INFERENCE FOR INDIVIDUAL MEDIATION EFFECTS AND INTERVENTIONAL EFFECTS IN SPARSE HIGH-DIMENSIONAL CAUSAL GRAPHICAL MODELS

By Abhishek Chakrabortty[1], Preetam Nandy[1] and Hongzhe Li*

University of Pennsylvania

We consider the problem of identifying intermediate variables (or mediators) that regulate the effect of a treatment on a response variable. While there has been significant research on this topic, little work has been done when the set of potential mediators is high-dimensional. A further complication arises when the potential mediators are interrelated. In particular, we assume that the causal structure of the treatment, the potential mediators and the response is a directed acyclic graph (DAG). High-dimensional DAG models have previously been used for the estimation of causal effects from observational data. In particular, methods called IDA and joint-IDA have been developed for estimating the effect of single interventions and the effect of multiple simultaneous interventions respectively. In this paper, we propose an IDA-type method, called MIDA, for estimating mediation effects from high-dimensional observational data. Although IDA and joint-IDA estimators have been shown to be consistent in certain sparse high-dimensional settings, their asymptotic properties such as convergence in distribution and inferential tools in such settings remained unknown. In this paper, we prove high-dimensional consistency of MIDA for linear structural equation models with sub-Gaussian errors. More importantly, we derive distributional convergence results for MIDA in similar high-dimensional settings, which are applicable to IDA and joint-IDA estimators as well. To the best of our knowledge, these are the first distributional convergence results facilitating inference for IDA-type estimators. These results have been built on our novel theoretical results regarding uniform bounds for linear regression estimators over varying subsets of high-dimensional covariates, which may be of independent interest. Finally, we empirically demonstrate the usefulness of our asymptotic theory in the identification of large mediation effects and we illustrate a practical application of MIDA in genomics with a real dataset.

* [1]These authors contributed equally to this work. Preetam Nandy is currently at LinkedIn Corp. This research was supported in part by the National Institutes of Health Grants GM123056 and GM129781.

MSC 2010 subject classifications: Primary 62F12, 62H05, 62H10, 62J05, 92B15, 62-09.

Keywords and phrases: Linear structural equation models, Directed acyclic graphs, Markov equivalence class, Sparsity, Post-selection inference, High-dimensional regression on varying subsets of covariates, Uniform bounds and asymptotic linearity.

1. Introduction. Although confirmatory causal inference from high-dimensional observational data is impossible due to identifiability issues, this topic has received great attention in the recent past. Intervention experiments are considered to be the gold-standard for making causal inference. However, experimental data cannot always be generated given the considerable ethical concerns, time constraints, and the high costs associated with performing appropriate experiments. Another major problem that can arise in many scientific disciplines is that the sheer number of causal hypotheses is simply too large to test experimentally. Good examples are gene knockout experiments, where potential candidate genes for knockout experiments typically lie in the thousands. In such a situation, causal predictions from observational data can be extremely useful in prioritizing intervention experiments [Maathuis et al., 2010; Stekhoven et al., 2012; Le et al., 2017].

There has been a lot of recent progress in estimating causal effects from high-dimensional observational data. Most of these methods assume that the data are generated from an unknown linear structural equation model (LSEM) with independent Gaussian errors and that the causal relationships among the variables can be represented by a directed cyclic graph (DAG). Under these assumptions, high-dimensional consistency results have been derived for the estimation of causal graph and causal effects. In particular, Maathuis, Kalisch and Bühlmann [2009] proposed Interventional calculus when the DAG is Absent (IDA) for estimating the total causal effect of a variable on another variable and the authors proved a high-dimensional consistency result for the IDA estimator. The IDA method has been extended to joint-IDA by Nandy, Maathuis and Richardson [2017] for estimating the effects of multiple simultaneous interventions and a similar high-dimensional consistency result for the joint-IDA estimator has been proved therein.

The IDA method estimates a multi-set of causal effects as follows. The first step is to estimate a partially directed graph, called Completed Partially Directed Acyclic Graph (CPDAG), from high-dimensional observational data. This can be done by applying a structure learning algorithm such as the PC algorithm [Spirtes, Glymour and Scheines, 2000; Colombo and Maathuis, 2014], greedy equivalence search [Chickering, 2002a] and adaptively restricted greedy equivalence search [Nandy, Hauser and Maathuis, 2018]. High-dimensional consistency results for these structure learning algorithms have been proved in [Kalisch and Bühlmann, 2007; Colombo and Maathuis, 2014; Nandy, Hauser and Maathuis, 2018]. The reason behind estimating a partially directed graph instead of estimating the underlying directed graph is that the causal DAG is not identifiable from observational data without making further stringent assumptions. A CPDAG uniquely represents a Markov equivalence class of DAGs that can generate the same joint distribution of the variables. The IDA method estimates a possible causal effect for each DAG in the Markov equivalence class represented by the estimated CPDAG and combines them to produce a multi-set (where each element can have multiple copies) of causal effects. The authors noted that the listing of all DAGs in the Markov equivalence class from a given CPDAG is typically computationally infeasible for large graphs with thousands of variables and they provided computational shortcuts to obtain the multi-set of possible effects without listing all DAGs in the Markov equivalence class of the estimated CPDAG. It is common practice to summarize the multi-set of possible effects by its average or the minimum absolute value.

In this paper, we propose an IDA-type method, called MIDA, for estimating the causal mediation effect of a treatment variable on a response variable through an intermediate variable (a.k.a. mediator) in high-dimensional settings. In particular, we consider a treatment (a.k.a. exposer) X_1, a response variable X_p and a set of potential mediators $\{X_2, \dots, X_{p-1}\}$. We assume that the causal relationships among the variables in $\mathbf{X} = \{X_1, X_2, \dots, X_p\}$ can be represented by a DAG, where X_i and X_j are connected by a directed edge if and only if X_i is a direct cause of X_j. Further, we assume that the joint distribution of $\mathbf{X}$ is generated from an LSEM characterized by the causal DAG (see Section 2.2).

As is the case with IDA, MIDA relies on the estimation of an underlying CPDAG, and it produces a multi-set of possible mediation effects which we summarize by taking average. We prove consistency of MIDA for certain sparse high-dimensional LSEMs with sub-Gaussian errors. Note that we relax the Gaussian errors requirement of the high-dimensional consistency results of IDA and joint-IDA estimators [Maathuis, Kalisch and Bühlmann, 2009; Nandy, Maathuis and Richardson, 2017]. Furthermore, we provide unified distributional convergence results for IDA-type estimators in similar high-dimensional settings. These results have been built on a uniform non-asymptotic theory for linear regression over varying subsets of high-dimensional covariates, which may be of independent interest. To the best of our knowledge, we propose the first estimation method of mediation effects when the data are generated from an unknown DAG, as well as the first high-dimensional distributional convergence results for IDA-type estimators of both interventional effects as well as mediation effects.

In the context of causal mediation analysis, a simpler problem considering only one potential mediator has been well studied within the framework of LSEM [Judd and Kenny, 1981; James, Mulaik and Brett, 1982; Sobel, 1982; Baron and Kenny, 1986; MacKinnon et al., 2002]. The goal of causal mediation analysis is to understand what portion of the total causal effect of a treatment on a response can be attributed to the potential mediator. In fact, the total effect can be decomposed as a sum of the direct effect and the indirect effect, where the indirect effect is the effect of the treatment on the response that goes through the potential mediator. Similarly, in the case of multiple potential mediators, we are interested in understanding what portion of the total effect of the treatment X_1 on the response X_p can be attributed to a potential mediator X_j. We refer to it as the individual mediation effect with respect to X_j. Note that the total effect of the treatment on the response may not be decomposed as the sum of all individual mediation effects and the direct effect, unless the potential mediators are conditionally independent of each other given the treatment variable.

The estimation and testing for mediation effects in causal models with conditionally independent mediators have been considered in the classical setting [Preacher and Hayes, 2008; Boca et al., 2014], as well as in high-dimensional settings [Zhang et al., 2016]. For causal models with conditionally dependent mediators, VanderWeele and Vansteelandt [2014] discussed estimation methods for the total effect of all mediators (or the total indirect effect), whereas Huang and Pan [2016] proposed to estimate the individual effects with respect to a transformed set of conditionally independent variables in high-dimensional settings. In contrast to these existing works, we are interested in separately evaluating the importance of each potential mediators. The identification of mediators corresponding to large individual mediation effects can be very useful in a variety of scientific applications including genomics, for instance, where it is often of interest to understand how an influential genotype regulates a phenotype of interest through gene expressions.

The rest of this paper is organized as follows. Section 2 provides some necessary background material. In Section 3, we propose the MIDA algorithm for estimating individual mediation effects from observational data. We prove consistency of MIDA in certain sparse high-dimensional settings in Section 4. Our non-asymptotic theoretical results on linear regression over varying subsets of high-dimensional covariates are given in Section 5, which can be read independently. Section 6 discusses distributional convergence results and inferential tools for the MIDA estimator. Section 7 contains simulation results, where we demonstrate the usefulness of MIDA and our asymptotic theory for the identification of non-zero mediation effects. In Section 8, we apply MIDA to a real dataset generated from a collection of yeast segregants. We end with a discussion and problems for future research in Section 9. All proofs, additional technical materials and additional numerical results are collected in the Supplementary Material.

2. Preliminaries.

2.1. *Graph Terminology.* We consider graphs $\mathcal{H} = (\mathbf{X}, E)$ with vertex (or node) set $\mathbf{X} = \{X_1, \ldots, X_p\}$ and edge set E. There is at most one edge between any pair of vertices and edges may be either directed ($X_i \to X_j$) or undirected ($X_i - X_j$). If $\mathcal{H}$ contains only (un)directed edges, it is called *(un)directed*. If $\mathcal{H}$ contains directed and/or undirected edges, it is called *partially directed*. A pair of nodes $\{X_i, X_j\}$ are *adjacent* if there is an edge between X_i and X_j. If $X_i \to X_j$, then X_i is a parent of X_j. We denote the set of all parents of X_j in $\mathcal{H}$ by $\mathbf{Pa}_{\mathcal{H}}(X_j)$. A path between X_i and X_j is a sequence of distinct nodes $\{X_i, \ldots, X_j\}$ such that all successive pairs of nodes are adjacent. A directed path from X_i to X_j is a path between X_i and X_j where all edges are directed towards X_j. A directed path from X_i to X_j together with the edge $X_j \to X_i$ forms a *directed cycle*. A (partially) directed graph that does not contain a directed cycle is called a *(partially) directed acyclic graph* or (P)DAG.

2.2. *Linear Structural Equation Models.*

Definition 2.1. *Let $\mathcal{G}_0 = (\mathbf{X}, E)$ be a DAG and let $B_{\mathcal{G}_0}$ be a $p \times p$ matrix such that $(B_{\mathcal{G}_0})_{ij} \neq 0$ if and only if $X_i \in \mathbf{Pa}_{\mathcal{G}_0}(X_j)$. Let $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_p)^T$ be a zero mean random vector of jointly independent error variables. Then $\mathbf{X} = (X_1, \ldots, X_p)^T$ is said to be generated from a linear structural equation model (LSEM) characterized by the pair $(B_{\mathcal{G}_0}, \boldsymbol{\epsilon})$ if*

$$(\mathbf{X} - \boldsymbol{\mu}) \leftarrow B_{\mathcal{G}_0}^T (\mathbf{X} - \boldsymbol{\mu}) + \boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{\mu} := \mathbb{E}(\mathbf{X}). \tag{2.1}$$

If $\mathbf{X}$ is generated from an LSEM characterized by the pair $(B_{\mathcal{G}_0}, \boldsymbol{\epsilon})$, then we call $\mathcal{G}_0$ the *causal DAG*. The symbol "$\leftarrow$" in (2.1) emphasizes that the expression should be understood as a generating mechanism rather than as a mere equation. We emphasize that we assume that there are no hidden confounders; hence the joint independence of the error terms. In the rest of the paper, we refer to LSEMs without explicitly mentioning the independent error assumption.

2.3. *Markov Equivalence Class of DAGs.* The causal DAG $\mathcal{G}_0$ is not identifiable from (observational data from) the distribution of $\mathbf{X}$. A DAG encodes conditional independence relationships via the notion of *d-separation* (Pearl [2000], Theorem 1.2.4, page 18). In general, several DAGs can encode the same conditional independence relationships and such DAGs form a *Markov equivalence class.* Two DAGs belong to the same Markov equivalence class if and only if they have the same skeleton and the same v-structures [Verma and Pearl, 1990]. A Markov equivalence class of DAGs can be uniquely represented by a *completed partially directed acyclic graph* (CPDAG) [Spirtes, Glymour and Scheines, 2000; Chickering, 2002b], which is a graph that can contain both directed and undirected edges. A CPDAG satisfies the following: $X_i \to X_j$ in the CPDAG if $X_i \to X_j$ in every DAG in the Markov equivalence class, and $X_i - X_j$ in the CPDAG if the Markov equivalence class contains a DAG for which $X_i \to X_j$ as well as a DAG for which $X_i \leftarrow X_j$. CPDAGs can be estimated from observational data using various algorithms [Spirtes, Glymour and Scheines, 2000; Chickering, 2002a; Tsamardinos, Brown and Aliferis, 2006; Nandy, Hauser and Maathuis, 2018].

