	0.9812	0.9545	0.9919	0.9276	0.9933	0.9290
145	0.9812	0.9549	0.9933	0.9333	0.9933	0.9298
160	0.9852	0.9597	0.9973	0.9411	0.9946	0.9376
175	0.9839	0.9656	0.9973	0.9458	0.9960	0.9513
$d = 500$						
100	0.9745	0.9455	0.9812	0.9097	0.9879	0.9126
115	0.9758	0.9758	0.9866	0.9320	0.9893	0.9320
130	0.9758	0.9813	0.9919	0.9467	0.9906	0.9343
145	0.9826	0.9826	0.9946	0.9470	0.9933	0.9477
160	0.9879	0.9879	0.9946	0.9502	0.9946	0.9537
175	0.9893	0.9893	0.9973	0.9619	0.9960	0.9604
Baseline	0.3356	0.1834	0.3765	0.0497	0.8188	0.8538

Table 4: Recall rate of edge selection on real-world databases, namely, GSE148426 Bhattacharya et al. (2020), GSE137140 Asakura et al. (2020), GSE106817 Yokoi et al. (2018), GSE115513 Slattery et al. (2016), and GSE122497 Sudo et al. (2019). The precision matrix $\mathbf{\Theta}$ learned by direct observations is assumed to be the ground-truth.

	Healthy group					Unhealthy group				
d/p	GSE 148426	GSE 137140	GSE 106817	GSE 115513	GSE 122497	GSE 148426	GSE 137140	GSE 106817	GSE 115513	GSE 122497
2	0.9494	0.8892	0.9	0.9856	0.7659	0.9424	0.9692	1	0.9856	0.8379
5	1	0.9950	1	1	0.9220	1	0.9692	1	1	0.9632
10	1	1	1	1	0.9707	1	1	1	1	0.9963
12	1	1	1	1	0.9805	1	1	1	1	1
15	1	1	1	1	0.9902	1	1	1	1	1
20	1	1	1	1	1	1	1	1	1	1

Lemma 30 ((Wainwright, 2019) (Example 2.11, P29))**.** *For a χ^2-RV Z with ℓ degrees of freedom, we have*

$$\mathbb{P}\left(|Z-\ell| \geq t\right) \leq 2\exp\left(-\left(\frac{t^2}{8\ell} \wedge \frac{t}{8}\right)\right), \quad \forall\, t \geq 0.$$

Theorem 8 (Theorem 2.35 in Tulino et al. (2004))**.** *For a $d \times p$ matrix $\mathbf{A}$ whose entries are independent zero-mean real RVs with variance d^{-1} and fourth moment of order $O(d^{-2})$, we have the empirical distribution of the eigenvalues of $\mathbf{A}^\top\mathbf{A}$ converge to the distribution with density*

$$f_\tau(x) = \left[0 \vee \left(1-\tau^{-1}\right)\right]\mathbb{1}(x) + \frac{\sqrt{\left[\left(\left(1+\sqrt{\tau}\right)^2 - x\right) \vee 0\right] \times \left[\left(x - \left(1-\sqrt{\tau}\right)^2\right) \vee 0\right]}}{2\pi\tau x},$$

as d and p approaches to infinity with $p/d \to \tau$.

Table 5: Precision rate of edge selection on real-world databases, namely, GSE148426 Bhattacharya et al. (2020), GSE137140 Asakura et al. (2020), GSE106817 Yokoi et al. (2018), GSE115513 Slattery et al. (2016), and GSE122497 Sudo et al. (2019). The precision matrix $\mathbf{\Theta}$ learned by direct observations is assumed to be the ground-truth.