2.4. *Setup.* We assume that $\mathbf{X}$ is generated from an LSEM characterized by the pair $(B_{\mathcal{G}_0}, \boldsymbol{\epsilon})$ and that the distribution $\mathbf{X}$ is faithful to $\mathcal{G}_0$. The faithfulness condition states that every independence constraint that holds in the distribution is encoded by $\mathcal{G}_0$ (see, for example, Definition 3.8 of Koller and Friedman [2009]) and it is an essential condition for learning causal structures from observational data [Spirtes, Glymour and Scheines, 2000; Chickering, 2002b]. Without loss of generality, let X_1 be the treatment variable and X_p the response variable. For $j = 2, \ldots, p-1$, we refer to X_j as a potential mediator. We assume that no potential mediator is a direct cause of the treatment variable X_1 (i.e. $(B_{\mathcal{G}_0})_{j1} = 0$). Further, we assume that the response variable X_p is not a direct cause of any potential mediator and the treatment variable (i.e. $(B_{\mathcal{G}_0})_{pj} = 0$). For $j = 2, \ldots, p$, we restrict ϵ_j to be continuous. However, we allow $X_1 = \epsilon_1$ to be a continuous variable, a discrete variable or a binary categorical variable (coded with 0 and 1).

In order to define the total causal effect of a variable X_i on another variable X_k, we consider a hypothetical outside intervention to the system where we set a variable X_i to some value x_i uniformly over the entire population. This can be denoted by Pearl's do-operator: $do(X_i = x_i)$ [Pearl, 2009], which corresponds to removing the edges into X_i in $\mathcal{G}_0$ (or equivalently, setting the ith column of $B_{\mathcal{G}_0}$ equal to zero) and replacing ϵ_i by the constant x_i. The post-interventional expectation of X_k is denoted by $\mathbb{E}[X_k \mid do(X_i = x_i)]$.

Under the linearity assumption, $\mathbb{E}[X_k \mid do(X_i = x_i)]$ is a linear function of x_i and the *total causal effect* of X_i on X_k is defined as [Maathuis, Kalisch and Bühlmann, 2009]

$$\theta_{ik} := \frac{\partial}{\partial x_i} \mathbb{E}[X_k \mid do(X_i = x_i)].$$

In order to provide a graphical interpretation of θ_{ik}, we define the effect of X_{i_0} to $X_{i_{k+1}}$ through a directed path $\{X_{i_0}, X_{i_1}, \ldots, X_{i_k}, X_{i_{k+1}}\}$ equals $\prod_{r=0}^{k}(B_{\mathcal{G}_0})_{j_r j_{r+1}}$. Then the total causal effect θ_{ik} is given by the sum of the effects of X_i to X_k through all directed paths from X_i to X_k. This is known as the *path method* for computing the total causal effects in an LSEM [Wright, 1921].

We denote a joint-intervention on X_i and X_j by $do(X_i = x_i,\ X_j = x_j)$. Again, the post-interventional expectation $\mathbb{E}[X_k \mid do(X_i = x_i,\ X_j = x_j)]$ is a linear function of (x_i, x_j) and the effect of X_i on X_k in the joint intervention $do(X_i = x_i,\ X_j = x_j)$ is defined as [Nandy, Maathuis and Richardson, 2017]

$$\theta_{ik}^{(i,j)} := \frac{\partial}{\partial x_i}\, \mathbb{E}[X_k \mid do(X_i = x_i,\ X_j = x_j)].$$

Note that $\theta_{ik}^{(i,j)}$ can be interpreted as the total causal effect of X_i on X_k when we set $X_j = x_j$ uniformly over the entire population, that is the portion of the total effect of X_i on X_k that does not go though X_j.

Finally, we define the *individual mediation effect* to be the portion of total effect of the treatment variable X_1 on the response X_p that goes through a potential mediator as follows.

DEFINITION 2.2. *The individual mediation effect with respect to a potential mediator X_j is*

$$\eta_j := \frac{\partial}{\partial x_1}\, \mathbb{E}[X_p \mid do(X_1 = x_1)] - \frac{\partial}{\partial x_1}\, \mathbb{E}[X_p \mid do(X_1 = x_1,\ X_j = x_j)].$$

The individual mediation effect η_j is given by the sum of the effects of X_1 to X_p through all directed paths from X_1 to X_p that go through X_j. It follows from Theorem 3.1 of Nandy, Maathuis and Richardson [2017] that η_j equals the product of the total causal effect of X_1 on X_j and the total causal effect of X_j on X_p.

PROPOSITION 2.1. *Let $\mathbf{X}$ be generated from an LSEM. The individual mediation effect with respect to a potential mediator X_j is given by*

$$\eta_j = \theta_{1j}\theta_{jp},$$

where θ_{ik} denote the total causal effect of X_i on X_k.

2.5. *Notation.* We denote the vector of potential mediators $(X_2, \ldots, X_{(p-1)})^T$ by $\mathbf{X}'$ and the corresponding subgraph of $\mathcal{G}_0$ by $\mathcal{G}_0'$. Further, we denote the CPDAG representing the Markov equivalence class of $\mathcal{G}_0'$ by $\mathcal{C}_0'$ and the Markov equivalence class by $\mathrm{MEC}(\mathcal{C}_0')$.

We will often treat sets as vectors and vice versa, where we consider an arbitrary ordering of the elements in a vector unless specified otherwise. For example, (i, S, k) denotes a vector where the first element is i, the last element is k, but elements of the set S are ordered arbitrarily in (i, S, k). We denote the covariance matrix of $\mathbf{X}$ by Σ_0. For any set $S \subseteq \{1, \ldots, p\}$, we denote the corresponding random vector $\{X_r : r \in S\}$ by $\mathbf{X}_S$. Further, we denote $\mathrm{Cov}(\mathbf{X}_S)$ (which is shorthand for $\mathrm{Cov}(\mathbf{X}_S, \mathbf{X}_S)$) and $\mathrm{Cov}(\mathbf{X}_{S_1}, \mathbf{X}_{S_2})$ by $(\Sigma_0)_S$ and $(\Sigma_0)_{S_1 S_2}$ respectively. For simplicity, we denote $(\Sigma_0)_{\{i\}}$ and $(\Sigma_0)_{\{i\}S}$ by $(\Sigma_0)_i$ and $(\Sigma_0)_{iS}$ respectively.

We denote the i-th column of the $k \times k$ identity matrix by $e_{i,k}$. For $i \neq k$ and any set $S \subseteq \{1, \ldots, p\} \setminus \{i, k\}$, we denote the coefficient of X_i in the linear regression of X_k on $\mathbf{X}_{\{i\} \cup S}$ by $\beta_{ik|S}$ or by $\beta_{ik|\mathbf{X}_S}$. For simplicity, we denote $\beta_{ik|\emptyset}$ by β_{ik}. Note that $\beta_{ik|S}$ is well-defined regardless of whether or not the conditional expectation $\mathbb{E}[X_k \mid \mathbf{X}_{\{i\} \cup S}]$ is a linear function of $\{X_i\} \cup \mathbf{X}_S$, and in general, $\beta_{ik|S} = e^T_{1,|S|+1} (\Sigma_0)^{-1}_{(i,S)} (\Sigma_0)_{(i,S)k}$.

3. Estimating Individual Mediation Effects.

3.1. *Individual Mediation Effects via Covariate Adjustment.* Under the linearity assumption, the total causal effect of X_i on X_j can be obtained from a linear regression as follows [Maathuis, Kalisch and Bühlmann, 2009; Nandy, Maathuis and Richardson, 2017]:

$$\theta_{ij} = \begin{cases} 0 & \text{if } X_j \in \mathbf{Pa}_{\mathcal{G}_0}(X_i) \\ \beta_{ij|\mathbf{Pa}_{\mathcal{G}_0}(X_i)} & \text{otherwise} \end{cases} .$$

Since the treatment variable X_1 is assumed to have no parent in $\mathcal{G}_0$, we have $\theta_{1j} = \beta_{1j}$. Further, since the response variable is not a parent of any potential mediator X_j, we have

$$\theta_{jp} = \beta_{jp|\mathbf{Pa}_{\mathcal{G}_0}(X_j)} = \beta_{jp|\mathbf{Pa}_{\mathcal{G}_0}(X_j) \cup \{X_1\}},$$

where the last equality follows from the fact that X_1 and X_j are conditionally independent given $\mathbf{Pa}_{\mathcal{G}_0}(X_j)$ whenever $X_1 \notin \mathbf{Pa}_{\mathcal{G}_0}(X_j)$.

Let $\mathcal{G}_0'$ denote the DAG on the set of potential mediators that is obtained by deleting the nodes X_1 and X_p and the corresponding edges from $\mathcal{G}_0$. Then it is easy to see that $\mathbf{Pa}_{\mathcal{G}_0}(X_j) \cup \{X_1\} = \mathbf{Pa}_{\mathcal{G}_0'}(X_j) \cup \{X_1\}$. Hence, we have

$$\eta_j = \theta_{1j}\theta_{jp} = \beta_{1j}\ \beta_{jp|\mathbf{Pa}_{\mathcal{G}_0'}(X_j) \cup \{X_1\}}. \tag{3.1}$$

3.2. *The MIDA Estimator.* Our goal is to estimate $\eta_j = \theta_{1j}\theta_{jp}$ based on n i.i.d. data from the distribution of $\mathbf{X}$, for $j = 2, \ldots, p-1$. Note that if

$\mathbf{Pa}_{\mathcal{G}_0'}(X_j)$ were known, then we could estimate η_j by plugging in the sample regression coefficients $\hat{\beta}_{1j}$ and $\hat{\beta}_{jp|\mathbf{Pa}_{\mathcal{G}_0'}(X_j)\cup\{X_1\}}$ in (3.1). When $\mathcal{G}_0'$ is unknown, we need to estimate it from the data. However, as we mentioned in Section 2.3, we can only estimate the CPDAG $\mathcal{C}_0'$ that represents the Markov equivalence class of $\mathcal{G}_0'$. Consequently, θ_{jp} and η_j are not identifiable from observational data. Therefore, following the IDA approach of Maathuis, Kalisch and Bühlmann [2009]; Nandy, Maathuis and Richardson [2017], we estimate the identifiable parameter $\eta_j(\mathcal{C}_0')$ defined as

$$\begin{aligned} \eta_j(\mathcal{C}_0') &:= \beta_{1j}\ \mathrm{aver}(\Theta_{jp}(\mathcal{C}_0')), \quad \text{where} \\ \Theta_{jp}(\mathcal{C}_0') &:= \{\beta_{jp|\mathbf{Pa}_{\mathcal{G}}(X_j)\cup\{X_1\}} : \mathcal{G} \in \mathrm{MEC}(\mathcal{C}_0')\} \end{aligned}$$

is a multi-set of possible causal effects of X_j on X_p and $\mathrm{aver}(A)$ denote the average of all numbers (respecting multiple occurrences) in the multi-set A. We will empirically verify that $\eta_j(\mathcal{C}_0')$ serves as a reasonable proxy for η_j in sparse high-dimensional settings (see Section 7).

In order to estimate $\mathcal{C}_0'$, we first remove the effect of X_1 on each potential mediator X_j by replacing the data that corresponds to X_j by the residuals of the regression of X_j on X_1. Then we apply a structure learning algorithm on the transformed data for estimating $\mathcal{C}_0'$ (see Lemma 3.1). In Sections 7 and 8, we will use the adaptively restricted greedy equivalence search (ARGES) method of Nandy, Hauser and Maathuis [2018] with an ℓ_0-penalized Gaussian log-likelihood score (see Definition 5.1 of Nandy, Hauser and Maathuis [2018]) for estimating CPDAGs.

Algorithm 3.1 MIDA

Input: n i.i.d. observations of $\mathbf{X}$ (data)
Output: Estimate of $\eta_j(\mathcal{C}_0')$

1: for $j = 2, \ldots, p-1$, obtain the vector of residuals $\mathbf{r}_j = (r_j^{(1)}, \ldots, r_j^{(n)})$ from the regression of X_j on X_1;
2: apply a structure learning algorithm (such as (AR)GES) on the data $\{\mathbf{r}_2, \ldots, \mathbf{r}_{p-1}\}$ to obtain an estimate $\hat{\mathcal{C}}_0'$ of the CPDAG $\mathcal{C}_0'$;
3: obtain a multi-set $\hat{\Theta}_{jp}(\hat{\mathcal{C}}_0') := \{\hat{\beta}_{jp|\mathbf{Pa}_{\mathcal{G}}(X_j)\cup\{X_1\}} : \mathcal{G} \in \mathrm{MEC}(\hat{\mathcal{C}}_0')\}$ of possible causal effects based on the original data;
4: **return** $\hat{\eta}_j(\hat{\mathcal{C}}_0') := \hat{\beta}_{1j}\ \mathrm{aver}(\hat{\Theta}_{jp}(\hat{\mathcal{C}}_0'))$.

LEMMA 3.1. *Suppose* $\mathbf{X}$ *is generated from an LSEM characterized by* $(B_{\mathcal{G}_0}, \boldsymbol{\epsilon})$ *such that* $(B_{\mathcal{G}_0})_{j1} = (B_{\mathcal{G}_0})_{pj} = 0$ *for all* j. *Let* $\mathbf{X}' = (X_2, \ldots, X_{p-1})^T$, $B_{\mathcal{G}_0'}$ *be the submatrix of* $B_{\mathcal{G}_0}$ *that corresponds to* $\mathbf{X}'$ *and* $\boldsymbol{\epsilon}' = (\epsilon_2, \ldots, \epsilon_{p-1})^T$. *Then the conditional expectation* $\mathbb{E}[\mathbf{X}'|X_1]$ *is linear in* X_1. *Further,* $\mathbf{X}^\dagger := \mathbf{X}' - \mathbb{E}[\mathbf{X}'|X_1]$ *satisfies* $\mathbf{X}^\dagger = B_{\mathcal{G}_0'}^T \mathbf{X}^\dagger + \epsilon'$ *and the distribution of* $\mathbf{X}^\dagger$ *is faithful to* $\mathcal{G}_0'$.

The main difference between $\hat{\Theta}_{jp}(\hat{\mathcal{C}}_0')$ and the corresponding original IDA estimator of Maathuis, Kalisch and Bühlmann [2009] is that we *always include* X_1 in the adjustment set, leveraging the fact that X_j is not a direct cause of the treatment variable X_1. Further, note that computing MEC($\hat{\mathcal{C}}_0'$) can be computationally infeasible for a large CPDAG $\hat{\mathcal{C}}_0'$ [Maathuis, Kalisch and Bühlmann, 2009]. This computation bottleneck can be relieved by directly obtaining the multi-set of parent sets $\mathcal{PA}_{\hat{\mathcal{C}}_0'}(X_j) = \{\mathbf{Pa}_{\mathcal{G}}(X_j) : \mathcal{G} \in \text{MEC}(\hat{\mathcal{C}}_0')\}$ from $\hat{\mathcal{C}}_0'$ without computing MEC($\hat{\mathcal{C}}_0'$) via Algorithm 3 of Nandy, Maathuis and Richardson [2017]. We note that the output of Algorithm 3 of Nandy, Maathuis and Richardson [2017] and $\mathcal{PA}_{\hat{\mathcal{C}}}(X_j)$ may not be the same multi-set, but Theorem 5.1 of Nandy, Maathuis and Richardson [2017] guarantees that they are equivalent multi-sets in the sense that they have the same distinct elements and the ratio of the multiplicities of any two elements in the output of Algorithm 3 of Nandy, Maathuis and Richardson [2017] equals the ratio of their multiplicities in $\mathcal{PA}_{\hat{\mathcal{C}}}(X_j)$. Therefore, using the output of Algorithm 3 of Nandy, Maathuis and Richardson [2017] instead of MEC($\hat{\mathcal{C}}_0'$) makes no difference in obtaining aver($\hat{\Theta}_{jp}(\hat{\mathcal{C}}_0')$). Thus, for simplicity, we can safely pretend that we use MEC($\hat{\mathcal{C}}_0'$) for computing aver($\hat{\Theta}_{jp}(\hat{\mathcal{C}}_0')$) in the rest of the paper.

4. Consistency in High-Dimensional Settings. We now consider an asymptotic scenario where the sample size n and the number of potential mediators $(p-2)$ in $\mathbf{X}$ grows to infinity. We prove high-dimensional consistency of the MIDA estimator $\hat{\eta}(\hat{\mathcal{C}}_0')$ defined in Algorithm 3.1 when the CPDAG $\hat{\mathcal{C}}_0'$ is estimated using GES. We note that similar high-dimensional consistency results hold when the CPDAG $\hat{\mathcal{C}}_0'$ is estimated using ARGES [Nandy, Hauser and Maathuis, 2018] or PC [Spirtes, Glymour and Scheines, 2000]. Following Nandy, Hauser and Maathuis [2018], we make the following assumptions.

Assumption 4.1 (LSEM with sub-Gaussian error variables). $\mathbf{X}$ *is generated from a linear SEM* $(B_{\mathcal{G}_0}, \boldsymbol{\epsilon})$ *with sub-Gaussian error variables satisfying* $\max_{1\le i\le p} ||\epsilon_i||_{\psi_2} \le C_1$ *for some absolute constant* $C_1 > 0$, *where* $||\cdot||_{\psi_2}$ *denotes the sub-Gaussian norm, as defined in Definition A.1.*

Assumption 4.2 (High-dimensional setting). $p = O(n^a)$ *for some* $0 \le a < \infty$.

Assumption 4.3 (Sparsity condition). *Let* $q = \max_{2\le j\le p-1} |\mathbf{Adj}_{\mathcal{C}_0'}(X_j)|$ *be the maximum degree in* $\mathcal{C}_0'$. *Then* $q = O(n^{1-b_1})$ *for some* $0 < b_1 \le 1$.

ASSUMPTION 4.4 (Bounds on the growth of oracle versions). *The maximum degree in the output of the forward phase of a δ-optimal oracle version of GES (Definition 5.3 of Nandy, Hauser and Maathuis [2018]) based on an ℓ_0-penalized Gaussian log-likelihood oracle scoring criterion $\mathcal{S}^*_\lambda$ with respect to $\Sigma^\dagger := \mathrm{Cov}(\mathbf{X}' - \mathbb{E}[\mathbf{X}' \mid X_1])$ (see Definition 4.1 below) is bounded by $\tilde{q} = O(n^{1-b_2})$, for all $\lambda \geq 0$ and $\delta^{-1} = O(n^{d_1})$ such that $0 \leq 2d_1 < b_2 \leq 1$.*

ASSUMPTION 4.5 (Bounds on partial correlations). *The partial correlations $\rho_{ij|S}$ between X_i and X_j given $\mathbf{X}_S$ satisfy the following upper and lower bounds for all n, uniformly over $i, j \in \{2, \ldots, p-1\}$ and $S \subseteq \{2, \ldots, p-1\} \setminus \{i, j\}$ with $|S| \leq \tilde{q}$ (where $\tilde{q}$ is given by Assumption 4.4):*

$$\sup_{i \neq j, S} |\rho_{ij|S\cup\{1\}}| \leq M < 1 \quad \textit{and} \quad \inf_{i,j,S} \{|\rho_{ij|S\cup\{1\}}| : \rho_{ij|S\cup\{1\}} \neq 0\} \geq c,$$

such that $c^{-1} = O(n^{d_2})$ for some constant d_2 satisfying $0 < 2d_2 < b_2$, where b_2 is given by Assumption 4.4.

ASSUMPTION 4.6 (Bounds on the eigenvalues of covariance matrices). *For any $(\tilde{q}+3) \times (\tilde{q}+3)$ principal submatrix Σ of $\Sigma_0 = \mathrm{Cov}(\mathbf{X})$,*

$$C_2 \leq 1/\left|\left|\Sigma^{-1}\right|\right|_2 \leq ||\Sigma||_2 \leq C_3$$

for some absolute constants $C_2, C_3 > 0$, where $||\mathbf{M}||_2$ denotes the spectral norm of $\mathbf{M}$, for any matrix $\mathbf{M}$, and $\tilde{q}$ is given by Assumption 4.4.

DEFINITION 4.1 (Oracle score). *We define the ℓ_0 penalized Gaussian log-likelihood oracle scoring criterion oracle score of a DAG $\mathcal{H} = (\mathbf{X}', E)$ with respect to a covariance matrix Σ and the penalty parameter λ as*

$$\mathcal{S}^*_\lambda(\mathcal{H}, \Sigma) = -\sum_{i=1}^{p} \tilde{\mathbb{E}}\left[\log\left(L(\boldsymbol{\theta}^*_i(\mathcal{H}), X_i|\mathbf{Pa}_{\mathcal{H}}(X_i))\right)\right] + \lambda|E|, \quad \textit{where}$$

$\tilde{\mathbb{E}}[\cdot]$ denotes expectation with respect to a zero mean Gaussian distribution with covariance matrix Σ, $L(\boldsymbol{\theta}_i, X_i|\mathbf{Pa}_{\mathcal{H}}(X_i))$ is the Gaussian likelihood that corresponds to the conditional density of X_i given $\mathbf{Pa}_{\mathcal{H}}(X_i)$, and

$$\boldsymbol{\theta}^*_i(\mathcal{H}) = \operatorname*{argmax}_{\boldsymbol{\theta}_i} \tilde{\mathbb{E}}\left[\log\left(L(\boldsymbol{\theta}_i, X_i|\mathbf{Pa}_{\mathcal{H}}(X_i))\right)\right].$$

THEOREM 4.1. *Assume 4.1 - 4.6. Let $\hat{\mathcal{C}}'_0(\lambda_n)$ denote the estimated CPDAG in Algorithm 3.1, where λ_n denotes the penalty parameter of the ℓ_0-penalized Gaussian log-likelihood scoring criterion. Let $\hat{\eta}_j(\hat{\mathcal{C}}'_0) = \hat{\beta}_{1j} \cdot \mathrm{aver}(\hat{\Theta}_{jp}(\hat{\mathcal{C}}'_0))$ denote the output of Algorithm 3.1 based on the estimated CPDAG $\hat{\mathcal{C}}'_0$, where we simplify the notation by dropping λ_n. Then,* *1.* $\mathbb{P}(\hat{\mathcal{C}}'(\lambda_n) \neq \mathcal{C}'_0) \longrightarrow 0,$

2. $\max_{1<j<p} \left|\text{aver}(\hat{\Theta}_{jp}(\hat{\mathcal{C}}'_0)) - \text{aver}(\Theta_{jp})\right| \xrightarrow{\mathbb{P}} 0$ *and*

3. $\max_{1<j<p} \left|\hat{\eta}_j(\hat{\mathcal{C}}'_0) - \theta_{1j} \cdot \text{aver}(\Theta_{jp})\right| \xrightarrow{\mathbb{P}} 0,$

for any sequence $\{\lambda_n\}$ *satisfying* $\lambda_n < \frac{1}{8}\log(1-c^2)$ *and* $\frac{\tilde{q}\log(p)}{n\lambda_n} \to 0$*, where* $\tilde{q}$ *is as in Assumption 4.4.*

Note that Nandy, Hauser and Maathuis [2018] made the same assumptions to prove high-dimensional consistency of GES for linear SEMs with sub-Gaussian errors. However, the high dimensional consistency results of the (joint)-IDA estimators were proven only for linear SEMs with Gaussian errors [Maathuis, Kalisch and Bühlmann, 2009; Nandy, Maathuis and Richardson, 2017]. Thus, Theorem 4.1 extends the existing high-dimensional consistency results of the IDA estimators to linear SEMs with sub-Gaussian errors. We note that if we apply the PC algorithm [Spirtes, Glymour and Scheines, 2000; Kalisch and Bühlmann, 2007] instead of GES (or ARGES), we no longer require Assumption 4.4 and $\tilde{q}$ in Assumption 4.5 and Theorem 4.1 can be replaced by q (which is smaller or equal to $\tilde{q}$). We refer to Nandy, Hauser and Maathuis [2018] for a detailed discussion on these assumptions.

5. Linear Regression over Varying Subsets of High-Dimensional Covariates. This section considers linear regression over varying subsets of high-dimensional covariates in a general setting. The results will be used in Section 6 to derive the asymptotic distributions of the estimates of interventional and mediation effects. While derived primarily for establishing (uniform) asymptotic normality of our proposed estimators, these results are applicable far more generally to any setting involving linear regressions over varying (non-random) subsets of high-dimensional regressors and may be of independent interest. For notational simplicity and clarity of exposition, we therefore derive them under more general and standard notations where Y and $\mathbf{X}$ denote a generic response and a (high-dimensional) covariate vector respectively. All our results here are non-asymptotic.

Let $\{\mathbf{Z}_i \equiv (Y_i, \mathbf{X}_i)\}_{i=1}^n$ denote a collection of n independent and identically distributed (i.i.d.) observations of any random vector $\mathbf{Z} := (Y, \mathbf{X})$, where $Y \in \mathbb{R}$ and $\mathbf{X} \in \mathbb{R}^p$, and $p \equiv p_n$ is allowed to diverge n. We consider the linear regression of Y over a collection of subsets (each of size $< n$) of $\mathbf{X}$ via ordinary least squares (OLS) estimator(s), and derive uniform first order expansions of the resulting estimators, where the uniformity is in the control of the error terms over all the subsets considered. Our analysis is completely *model free*, i.e. for none of the regression problems involved, a true corresponding linear model is assumed, and further, neither Y nor $\mathbf{X}$ are needed to be continuous and/or centered.