	Healthy group					Unhealthy group				
d/p	GSE 148426	GSE 137140	GSE 106817	GSE 115513	GSE 122497	GSE 148426	GSE 137140	GSE 106817	GSE 115513	GSE 122497
2	1	1	1	0.9548	1	0.9704	0.9692	1	0.9581	1
5	0.9080	0.9900	1	0.9176	1	0.9205	0.9692	1	0.9207	0.9924
10	0.9080	0.9341	0.9091	0.9207	0.9900	0.9205	0.9420	1	0.9268	0.9854
12	0.9080	0.9475	0.9259	0.9237	0.9901	0.9329	0.9420	1	0.9268	0.9645
15	0.9080	0.9566	0.9434	0.9299	0.9621	0.9456	0.9420	1	0.9299	0.9645
20	0.9518	0.9613	0.9615	0.9393	0.9535	0.9586	0.9420	1	0.9457	0.9784

Theorem 9 (Corol. 13.2 in Boucheron et al. (2013))**.** *Consider a totally bounded pseudo-metric space* $(\mathcal{T}, \mathrm{dist}(\cdot,\cdot))$*. Provided a collection of RVs* $\{Z_u\}_{u\in\mathcal{T}}$ *satisfying*

$$\mathbb{E}e^{\lambda(Z_u - Z_v)} \leq \frac{\nu\lambda^2 \mathrm{dist}^2(u,v)}{2}, \quad \forall\ \lambda \geq 0,$$

we have

$$\mathbb{E}\sup_u (Z_u - Z_{u_0}) \leq 12\sqrt{\nu}\int_0^{\mathrm{diam}(T)/2} \sqrt{\mathcal{H}(\delta,\mathcal{T})}d\delta,$$

where $\mathcal{H}(t,\mathcal{T})$ *denotes the* δ*-covering entropy with pseudo-metric* $\mathrm{dist}(\cdot,\cdot)$*, and* $\mathrm{diam}(\cdot) \triangleq \sup \mathrm{dist}(u,u_0)$ *is the diameter of the set* $\mathcal{T}$*.*

Lemma 31 (Lemma 11 in Liu et al. (2009))**.** *Denote the distribution function and density function of the standard normal RV as* $\Phi(\cdot)$ *and* $\phi(\cdot)$*, respectively. We have*

$$\left(\Phi^{-1}\right)'(\eta) = \frac{1}{\phi\left(\Phi^{-1}(\eta)\right)}.$$

Furthermore, we have

$$\Phi^{-1}(\eta) \leq \sqrt{2\log\frac{1}{1-\eta}},$$

for $\eta \geq 0.99$*.*

Theorem 10 (Wick's theorem, Thm. 1.28 (P11) in Janson et al. (1997))**.** *Considering the centered jointly normal variables* $\xi_1, \xi_2, \cdots, \xi_n$*, we conclude*

$$\mathbb{E}\left(\xi_1\xi_2\cdots\xi_n\right) = \sum_{\substack{\text{all possible disjoint}\\ \text{pairs } (i_k,j_k) \text{ of } [n]}} \prod_k \mathbb{E}\left(\xi_{i_k}\xi_{j_k}\right).$$

Notice that the variables $\{\xi_i\}_{1\leq i\leq n}$ are not necessarily different nor independent. To illustrate this theorem, we consider two special cases. First we let $\xi_1 = \xi_2 = \xi_3 = \xi_4 = \xi \sim \mathsf{N}(0,1)$, then we have

$$\mathbb{E}\xi^4 = \mathbb{E}\left(\xi_1\xi_2\xi_3\xi_4\right) = \mathbb{E}(\xi_1\xi_2)\mathbb{E}(\xi_3\xi_4) + \ \mathbb{E}(\xi_1\xi_3)\mathbb{E}(\xi_2\xi_4) + \ \mathbb{E}(\xi_1\xi_4)\mathbb{E}(\xi_2\xi_3) = 3.$$

Second we consider the case where n is odd. Since we cannot partition $\{1,2,\cdots,n\}$ into disjoint pairs (ξ_{i_k},ξ_{j_k}), we always have $\mathbb{E}(\xi_1\cdots\xi_n) = 0$.