Notations. The following notations will be used throughout. For any vector $\mathbf{v} \in \mathbb{R}^d$, $\mathbf{v}_{[j]}$ denotes the j^{th} coordinate of $\mathbf{v}$ $\forall\ 1 \leq j \leq d$. For any $1 \leq p < \infty$, $\|\mathbf{v}\|_p := (\sum_{j=1}^{d} |\mathbf{v}_{[j]}|^p)^{1/p}$ denotes the L_p norm of $\mathbf{v}$, and $\|\mathbf{v}\|_\infty := \max\{|\mathbf{v}_{[j]}| : j = 1, \ldots, d\}$ denotes the L_∞ norm of $\mathbf{v}$. For any matrix $\mathbf{M} \in \mathbb{R}^{d\times d}$, $\mathbf{M}_{[ij]}$ denotes the $(i,j)^{th}$ entry of $\mathbf{M}$ $\forall\ 1 \leq i, j \leq d$, and $||\mathbf{M}||_2 := \sup_{\|\mathbf{v}\|_2 \leq 1} \|\mathbf{M}v\|_2$, $||\mathbf{M}||_\infty := \max\limits_{1\leq i\leq d} \sum_{j=1}^{d} |\mathbf{M}_{[ij]}|$ and $||\mathbf{M}||_{\max} := \max\limits_{1\leq i,j\leq d} |\mathbf{M}_{[ij]}|$ respectively denote the spectral norm, the matrix-L_∞ norm and the maximum norm of $\mathbf{M}$. Further, we denote any symmetric positive definite (p.d.) matrix $\mathbf{M} \in \mathbb{R}^{d\times d}$ as $\mathbf{M} \succ 0$, and its minimum and maximum eigenvalues (also singular values) as $\lambda_{\min}(\mathbf{M}) > 0$ and $\lambda_{\max}(\mathbf{M}) \equiv ||\mathbf{M}||_2 > 0$ respectively.

Basic Set-Up and Definitions. Let $\mathcal{D}_n := \{\mathbf{Z}_i \equiv (Y_i, \mathbf{X}_i)\}_{i=1}^n$ denote the observed data consisting of n i.i.d. realizations of $\mathbf{Z} := (Y, \mathbf{X})$, where $Y \in \mathbb{R}$, $\mathbf{X} \in \mathbb{R}^p$ and $p \equiv p_n$ is allowed to diverge with n. Let $\mu_Y := \mathbb{E}(Y)$, $\boldsymbol{\mu} := \mathbb{E}(\mathbf{X})$ and $\boldsymbol{\Sigma} := \mathbb{E}\{(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\} \equiv \mathrm{Cov}(\mathbf{X})$, where we assume $\boldsymbol{\Sigma} \succ \mathbf{0}$. Let $\overline{Y} := n^{-1} \sum_{i=1}^n Y_i$, $\overline{\mathbf{X}} := n^{-1} \sum_{i=1}^n \mathbf{X}_i$ and $\mathcal{J} := \{1, \ldots, p\}$.

For any $J \subseteq \mathcal{J}$, let Ω_J denote the power set (the collection of all possible subsets) of J. For any $S \in \Omega_{\mathcal{J}}$ with $|S| = s \leq p$, and for any vector $\mathbf{v} \in \mathbb{R}^p$, let $\mathbf{v}_S \in \mathbb{R}^s$ denote the restriction of $\mathbf{v}$ onto S, i.e. for $S = \{i_1, \ldots, i_s\} \subseteq \mathcal{J}$, $\mathbf{v}_{S[j]} = \mathbf{v}_{[i_j]}$ $\forall\ 1 \leq j \leq s$. For any $S \in \Omega_{\mathcal{J}}$, let $\mathbf{X}_S$, $\{\mathbf{X}_{S,i}\}_{i=1}^n$, $\boldsymbol{\mu}_S$ and $\overline{\mathbf{X}}_S$ respectively denote the restrictions of $\mathbf{X}$, $\{\mathbf{X}_i\}_{i=1}^n$, $\boldsymbol{\mu}$ and $\overline{\mathbf{X}}$ onto S, and let $\boldsymbol{\Sigma}_S := \mathrm{Cov}(\mathbf{X}_S)$ and $\boldsymbol{\Sigma}_{S,Y} := \mathrm{Cov}(Y, \mathbf{X}_S)$. Further, for any $S \in \Omega_{\mathcal{J}}$, define:

$$\begin{aligned}
\widehat{\boldsymbol{\Sigma}}_S &:= \frac{1}{n}\sum_{i=1}^n (\mathbf{X}_{S,i} - \overline{\mathbf{X}}_S)(\mathbf{X}_{S,i} - \overline{\mathbf{X}}_S)^T, \;\; \widehat{\boldsymbol{\Gamma}}_S \;:=\; (\overline{\mathbf{X}}_S - \boldsymbol{\mu}_S)(\overline{\mathbf{X}}_S - \boldsymbol{\mu}_S)^T, \\
\widetilde{\boldsymbol{\Sigma}}_S &:= \frac{1}{n}\sum_{i=1}^n (\mathbf{X}_{S,i} - \boldsymbol{\mu}_S)(\mathbf{X}_{S,i} - \boldsymbol{\mu}_S)^T \;\equiv\; \widehat{\boldsymbol{\Sigma}}_S + \widehat{\boldsymbol{\Gamma}}_S; \quad \text{and} \\
\widehat{\boldsymbol{\Sigma}}_{S,Y} &:= \frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y})(\mathbf{X}_{S,i} - \overline{\mathbf{X}}_S), \;\; \widehat{\boldsymbol{\Gamma}}_{S,Y} \;:=\; (\overline{Y} - \mu_Y)(\overline{\mathbf{X}}_S - \boldsymbol{\mu}_S), \\
\widetilde{\boldsymbol{\Sigma}}_{S,Y} &:= \frac{1}{n}\sum_{i=1}^n (Y_i - \mu_Y)(\mathbf{X}_{S,i} - \boldsymbol{\mu}_S) \;\equiv\; \widehat{\boldsymbol{\Sigma}}_{S,Y} + \widehat{\boldsymbol{\Gamma}}_{S,Y}.
\end{aligned}$$

Varying Subset Linear Regression Estimator(s). Let $\mathcal{S} \subseteq \Omega_{\mathcal{J}}$ denote any collection of subsets of $\mathcal{J}$ such that $\max\{s := |S| : S \in \mathcal{S}\} \leq q_n$ for some $q_n \equiv q_{n,\mathcal{S}} \leq \min(n, p_n)$ and let $L_n \equiv L_{n,\mathcal{S}} := |\mathcal{S}|$. We now consider linear regressions of Y on $\mathbf{X}_S$, for all $S \in \mathcal{S}$, via the ordinary least squares (OLS) *estimator* $\widehat{\boldsymbol{\beta}}_S$ which we formally define, along with its corresponding *target parameter* $\boldsymbol{\beta}_S$, as follows. For any $S \in \mathcal{S}$, define:

$$\boldsymbol{\beta}_S \ := \ \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^s} \mathbb{E}[\{(Y - \mu_Y) - (\mathbf{X}_S - \boldsymbol{\mu}_S)^T \boldsymbol{\beta}\}^2] \ \equiv \ \boldsymbol{\Sigma}_S^{-1} \boldsymbol{\Sigma}_{S,Y}, \tag{5.1}$$

and let $\widehat{\boldsymbol{\beta}}_S$ be the corresponding OLS estimator given by:

$$\widehat{\boldsymbol{\beta}}_S \ := \ \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^s} \frac{1}{n} \sum_{i=1}^{n} \{(Y_i - \overline{Y}) - (\mathbf{X}_{S,i} - \overline{\mathbf{X}}_S)^T \boldsymbol{\beta}\}^2 \ \equiv \ \widehat{\boldsymbol{\Sigma}}_S^{-1} \widehat{\boldsymbol{\Sigma}}_{S,Y}. \tag{5.2}$$

Since we are only interested in the vector of regression coefficients, we circumvent the need for any nuisance intercept terms by appropriately centering Y and $\mathbf{X}_S$ in both (5.1) and (5.2). The existence and uniqueness of $\boldsymbol{\beta}_S \equiv \boldsymbol{\Sigma}_S^{-1} \boldsymbol{\Sigma}_{S,Y}$ and $\widehat{\boldsymbol{\beta}}_S \equiv \widehat{\boldsymbol{\Sigma}}_S^{-1} \widehat{\boldsymbol{\Sigma}}_{S,Y}$ in (5.1)-(5.2) are both guaranteed for any $S \in \mathcal{S}$ since $\boldsymbol{\Sigma} \succ \mathbf{0}$, so that $\boldsymbol{\Sigma}_S$ is invertible, and $\widehat{\boldsymbol{\Sigma}}_S$ is invertible almost surely (a.s.) as $|S| \equiv s \leq n$. Further note that, throughout the formulations in (5.1)-(5.2), we make *no* assumptions on the existence of a true linear model between Y and $\mathbf{X}_S$ for any S. The target parameter $\boldsymbol{\beta}_S$ is well-defined *regardless* of any such model assumptions and simply denotes the coefficients in the best (in the L_2 sense) linear predictor of Y given $\mathbf{X}_S$.

Decomposition of $(\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S)$. For notational simplicity, let us define: $\widetilde{\mathbf{X}}_S := (\mathbf{X}_S - \boldsymbol{\mu}_S)$ and $\widetilde{Y} := (Y - \mu_Y)$. Using the estimating equations in (5.1)-(5.2), it is then straightforward to show that $(\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S)$, for any $S \in \mathcal{S}$, satisfies a deterministic decomposition as follows:

$$(\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S) \ = \ \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\Psi}_S(\mathbf{Z}_i) + (\mathbf{T}_{n,S} + \mathbf{R}_{n,S}), \quad \text{where} \tag{5.3}$$

$$\boldsymbol{\Psi}_S(\mathbf{Z}) \ := \ \boldsymbol{\Sigma}_S^{-1} \boldsymbol{\psi}_S(\mathbf{Z}) \ \equiv \ \boldsymbol{\Sigma}_S^{-1} \widetilde{\mathbf{X}}_S (\widetilde{Y} - \widetilde{\mathbf{X}}_S^T \boldsymbol{\beta}_S) \ \text{ with } \ \mathbb{E}\{\boldsymbol{\psi}_S(\mathbf{Z})\} = \mathbf{0},$$

$$\mathbf{T}_{n,S} \ := \ \frac{1}{n} (\widehat{\boldsymbol{\Sigma}}_S^{-1} - \boldsymbol{\Sigma}_S^{-1}) \sum_{i=1}^{n} \boldsymbol{\psi}_S(\mathbf{Z}_i) \ \equiv \ (\widehat{\boldsymbol{\Sigma}}_S^{-1} - \boldsymbol{\Sigma}_S^{-1})(\widetilde{\boldsymbol{\Sigma}}_{S,Y} - \widetilde{\boldsymbol{\Sigma}}_S \boldsymbol{\beta}_S),$$

$$\text{and } \ \mathbf{R}_{n,S} \ := \ \widehat{\boldsymbol{\Sigma}}_S^{-1} (\widehat{\boldsymbol{\Gamma}}_S \boldsymbol{\beta}_S) - \widehat{\boldsymbol{\Sigma}}_S^{-1} \widehat{\boldsymbol{\Gamma}}_{S,Y} \ =: \ \mathbf{R}_{n,S}^{(1)} - \mathbf{R}_{n,S}^{(2)} \ \text{(say)}.$$

For a *single* S and under classical asymptotics (i.e. s is fixed and $n \to \infty$), standard results from M-estimation theory (see for instance Van der Vaart [1998]; Van der Vaart and Wellner [1996]) imply that under very mild regularity conditions, $\widehat{\boldsymbol{\beta}}_S$ is a $\sqrt{n}$-consistent and asymptotically normal (CAN) estimator of $\boldsymbol{\beta}_S$ and admits an asymptotically linear expansion (ALE):

$(\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S) = n^{-1}\sum_{i=1}^n \boldsymbol{\Psi}_S(\mathbf{Z}_i) + o_{\mathbb{P}}(n^{-1/2})$, with influence function (IF) $\boldsymbol{\Psi}_S(\mathbf{Z})$, so that $n^{1/2}(\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S)$ is asymptotically normal with mean $\mathbf{0}$ and variance $\mathrm{Cov}\{\boldsymbol{\Psi}_S(\mathbf{Z})\}$. Further, even when $s \equiv |S|$ is allowed to diverge, it is also well known (see for instance Portnoy [1984, 1985, 1986, 1988]) that under suitable regularity conditions and if $s = o(n)$, $\left\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S\right\|_2 = O_{\mathbb{P}}(s^{1/2}/n^{1/2})$ and $\left\|(\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S) - n^{-1}\sum_{i=1}^n \boldsymbol{\Psi}_S(\mathbf{Z}_i)\right\|_2 = O_{\mathbb{P}}(s/n)$, so that whenever $s = o(n^{1/2})$, $\widehat{\boldsymbol{\beta}}_S$ is a CAN estimator of $\boldsymbol{\beta}_S$ and admits an ALE with IF $\boldsymbol{\Psi}_S(\mathbf{Z})$. However, these results are all asymptotic and more importantly, apply only to a single S.

Our main challenges stem from the fact(s) that we have a *family* of estimators based on a collection of subsets $\{\mathbf{X}_S\}_{S\in\mathcal{S}}$ of $\mathbf{X}$, where $|\mathcal{S}|$ itself is possibly large and further, for each $S \in \mathcal{S}$, $\mathbf{X}_S$ may be high-dimensional with $s \leq q_n$ allowed to diverge with n. Despite such a multi-fold large scale setting, we desire/need to provide inferential tools for our family of estimators $\{\widehat{\boldsymbol{\beta}}_S\}_{S\in\mathcal{S}}$ and their derived functionals. We achieve this by providing (uniform) ALEs for $\{\widehat{\boldsymbol{\beta}}_S\}_{S\in\mathcal{S}}$, whereby we control the remainder terms $\mathbf{T}_{n,S}$ and $\mathbf{R}_{n,S}$ in (5.3) *uniformly over* $S \in \mathcal{S}$ based on non-asymptotic bounds for: $\sup_{S\in\mathcal{S}} \|\mathbf{T}_{n,S} + \mathbf{R}_{n,S}\|_2$ establishing their uniform convergence rates. Note that the potentially diverging sizes of $\mathcal{S}$ and each $S \in \mathcal{S}$ necessitate such non-asymptotic analyses. Lastly, apart from the (second order) error terms $\mathbf{T}_{n,S}$ and $\mathbf{R}_{n,S}$, we also provide uniform convergence rates of the first order term: $n^{-1}\sum_{i=1}^n \boldsymbol{\Psi}_S(\mathbf{Z}_i)$ in (5.3) under the L_2 norm, thereby establishing the rate of $\sup_{S\in\mathcal{S}} \|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S\|_2$. Further, for linear functionals of $\{\boldsymbol{\beta}_S\}_{S\in\mathcal{S}}$, we also provide results on $\sqrt{n}$-consistency and asymptotic normality for the corresponding linear functionals of the estimators $\{\widehat{\boldsymbol{\beta}}_S\}_{S\in\mathcal{S}}$. Such results would be useful for establishing our results in Section 6 regarding asymptotic distribution of the IDA based estimators.

It is worth mentioning that some results on 'uniform-in-model' bounds, similar in flavor to those mentioned above, were also obtained independently in the recent work of Kuchibhotla et al. [2018] on post-selection inference in linear regression, although their results are not directly comparable to ours. We refer to Section 9 for further discussions on their results.

5.1. *Uniform ALEs for OLS: Non-Asymptotic Bounds and Uniform Convergence Rates for All Terms in (5.3).* We first state our main assumptions and define a few related quantities that will appear in our results. Throughout this section, all notations, definitions and conditions introduced so far in Section 5.1 will be automatically adopted without any special mention.

ASSUMPTION 5.1 (Main assumptions and some definitions). *(i) We assume that* $\widetilde{Y} \equiv (Y - \mu_Y)$ *is sub-Gaussian and* $\widetilde{\mathbf{X}}_S \equiv (\mathbf{X}_S - \boldsymbol{\mu}_S)$ *is sub-Gaussian uniformly in* $S \in \mathcal{S}$. *We also assume that* $\boldsymbol{\Sigma}_S$ *is well conditioned, in terms of its extremal eigenvalues, uniformly in* $S \in \mathcal{S}$. *Specifically, for some constants* $\sigma_Y, \sigma_{\mathbf{X},\mathcal{S}} \in [0,\infty)$ *and* $0 < \lambda_{\inf,\mathcal{S}} \leq \lambda_{\sup,\mathcal{S}} < \infty$, *we assume*

$$\|Y - \mu_Y\|_{\psi_2} \leq \sigma_Y, \quad \sup_{S\in\mathcal{S}} \|\mathbf{X}_S - \boldsymbol{\mu}_S\|^*_{\psi_2} \leq \sigma_{\mathbf{X},\mathcal{S}}, \quad \text{and}$$

$$0 < \lambda_{\inf,\mathcal{S}} \leq \inf_{S\in\mathcal{S}} \lambda_{\min}(\boldsymbol{\Sigma}_S) \leq \sup_{S\in\mathcal{S}} \lambda_{\max}(\boldsymbol{\Sigma}_S) \leq \lambda_{\sup,\mathcal{S}} < \infty, \tag{5.4}$$

where $\|\cdot\|_{\psi_2}$ *and* $\|\cdot\|^*_{\psi_2}$ *denote the sub-Gaussian norms as defined in A.1 and A.2 respectively. Let us further define the constant* $K_{\mathcal{S}} := C_1\sigma^2_{\mathbf{X},\mathcal{S}}\frac{\lambda_{\sup,\mathcal{S}}}{\lambda_{\inf,\mathcal{S}}} > 0$, *where* C_1 *is the same absolute constant as in Lemma C.6 (and also same as the constant given in Theorem 4.7.1 and Exercise 4.7.3 of Vershynin [2018]).*

(ii) Let $\mathbf{Z}_S := (Y, \mathbf{X}_S)$, $\boldsymbol{\nu}_S := \mathbb{E}(\mathbf{Z}_S)$ *and* $\boldsymbol{\Xi}_S := \mathrm{Cov}(\mathbf{Z}_S)$. *Then, we also assume that* $\boldsymbol{\Xi}_S$ *is well-conditioned uniformly in* $S \in \mathcal{S}$. *Specifically, for some constants* $\widetilde{\lambda}_{\inf,\mathcal{S}}, \widetilde{\lambda}_{\sup,\mathcal{S}} \in (0,\infty)$,

$$0 < \widetilde{\lambda}_{\inf,\mathcal{S}} \leq \inf_{S\in\mathcal{S}} \lambda_{\min}(\boldsymbol{\Xi}_S) \leq \sup_{S\in\mathcal{S}} \lambda_{\max}(\boldsymbol{\Xi}_S) \leq \widetilde{\lambda}_{\sup,\mathcal{S}} < \infty.$$

Further, let $\widetilde{\sigma}_{\mathbf{Z},\mathcal{S}} := (\sigma_Y + \sigma_{\mathbf{X},\mathcal{S}})$ *and define the constant* $\widetilde{K}_{\mathcal{S}} := C_1\widetilde{\sigma}^2_{\mathbf{Z},\mathcal{S}}\frac{\widetilde{\lambda}_{\sup,\mathcal{S}}}{\widetilde{\lambda}_{\inf,\mathcal{S}}} > 0$, *where* $C_1 > 0$ *is the same absolute constant as in part (i) above.*

We present our main results in Theorem 5.1 below. Its proof also involves two useful supporting lemmas, Lemma B.2 and Lemma B.3, which may be of independent interest themselves. These lemmas are given in Section B.4.

THEOREM 5.1 (Uniform bounds and convergence rates for all the terms in (5.3)). *Consider any* $\mathcal{S} \subseteq \Omega_{\mathcal{J}}$ *with* $|\mathcal{S}| := L_n \equiv L_{n,\mathcal{S}}$ *and* $\sup_{S\in\mathcal{S}} |S| \leq q_n \equiv q_{n,\mathcal{S}} \leq \min(n, p_n)$, *and suppose Assumption 5.1 holds. Let* $r_n := q_n + \log L_n$, $\widetilde{r}_n := r_n + 1$ *and* $C_{\mathcal{S}} := \sqrt{2}\sigma_Y\lambda^{-1/2}_{\inf,\mathcal{S}}$. *For any* $c > 0$, *let* $\bar{c} := c + 1$ *and define:*

$$\epsilon_{n,1}(c, r_n) := \bar{c}K_{\mathcal{S}}\left(\sqrt{\frac{r_n}{n}} + \frac{r_n}{n}\right), \quad \epsilon_{n,2}(c, r_n) := \bar{c}\widetilde{K}_{\mathcal{S}}\left(\sqrt{\frac{\widetilde{r}_n}{n}} + \frac{\widetilde{r}_n}{n}\right);$$

$$\eta_{n,1}(c, r_n) := 32\bar{c}K_{\mathcal{S}}\frac{r_n}{n} + \frac{\lambda_{\sup,\mathcal{S}}}{n}, \quad \eta_{n,2}(c, r_n) := 32\bar{c}\widetilde{K}_{\mathcal{S}}\frac{\widetilde{r}_n}{n} + \frac{\widetilde{\lambda}_{\sup,\mathcal{S}}}{n}; \quad \text{and}$$

$$\delta_n(c, r_n) := \bar{c}K^*_{\mathcal{S}}\left(\sqrt{\frac{r_n}{n}} + \frac{33r_n}{n}\right) + \frac{2}{n}\frac{\lambda_{\sup,\mathcal{S}}}{\lambda^2_{\inf,\mathcal{S}}},$$

where $(K_{\mathcal{S}}, \widetilde{K}_{\mathcal{S}}, \lambda_{\sup,\mathcal{S}}, \lambda_{\inf,\mathcal{S}}, \widetilde{\lambda}_{\sup,\mathcal{S}}, \widetilde{\lambda}_{\inf,\mathcal{S}}, \sigma_Y)$ *are as in Assumption 5.1 and* $K^*_{\mathcal{S}} := 2\lambda^{-2}_{\inf,\mathcal{S}} K_{\mathcal{S}}$*. Further, let* $c^* > 0$ *be any constant that satisfies:*

$$(c^* + 1)K_{\mathcal{S}} \left(\sqrt{\frac{r_n}{n}} + \frac{33 r_n}{n} \right) + \frac{\lambda_{\sup,\mathcal{S}}}{n} \;\leq\; \frac{1}{2}\lambda_{\inf,\mathcal{S}}.$$

(i) Then, for any such constant $c^* > 0$*, and for any* $c > 0$*, we have the following bounds. With probability at least* $1 - 8\exp(-c r_n) - 4\exp(-c^* r_n)$*,*

$$\begin{aligned}
\sup_{S \in \mathcal{S}} \|\mathbf{T}_{n,S}\|_2 &\leq \delta_n(c, r_n)\left\{\epsilon_{n,1}(c, r_n) C_{\mathcal{S}} + \epsilon_{n,2}(c, r_n)\right\} \\
&\lesssim \mathbb{C}_{\mathcal{S}}(c+1)^2 \frac{r_n}{n}\left\{1 + o(1)\right\}, \quad \textit{and} \\
\sup_{S \in \mathcal{S}} \|\mathbf{R}_{n,S}\|_2 &\leq \{\delta_n(c, r_n) + \lambda^{-1}_{\inf,\mathcal{S}}\}\{\eta_{n,1}(c, r_n) C_{\mathcal{S}} + \eta_{n,2}(c, r_n)\} \\
&\lesssim \mathbb{C}_{\mathcal{S}}(c+1)\frac{r_n}{n}\{1 + o(1)\}.
\end{aligned}$$

(ii) Further, for any $c > 0$*, the first order term* $n^{-1}\sum_{i=1}^n \boldsymbol{\Psi}_S(\mathbf{Z}_i)$ *in (5.3) satisfies the following bound. With probability at least* $1 - 4\exp(-c r_n)$*,*

$$\begin{aligned}
\sup_{S \in \mathcal{S}} \left\| \frac{1}{n}\sum_{i=1}^n \boldsymbol{\Psi}_S(\mathbf{Z}_i) \right\|_2 &\leq \lambda^{-1}_{\inf,\mathcal{S}}\left\{\epsilon_{n,1}(c, r_n) C_{\mathcal{S}} + \epsilon_{n,2}(c, r_n)\right\} \\
&\lesssim \mathbb{C}_{\mathcal{S}}(c+1)\sqrt{\frac{r_n}{n}}\left\{1 + o(1)\right\},
\end{aligned}$$

Here, $\mathbb{C}_{\mathcal{S}}$ *denotes a generic constant (possibly different in each bound) depending only on* $\mathcal{S}$*, and '* $\lesssim$*' denotes inequality upto multiplicative constants.*

REMARK 5.1. The two bounds in result (i) of Theorem 5.1 also imply in particular that with high probability,

$$\sup_{S \in \mathcal{S}} \left\| (\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S) - \frac{1}{n}\sum_{i=1}^n \boldsymbol{\Psi}_S(\mathbf{Z}_i) \right\|_2 \;\equiv\; \sup_{S \in \mathcal{S}} \|\mathbf{T}_{n,S} + \mathbf{R}_{n,S}\|_2 \;\lesssim\; \frac{r_n}{n},$$

thereby (non-asymptotically) establishing the uniform (in $S \in \mathcal{S}$) convergence rates (under the L_2 norm) of the second order terms in the ALE (5.3) of $(\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S)$ to be $O_{\mathbb{P}}(r_n/n)$. Further, the bound in result (ii) establishes (non-asymptotically) the uniform (in $S \in \mathcal{S}$) convergence rate (under the L_2 norm) of the first order term in the ALE (5.3) to be $O_{\mathbb{P}}(\sqrt{r_n/n})$. As a consequence, it also establishes that $\sup_{S \in \mathcal{S}} \left\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S\right\|_2 = O_{\mathbb{P}}(\sqrt{r_n/n} + r_n/n)$.

Remark 5.2 (ALEs and asymptotic normality for linear functionals of $\{\widehat{\boldsymbol\beta}_S\}_{S\in\mathcal{S}}$). Let $\mathcal{A}_\mathcal{S} := \{\mathbf{a}_S \in \mathbb{R}^s : S \in \mathcal{S}\}$ be any collection of (known) vectors with $\sum_{S\in\mathcal{S}} \|\mathbf{a}_S\|_2 = O(1)$. Consider the functional given by $\beta(\mathcal{A}_\mathcal{S}) := \sum_{S\in\mathcal{S}} \mathbf{a}_S^T \boldsymbol\beta_S$ and its estimator $\widehat\beta(\mathcal{A}_\mathcal{S}) := \sum_{S\in\mathcal{S}} \mathbf{a}_S^T \widehat{\boldsymbol\beta}_S$. Then, as a direct consequence of Theorem 5.1, $\widehat\beta(\mathcal{A}_\mathcal{S}) - \beta(\mathcal{A}_\mathcal{S})$ satisfies the following ALE.

$$\sqrt{n}\{\widehat\beta(\mathcal{A}_\mathcal{S}) - \beta(\mathcal{A}_\mathcal{S})\} \;=\; \frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_{\mathcal{A}_\mathcal{S}}(\mathbf{Z}_i) + \mathcal{R}_{n,\mathcal{A}_\mathcal{S}} \;=:\; \sqrt{n}\mathbb{S}_{n,\mathcal{A}_\mathcal{S}} + \mathcal{R}_{n,\mathcal{A}_\mathcal{S}},$$

$$\text{where } \xi_{\mathcal{A}_\mathcal{S}}(\mathbf{Z}) := \sum_{S\in\mathcal{S}} \mathbf{a}_S^T \boldsymbol\Psi_S(\mathbf{Z}) \text{ and } \mathcal{R}_{n,\mathcal{A}_\mathcal{S}} := \sum_{S\in\mathcal{S}} \mathbf{a}_S^T(\mathbf{T}_{n,S} + \mathbf{R}_{n,S}),$$

$$\text{with } |\mathcal{R}_{n,\mathcal{A}_\mathcal{S}}| \;\le\; \sup_{S\in\mathcal{S}} \|\mathbf{T}_{n,S} + \mathbf{R}_{n,S}\|_2 \left(\sum_{S\in\mathcal{S}} \|\mathbf{a}_S\|_2\right) \;=\; O_\mathbb{P}\left(\frac{r_n}{\sqrt{n}}\right). \quad \square$$

Thus, $\sqrt{n}\{\widehat\beta(\mathcal{A}_\mathcal{S}) - \beta(\mathcal{A}_\mathcal{S})\} = \sqrt{n}\mathbb{S}_{n,\mathcal{A}_\mathcal{S}} + o_\mathbb{P}(1)$, as long as $r_n = o(\sqrt{n})$ and $\sum_{S\in\mathcal{S}} \|\mathbf{a}_S\|_2 = O(1)$. Note that $\mathbb{S}_{n,\mathcal{A}_\mathcal{S}}$ is an average of the centered i.i.d. random variables $\{\xi_{\mathcal{A}_\mathcal{S}}(\mathbf{Z}_i)\}_{i=1}^n$. Hence, $\sqrt{n}\mathbb{S}_{n,\mathcal{A}_\mathcal{S}}$, when appropriately scaled to have unit variance, is expected to converge to a $\mathcal{N}(0,1)$ distribution under suitable Lyapunov-type moment conditions on $\xi_{\mathcal{A}_\mathcal{S}}(\mathbf{Z})$. We characterize this more explicitly through a stronger non-asymptotic statement using the Berry-Esseen bound. Let $\sigma^2_{\xi_{\mathcal{A}_\mathcal{S}}} := \mathbb{E}[\{\xi_{\mathcal{A}_\mathcal{S}}(\mathbf{Z})\}^2] \equiv \mathrm{Var}\{\xi_{\mathcal{A}_\mathcal{S}}(\mathbf{Z})\}$ and $\rho^3_{\xi_{\mathcal{A}_\mathcal{S}}} := \mathbb{E}\{|\xi_{\mathcal{A}_\mathcal{S}}(\mathbf{Z})|^3\}$ with $0 < \sigma_{\xi_{\mathcal{A}_\mathcal{S}}} \le \rho_{\xi_{\mathcal{A}_\mathcal{S}}}$, and assume that $\sigma_{\xi_{\mathcal{A}_\mathcal{S}}} = \Omega(1)$ and $\rho_{\xi_{\mathcal{A}_\mathcal{S}}} = O(1)$, so that $\rho_{\xi_{\mathcal{A}_\mathcal{S}}}/\sigma_{\xi_{\mathcal{A}_\mathcal{S}}} = O(1)$ (verifications of these conditions will be discussed later). Finally, let $F_{\mathbb{S}_{n,\mathcal{A}_\mathcal{S}}}(x) := \mathbb{P}(\sqrt{n}\mathbb{S}_{n,\mathcal{A}_\mathcal{S}}/\sigma_{\xi_{\mathcal{A}_\mathcal{S}}} \le x)\ \forall\, x \in \mathbb{R}$ denote the cumulative distribution function (CDF) of $\sqrt{n}\mathbb{S}_{n,\mathcal{A}_\mathcal{S}}/\sigma_{\xi_{\mathcal{A}_\mathcal{S}}}$ and let $\Phi(x) := \mathbb{P}(Z \le x)\ \forall\, x \in \mathbb{R}$, where $Z \sim \mathcal{N}(0,1)$, denote the standard normal CDF. Then, the Berry-Esseen theorem [Shevtsova, 2011] implies

$$\sup_{x\in\mathbb{R}} \left| F_{\mathbb{S}_{n,\mathcal{A}_\mathcal{S}}}(x) - \Phi(x) \right| \;\le\; \frac{0.48\rho^3_{\xi_{\mathcal{A}_\mathcal{S}}}}{\sqrt{n}\sigma^3_{\xi_{\mathcal{A}_\mathcal{S}}}} \;=\; O(n^{-1/2}),$$

and hence, $\frac{\sqrt{n}\mathbb{S}_{n,\mathcal{A}_\mathcal{S}}}{\sigma_{\xi_{\mathcal{A}_\mathcal{S}}}} \xrightarrow{d} \mathcal{N}(0,1)$. □

Thus, as long as $r_n = o(\sqrt{n})$, $\sum_{S\in\mathcal{S}} \|\mathbf{a}_S\|_2 = O(1)$, $\sigma_{\xi_{\mathcal{A}_\mathcal{S}}} = \Omega(1)$ and $\rho_{\xi_{\mathcal{A}_\mathcal{S}}} = O(1)$, we have: $\sqrt{n}\{\widehat\beta(\mathcal{A}_\mathcal{S}) - \beta(\mathcal{A}_\mathcal{S})\}/\sigma_{\xi_{\mathcal{A}_\mathcal{S}}} = \sqrt{n}\mathbb{S}_{n,\mathcal{A}_\mathcal{S}}/\sigma_{\xi_{\mathcal{A}_\mathcal{S}}} + o_\mathbb{P}(1)$ and $\sqrt{n}\mathbb{S}_{n,\mathcal{A}_\mathcal{S}}/\sigma_{\xi_{\mathcal{A}_\mathcal{S}}} \xrightarrow{d} \mathcal{N}(0,1)$. Invoking Slutsky's theorem, we therefore have: $\sqrt{n}\{\widehat\beta(\mathcal{A}_\mathcal{S}) - \beta(\mathcal{A}_\mathcal{S})\}/\sigma_{\xi_{\mathcal{A}_\mathcal{S}}} \xrightarrow{d} \mathcal{N}(0,1)$. Furthermore, for any consistent estimator $\widehat\sigma_{\xi_{\mathcal{A}_\mathcal{S}}}$ of $\sigma_{\xi_{\mathcal{A}_\mathcal{S}}}$, it also holds, via another application of Slutsky's theorem, that $\sqrt{n}\{\widehat\beta(\mathcal{A}_\mathcal{S}) - \beta(\mathcal{A}_\mathcal{S})\}/\widehat\sigma_{\xi_{\mathcal{A}_\mathcal{S}}} \xrightarrow{d} \mathcal{N}(0,1)$. □

We conclude the remark with some discussions regarding verifications of the moment conditions: $\rho_{\xi_{\mathcal{A}_\mathcal{S}}} = O(1)$ and $\sigma_{\xi_{\mathcal{A}_\mathcal{S}}} = \Omega(1)$. The first condition (and lot more) can be indeed verified generally under our basic assumptions in this section. To this end, note that under Assumption 5.1 (i), and through multiple uses of Lemma C.2 (i), Lemma C.1 (i) and Lemma C.1 (iv), as well as the last claim in Lemma B.2, we have: for each $S \in \mathcal{S}$,

$$\begin{aligned}
\left\|\mathbf{a}_S^T \mathbf{\Psi}_S(\mathbf{Z})\right\|_{\psi_1} &\leq \left\|\mathbf{a}_S^T \mathbf{\Sigma}_S^{-1} \widetilde{\mathbf{X}}_S\right\|_{\psi_2} \left\|\widetilde{Y} - \widetilde{\mathbf{X}}_S^T \boldsymbol{\beta}_S\right\|_{\psi_2} \\
&\leq \|\mathbf{a}_S\|_2 \{\lambda_{\min}(\mathbf{\Sigma}_S)\}^{-1} \|\mathbf{X}_S\|_{\psi_2}^* (\sigma_Y + \|\mathbf{X}_S\|_{\psi_2}^* \|\boldsymbol{\beta}_S\|_2) \\
&\leq \|\mathbf{a}_S\|_2 \left\{\lambda_{\inf,\mathcal{S}}^{-1} \sigma_{\mathbf{X},\mathcal{S}} (\sigma_Y + \sqrt{2}\sigma_{\mathbf{X},\mathcal{S}} \lambda_{\inf,\mathcal{S}}^{-1/2} \sigma_Y)\right\} \equiv \|\mathbf{a}_S\|_2 D_\mathcal{S} \text{ (say)},
\end{aligned}$$

where $D_\mathcal{S} := \lambda_{\inf,\mathcal{S}}^{-1} \sigma_{\mathbf{X},\mathcal{S}} \sigma_Y (1 + \sqrt{2}\sigma_{\mathbf{X},\mathcal{S}} \lambda_{\inf,\mathcal{S}}^{-1/2})$ depends only on the constants in Assumption 5.1 (i). Thus, as long as $D_\mathcal{S} = O(1)$ and $\sum_{S \in \mathcal{S}} \|\mathbf{a}_S\|_2 = O(1)$, as assumed before, we have:

$$\|\xi_{\mathcal{A}_\mathcal{S}}(\mathbf{Z})\|_{\psi_1} \equiv \left\|\sum_{S \in \mathcal{S}} \mathbf{a}_S^T \mathbf{\Psi}_S(\mathbf{Z})\right\|_{\psi_1} \leq D_\mathcal{S} \left(\sum_{S \in \mathcal{S}} \|\mathbf{a}_S\|_2\right) = O(1). \quad \square$$

Consequently, using Lemma C.1 (iii), we have $0 \leq \sigma_{\xi_{\mathcal{A}_\mathcal{S}}} \leq \rho_{\xi_{\mathcal{A}_\mathcal{S}}} \leq O(1)$. Among other implications, this also verifies one of the moment conditions.

Lastly, we provide some sufficient conditions for verifying the other moment condition: $\sigma_{\xi_{\mathcal{A}_\mathcal{S}}} = \Omega(1)$. To this end, suppose that for some positive constant $\eta_\mathcal{S} = \Omega(1)$, we have $\mathrm{Var}(Y \mid \cup_{S \in \mathcal{S}} \mathbf{X}_S) \geq \eta_\mathcal{S} > 0$. Then,

$$\begin{aligned}
\sigma_{\xi_{\mathcal{A}_\mathcal{S}}}^2 &\equiv \mathrm{Var}\{\xi_{\mathcal{A}_\mathcal{S}}(\mathbf{Z})\} \geq \mathbb{E}\left[\mathrm{Var}\left\{\xi_{\mathcal{A}_\mathcal{S}}(\mathbf{Z}) \mid \cup_{S \in \mathcal{S}} \mathbf{X}_S\right\}\right] \\
&\equiv \mathbb{E}\left[\mathrm{Var}\left\{\sum_{S \in \mathcal{S}} \mathbf{a}_S^T \mathbf{\Sigma}_S^{-1} \widetilde{\mathbf{X}}_S (\widetilde{Y} - \widetilde{\mathbf{X}}_S^T \boldsymbol{\beta}_S) \mid \cup_{S \in \mathcal{S}} \mathbf{X}_S\right\}\right] \\
= \mathbb{E}&\left\{\mathrm{Var}\left(Y \mid \cup_{S \in \mathcal{S}} \mathbf{X}_S\right) \left(\sum_{S \in \mathcal{S}} \mathbf{a}_S^T \mathbf{\Sigma}_S^{-1} \widetilde{\mathbf{X}}_S\right)^2\right\} \geq \eta_\mathcal{S} \mathbb{E}\left(\sum_{S \in \mathcal{S}} \mathbf{a}_S^T \mathbf{\Sigma}_S^{-1} \widetilde{\mathbf{X}}_S\right)^2.
\end{aligned}$$

Hence, as long as $\mathbb{E}\{(\sum_{S \in \mathcal{S}} \mathbf{a}_S^T \mathbf{\Sigma}_S^{-1} \widetilde{\mathbf{X}}_S)^2\} = \Omega(1)$, and $\eta_\mathcal{S} = \Omega(1)$ as assumed, we have $\sigma_{\xi_{\mathcal{A}_\mathcal{S}}}^2 = \Omega(1)$, as required in the second condition. $\square$

6. Asymptotic Distribution of the IDA Based Estimators.

Using the results from Section 5, we prove asymptotic linearity of the MIDA estimator under Assumptions 4.1, 4.2 and 4.3 and the following assumptions.

ASSUMPTION 6.1 (Structure learning consistency). *The estimated CPDA-G $\hat{\mathcal{C}}_0'$ in Algorithm 3.1 is a consistent estimator of $\mathcal{C}_0'$, i.e.* $\mathbb{P}(\hat{\mathcal{C}}_0' \neq \mathcal{C}_0') \longrightarrow 0$.

ASSUMPTION 6.2 (Bounds on the eigenvalues of covariance matrices). *For any $(q+3)\times(q+3)$ principal submatrix Σ of $\Sigma_0 = \mathrm{Cov}(\mathbf{X})$,*

$$C_2 \leq 1/\left|\left|\Sigma^{-1}\right|\right|_2 \leq ||\Sigma||_2 \leq C_3$$

for some absolute constants $C_2, C_3 > 0$, where $||\cdot||_2$ denotes the spectral norm and q is given by Assumption 4.3.

We emphasize that the results derived below does *not* depend on the nature of the CPDAG estimation procedure as long as it is consistent (Theorem 4.1 establishes this criteria for our particular CPDAG estimation procuedre). Assumption 6.2 is a slightly weaker version of Assumption 4.6, and unlike Assumption 4.6, it does not depend on the CPDAG estimation procedure.

For $j \in \{2, \ldots, p-1\}$, we define

$$E_{n,jp}(\hat{\mathcal{C}}_0', \mathcal{C}_0') := \mathrm{aver}(\hat{\Theta}_{jp}(\hat{\mathcal{C}}_0')) - \mathrm{aver}(\hat{\Theta}_{jp}(\mathcal{C}_0')),$$

where $\hat{\Theta}_{jp}(\mathcal{C}_0') := \{\hat{\beta}_{jp|\mathbf{Pa}_{\mathcal{G}'}(X_j)\cup\{X_1\}} : \mathcal{G}' \in \mathrm{MEC}(\mathcal{C}_0')\}$. Since $E_{n,jp}(\hat{\mathcal{C}}_0', \mathcal{C}_0') = 0$ whenever $\hat{\mathcal{C}}_0'$ and $\mathcal{C}_0'$ are identical, the consistency of the CPDAG estimation procedure implies that for any $\epsilon > 0$ and for *any* non-negative (and possibly diverging) sequence $\{a_n\}$ (e.g. $a_n = n$),

$$\mathbb{P}\left(a_n \left|E_{n,jp}(\hat{\mathcal{C}}_0', \mathcal{C}_0')\right| > \epsilon\right) \;\leq\; \mathbb{P}(\hat{\mathcal{C}}_0' \neq \mathcal{C}_0') \;\longrightarrow 0,$$

so that $E_{n,jp}(\hat{\mathcal{C}}_0', \mathcal{C}_0') = o_{\mathbb{P}}(a_n^{-1})$. This result allows us to use the uniform non-asymptotic theory developed in Section 5 for linear regression over *non-random* subsets of high-dimensional covariates, since

$$\begin{aligned}&\mathrm{aver}(\hat{\Theta}_{jp}(\hat{\mathcal{C}}_0')) - \mathrm{aver}(\Theta_{jp}(\mathcal{C}_0'))\\ &\quad = \; E_{n,jp}(\hat{\mathcal{C}}_0', \mathcal{C}_0') + \left\{\mathrm{aver}(\hat{\Theta}_{jp}(\mathcal{C}_0')) - \mathrm{aver}(\Theta_{jp}(\mathcal{C}_0'))\right\}.\end{aligned}$$

To present our results, we define the residual of the linear regression of X_i on $S \subseteq \{1, \ldots, p\} \setminus \{i\}$ as

$$R_{i|S} \;:=\; X_i - \mu_i - (\Sigma_0)_{iS}\left[(\Sigma_0)_{SS}\right]^{-1}(\mathbf{X}_S - \boldsymbol{\mu}_S),$$

where $\mathbf{X}_S$ is a vector consisting of $\{X_k : k \in S\}$, $(\Sigma_0)_{iS} = \mathrm{Cov}(X_i, \mathbf{X}_S)$ and $(\Sigma_0)_{SS} = \mathrm{Cov}(\mathbf{X}_S)$. Note that $(\Sigma_0)_{iS}\left[(\Sigma_0)_{SS}\right]^{-1}$ is the vector of regression coefficients in the linear regression of X_i on $\mathbf{X}_S$. Further, we define

$$Z_{jp} \;:=\; \frac{1}{L_j}\sum_{\ell=1}^{L_j} \mathbf{e}_{1,|S_{j\ell}|}^T \left((\Sigma_0)_{S_{j\ell}S_{j\ell}}\right)^{-1}(\mathbf{X}_{S_{j\ell}} - \boldsymbol{\mu}_{S_{j\ell}})R_{p|S_{j\ell}},$$

where $\{\mathbf{X}_{S_{j1}}, \ldots, \mathbf{X}_{S_{jL_j}}\}$ is the multi-set of vectors $\{(X_j, X_1, \mathbf{Pa}_{\mathcal{G}'}(X_j))^T : \mathcal{G}' \in \mathrm{MEC}(\mathcal{C}_0')\}$ containing $L_{\mathrm{distinct},j}$ distinct elements. For any random variable (or vector) Z, we will denote its n i.i.d. copies by $Z^{(1)}, \ldots, Z^{(n)}$. Finally, we define $q_j = |\mathbf{Adj}_{\mathcal{C}_0'}(X_j)|$. Note that by Assumption 4.3, we have $q_j \le q = O(n^{1-b_1})$ for some $0 < b_1 \le 1$.

THEOREM 6.1. *Under Assumptions 4.1, 4.2, 4.3 and 6.2, we have*

$$\begin{aligned} &\mathrm{aver}(\hat{\Theta}_{jp}(\hat{\mathcal{C}}_0')) - \mathrm{aver}(\Theta_{jp}(\mathcal{C}_0')) \\ &\qquad = E_{n,jp}(\hat{\mathcal{C}}_0', \mathcal{C}_0') + \frac{1}{n}\sum_{r=1}^{n} Z_{jp}^{(r)} + O_{\mathbb{P}}\left(\frac{q_j + \log(L_{\mathrm{distinct},j})}{n}\right), \text{ and} \end{aligned}$$

$$\begin{aligned} \hat{\eta}(\hat{\mathcal{C}}_0') - \eta(\mathcal{C}_0') \;&=\; \hat{\theta}_{1j}\; E_{n,jp}(\hat{\mathcal{C}}_0', \mathcal{C}_0') \\ &+ \frac{1}{n}\sum_{r=1}^{n}\left\{\theta_{1j} Z_{jp}^{(r)} + \mathrm{aver}(\Theta_{jp}(\mathcal{C}_0')) Z_{1j}^{(r)}\right\} + \left(\frac{1}{n}\sum_{r=1}^{n} Z_{jp}^{(r)}\right)\left(\frac{1}{n}\sum_{r=1}^{n} Z_{1j}^{(r)}\right) \\ &+ \hat{\theta}_{1j}\; O_{\mathbb{P}}\left(\frac{q_j + \log(L_{\mathrm{distinct},j})}{n}\right) + \mathrm{aver}(\hat{\Theta}_{jp}(\mathcal{C}_0'))\; O_{\mathbb{P}}\left(\frac{1}{n}\right). \end{aligned}$$

In order to establish asymptotic normality of the estimator of the total causal effect and the estimator of the individual mediation effects, we impose the following stronger sparsity condition.

ASSUMPTION 6.3 (Sparsity condition). *Let q_j and $L_{\mathrm{distinct,j}}$ be as above. Then we assume*

$$n^{-1/2}\{q_j + \log(L_{\mathrm{distinct},j})\} \longrightarrow 0.$$

We note that such stronger sparsity assumptions of a similar flavor have been frequently adopted in the literature, albeit for different but related problems, for deriving asymptotic normality results and confidence intervals in high-dimensional settings (see e.g. Portnoy [1988]; Van de Geer et al. [2014]; Javanmard and Montanari [2014]; Zhang and Zhang [2014]). To compare Assumption 6.3 with the sparsity condition given in Assumption 4.3, note that $q = O(n^{1-b_1})$ for some $1/2 < b_1 \le 1$ implies that the Assumption 6.3 holds, since

$$q_j + \log(L_{\mathrm{distinct},j}) \le q_j + \log(2^{q_j}) \le (1 + \log 2) \max_{2 \le j \le p-1} q_j = (1 + \log 2) q.$$

Additionally, we make the following assumption.

ASSUMPTION 6.4 (Nondegenrate conditional distributions). *The conditional variances satisfy the following lower bounds for some constant $v > 0$:*

1. $\mathrm{Var}(X_p \mid \mathbf{Adj}_{\mathcal{C}_0'}(X_j) \cup \{X_1, X_j\}) > v$ *and*
2. $\mathbb{E}\left[\mathrm{Var}(X_j \mid \mathbf{Adj}_{\mathcal{C}_0'}(X_j) \cup \{X_1\})\right] > v$.

We note that Assumption 6.4 resembles Assumption (F) of Maathuis, Kalisch and Bühlmann [2009]. Further, note that Assumption 6.4 follows from Assumption 6.2 when the error variables are normally distributed. This is because $\mathbf{X}$ is generated from an LSEM with normally distributed error variables implies that the joint distribution of $\mathbf{X}$ is multivariate Gaussian with covariance matrix Σ_0. Hence, for any $1 \le i \le p$ and $S \subseteq \{1, \ldots, p\} \setminus \{i\}$ such that $|S| \le q + 3$,

$$\mathrm{Var}(X_i \mid \mathbf{X}_S) = (\Sigma_0)_i - (\Sigma_0)_{iS}(\Sigma_0)_S^{-1}(\Sigma_0)_{Si} \ge \lambda_{\min}((\Sigma_0)_{S\cup\{i\}}) \ge C_2,$$

where the first inequality follows from the interlacing property of eigenvalues of a Hermitian matrix A and eigenvalues of the Schur complement of any principal submatrix of A (see, for example, Corollary 2.3 of Zhang [2005]), and the last inequality follows from Assumption 6.2.

COROLLARY 6.1 (Asymptotic normality of the estimator of the total causal effects). *Assume that* $\mathbb{P}(\hat{\mathcal{C}}_0' \neq \mathcal{C}_0') \longrightarrow 0$. *Then under Assumptions 4.1, 4.2, 6.2, 6.3 and 6.4, we have*

$$\frac{\sqrt{n}\left\{\mathrm{aver}(\hat{\Theta}_{jp}(\hat{\mathcal{C}}_0')) - \mathrm{aver}(\Theta_{jp}(\mathcal{C}_0'))\right\}}{\sqrt{\mathbb{E}[Z_{jp}^2]}} \xrightarrow{d} \mathcal{N}(0,1).$$

REMARK 6.1. Although we state Corollary 6.1 for our particular LSEM setting where $B_{j1} = B_{pj} = 0$ for all $j \in \{2, \ldots, p-1\}$, the same result holds for the original IDA estimator $\hat{\Theta}_{ik}(\hat{\mathcal{C}}_0) = \{\hat{\beta}_{ik|\mathbf{Pa}_{\mathcal{G}}(X_k)} : \mathcal{G} \in \mathrm{MEC}(\mathcal{C}_0)\}$, under the assumptions of Corollary 6.1.

COROLLARY 6.2 (Asymptotic normality of the estimator of the mediation effects). *Assume that* $\mathbb{P}(\hat{\mathcal{C}}_0' \neq \mathcal{C}_0') \longrightarrow 0$. *Then under Assumptions 4.1, 4.2, 6.2, 6.3 and 6.4, we have*

$$T_{jp} = \frac{\sqrt{n}\left(\hat{\eta}_{jp}(\hat{\mathcal{C}}_0') - \eta_{jp}(\mathcal{C}_0')\right)}{\sqrt{\theta_{1j}^2\,\mathbb{E}[Z_{jp}^2] + \mathrm{aver}(\Theta_{jp}(\mathcal{C}_0'))^2\,\mathbb{E}[Z_{1j}^2] + 2\theta_{1j}\mathrm{aver}(\Theta_{jp}(\mathcal{C}_0'))\,\mathbb{E}[Z_{jp}Z_{1j}]}} \xrightarrow{d} \mathcal{N}(0,1),$$

provided at least one of θ_{j1} *and* $\mathrm{aver}(\Theta_{jp}(\mathcal{C}_0'))$ *is nonzero.*

We conclude this section by addressing the following two major concerns for applying Corollary 6.2 in obtaining confidence intervals or in testing for zero mediation effects: (i) estimation of the asymptotic variance (ii) understanding the asymptotic behavior of $\hat{\eta}_{jp}(\hat{\mathcal{C}}_0')$ when both θ_{j1} and $\text{aver}(\Theta_{jp}(\mathcal{C}_0'))$ are zero.

To estimate the asymptotic variance, we propose to use the following plug-in estimator, $\widehat{\text{AVar}(\hat{\eta}_{jp})}$, given by:

$$\frac{1}{n}\sum_{r=1}^{n}\left\{\frac{\text{aver}(\hat{\Theta}_{jp}(\hat{\mathcal{C}}_0'))\ X_1^{(r)}\hat{R}_{j|\{1\}}^{(r)}}{(\hat{\Sigma}_0)_1}+\frac{\hat{\theta}_{1j}}{\hat{L}_j}\sum_{\ell=1}^{\hat{L}_j}\mathbf{e}_{1,|\hat{S}_\ell|}^T\left((\hat{\Sigma}_0)_{\hat{S}_\ell}\right)^{-1}\mathbf{X}_{\hat{S}_\ell}^{(r)}\hat{R}_{p|\hat{S}_\ell}^{(r)}\right\}^2,$$

where $\{\mathbf{X}_{\hat{S}_1},\ldots,\mathbf{X}_{\hat{S}_{\hat{L}_j}}\}$ is the multi-set of vectors $\{(X_j,X_1,\mathbf{Pa}_{\mathcal{G}'}(X_j))^T : \mathcal{G}'\in \text{MEC}(\hat{\mathcal{C}}_0')\}$, $\hat{\Sigma}_0$ is the sample covariance matrix of $\mathbf{X}$ and $\hat{R}_{p|S}^{(r)} := X_p^{(r)}-(\hat{\Sigma}_0)_{pS}\left((\hat{\Sigma}_0)_{SS}\right)^{-1}\mathbf{X}_S^{(r)}$, for $S\in\{\hat{S}_1,\ldots,\hat{S}_{\hat{L}_j}\}$ and $r=1,\ldots,n$.

When $\theta_{j1}=\text{aver}(\Theta_{jp}(\mathcal{C}_0'))=0$, it can be show that our test statistic

$$\hat{T}_{jp} := \frac{\sqrt{n}\left(\hat{\eta}_{jp}(\hat{\mathcal{C}}_0')-\eta_{jp}(\mathcal{C}_0')\right)}{\widehat{\text{AVar}(\hat{\eta}_{jp})}} \tag{6.1}$$

has a non-degenerate asymptotic distribution. For the sake of simplicity, we derive the asymptotic distribution by considering

$$\tilde{T}_{n,jp} := \frac{\sqrt{n}\left(\hat{\eta}_{jp}(\hat{\mathcal{C}}_0')-\eta_{jp}(\mathcal{C}_0')\right)}{\sqrt{\hat{\theta}_{1j}^2\,\mathbb{E}[Z_{jp}^2]+\text{aver}(\hat{\Theta}_{jp}(\mathcal{C}_0'))^2\,\mathbb{E}[Z_{1j}^2]+2\hat{\theta}_{1j}\text{aver}(\hat{\Theta}_{jp}(\mathcal{C}_0'))\,\mathbb{E}[Z_{jp}Z_{1j}]}}.$$

Note that $\tilde{T}_{jp}$ is always well-defined, and can be written as

$$\tilde{T}_{n,jp} := \frac{W_{n,1j}\ W_{n,jp}}{\sqrt{W_{n,1j}^2+W_{n,jp}^2+2\ \rho\ W_{n,1j}\ W_{n,jp}}},$$

where ρ is the correlation coefficient between Z_{jp} and Z_{1j}, and

$$(W_{n,1j},W_{n,jp})^T := \left(\frac{\sqrt{n}\ \hat{\theta}_{1j}}{\sqrt{\mathbb{E}[Z_{1j}^2]}},\frac{\sqrt{n}\ \text{aver}(\hat{\Theta}_{jp}(\mathcal{C}_0'))}{\sqrt{\mathbb{E}[Z_{jp}^2]}}\right)^T.$$

By following similar arguments as in Theorem 6.1 and in Corollary 6.2 it can be shown that

$$(W_{n,1j},W_{n,jp})^T \xrightarrow{d} (W_1,W_2)^T \sim \mathcal{N}\left(\mathbf{0},\ \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right).$$

Therefore, $\tilde{T}_{n,jp}$ is asymptotically distributed as

$$W := \frac{W_1 W_2}{\sqrt{W_1^2 + W_2^2 + 2\rho W_1 W_2}}. \tag{6.2}$$

The same result holds for our test statistic $\hat{T}_{jp}$. Figure 1 shows that the distribution of W is much more concentrated around zero compared to the standard normal distribution. Consequently, if we assume that the asymptotic distribution of $\hat{T}_{jp}$ is standard normal for all values of θ_{j1} and $\text{aver}(\Theta_{jp}(\mathcal{C}_0'))$, the resulting confidence interval would be conservative when both θ_{j1} and $\text{aver}(\Theta_{jp}(\mathcal{C}_0'))$ are zero or when they are very close to zero.

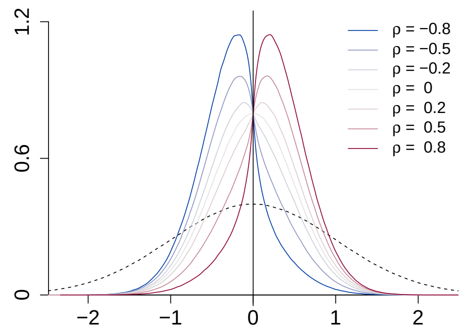

Fig 1: The probability density function of W for several values of ρ and the probability density function of the $\mathcal{N}(0,1)$ distribution (dashed line).

7. Simulations.

7.1. *Simulation Settings.* For each of the three settings given in Table 1, we use the R-package **pcalg** [Kalisch et al., 2012] to simulate r random weighted DAGs $\{\mathcal{G}'^{(1)}, \ldots, \mathcal{G}'^{(r)}\}$ with $p-2$ vertices $\{X_2, \ldots, X_{p-1}\}$ and $(p-2)d/2$ edges on average, where each pair of nodes in a randomly generated DAG has the probability $d/\binom{p-2}{2}$ of being adjacent (implying that the expected degree of each node is d). From each DAG $\mathcal{G}'^{(t)}$, we obtain the DAG $\mathcal{G}^{(t)}$, with p vertices $\{X_1, X_2 \ldots, X_{p-1}, X_p\}$ by randomly adding directed edges from X_1 to X_j with probability 0.2 and from X_j to X_p with probability 0.1, for $j = 1, \ldots, p-1$. The edge weights are drawn independently from a uniform distribution on $[-1, -0.5] \cup [0.5, 1]$.

Let $B_{\mathcal{G}^{(t)}}$ denote the weight matrix of the weighted DAG $\mathcal{G}^{(t)}$, i.e. $(B_{\mathcal{G}^{(t)}})_{ij} \neq 0$ if and only if the edge $X_i \to X_j$ is present in $\mathcal{G}^{(t)}$ and it then equals the corresponding edge weight. For $t = 1, \ldots, r$, the weight matrix $B_{\mathcal{G}^{(t)}}$

TABLE 1

Simulation settings where p denotes the number of potential mediators, d denotes the average degree in the causal DAG on the potential mediators and r denotes the number of randomly generated DAGs for a pair (p, d).

Setting	p	d	r
1	252	3	20
2	502	3.5	10
3	1002	4	5

and a random vector $\boldsymbol{\epsilon}^{(t)} = (\epsilon_1^{(t)}, \ldots, \epsilon_p^{(t)})^T$ define a distribution on $\mathbf{X}^{(t)} = (X_1^{(t)}, \ldots, X_p^{(t)})^T$ via the linear structural equation model

$$\mathbf{X}^{(t)} = B_{\mathcal{G}^{(t)}}^T \mathbf{X}^{(t)} + \boldsymbol{\epsilon}^{(t)}.$$

We choose $\epsilon_1^{(t)}, \ldots, \epsilon_p^{(t)}$ to be zero mean Gaussian random variables with variances independently drawn from a Uniform$[0.5, 1]$ distribution. Finally, we standardize the variables to have $\mathrm{Var}(X_i^{(t)}) = 1$, for all $i = 1, \ldots, p$.

For each setting, we generate 200 random samples $\{\mathcal{D}_{n1}^{(t)}, \ldots, \mathcal{D}_{n200}^{(t)}\}$ of size $n \in \{500, 1000, 5000\}$ from the joint distribution of $\mathbf{X}^{(t)}$ for $t = 1, \ldots, r$. For each $t \in \{1, \ldots, r\}$, we compute estimates and the corresponding p-values for $\eta_j^{(t)}$ based on the data $\mathcal{D}_{nk}^{(t)}$ using Algorithm 3.1 when the graph used in line 3 is either (i) an estimated CPDAG obtained by applying the ARGES algorithm [Nandy, Hauser and Maathuis, 2018], (ii) the true CPDAG $C_0'^{(t)}$, (iii) the true DAG $\mathcal{G}'^{(t)}$, or (iv) the empty graph. Note that the last case corresponds to a naive method that assumes that the potential mediators are conditionally independent given the treatment variable.

7.2. *Results.* As a finite sample validation of the asymptotic results obtained in Section 6, we record (from each replication) whether the true parameter value $\eta_j^{(t)}$ lies within the 95% standard Gaussian confidence interval constructed from the test statistic $\hat{T}_{jp}^{(t)}$ given by (6.1). To present the result, we split $\{\eta_j^{(t)} : t \in \{1, \ldots, r\},\ j \in \{2, \ldots, p-2\}\}$ into three equally sized groups according to the quantiles of $|\max(\theta_{1j}^{(t)}, \mathrm{aver}(\Theta_{jp}^{(t)}))|$'s distribution, and report the mean and median empirical coverage probabilities in each group. Tables 2 and 3 show that the 95% asymptotic confidence intervals exhibit an extremely high coverage in the first group where most of $|\max(\theta_{1j}^{(t)}, \mathrm{aver}(\Theta_{jp}^{(t)}))|$ equal zero, as well as in the second group where most of $|\max(\theta_{1j}^{(t)}, \mathrm{aver}(\Theta_{jp}^{(t)}))|$ are very close to zero. This is due to the fact that the correct asymptotic distribution of $\hat{T}_{jp}^{(t)}$ (see (6.2)) in these cases is much more concentrated around zero than the distribution of a standard

Gaussian random variable. The third group with reasonably high values of $|\max(\theta_{1j}^{(t)}, \text{aver}(\Theta_{jp}^{(t)}))|$ exhibits the correct coverage when the CPDAG is known (see Table 2), but we do see some loss of coverage when the CPDAG is estimated (see Table 3) due to finite sample graph estimation errors.

TABLE 2
Empirical coverage probabilities (mean and median) and average length of the 95% confidence intervals for each of the groups, when the true CPDAG is known.

p	n	$\lvert\max(\theta_{1j}^{(t)}, \text{aver}(\Theta_{jp}^{(t)}))\rvert$	Median Coverage	Mean Coverage	Length
252	500	Low	100 (0)	99.99 (0.07)	0.01
		Medium	99 (1.48)	98.25 (2.05)	0.04
		High	95 (1.48)	95.04 (1.62)	0.14
	1000	Low	100 (0)	99.99 (0.08)	0.01
		Medium	98 (2.22)	97.68 (2.07)	0.03
		High	95 (1.48)	95.08 (1.56)	0.10
	5000	Low	100 (0)	99.99 (0.08)	0.00
		Medium	96 (2.22)	96.37 (2.03)	0.01
		High	95 (1.48)	95.02 (1.52)	0.04
502	500	Low	100 (0)	99.98 (0.09)	0.01
		Medium	99.5 (0.74)	98.49 (2.08)	0.04
		High	95.5 (1.48)	95.14 (1.63)	0.14
	1000	Low	100 (0)	99.99 (0.07)	0.01
		Medium	98.5 (2.22)	97.94 (2.17)	0.03
		High	95 (1.48)	95.04 (1.53)	0.10
	5000	Low	100 (0)	99.99 (0.08)	0.00
		Medium	96.5 (2.22)	96.61 (2.13)	0.01
		High	95 (1.48)	94.97 (1.55)	0.04
1002	500	Low	100 (0)	99.98 (0.11)	0.01
		Medium	98.5 (2.22)	98.02 (2.12)	0.05
		High	95 (1.48)	95.14 (1.57)	0.15
	1000	Low	100 (0)	99.98 (0.14)	0.01
		Medium	97.5 (2.97)	97.38 (2.28)	0.04
		High	95 (1.48)	95.1 (1.51)	0.11
	5000	Low	100 (0)	99.92 (0.38)	0.00
		Medium	96 (2.22)	96.03 (1.93)	0.02
		High	95 (1.48)	95.01 (1.55)	0.05

Next, we investigate the effect of graph estimation error in identifying the set of true mediators $S^{(t)} = \{X_j^{(t)} : \eta_j^{(t)} \neq 0,\ j = 2, \ldots, p-1\}$. Figure 2 shows the averaged (over 200 iterations) Precision-Recall curves for estimating the target set $\cup_{t=1}^{r} S^{(t)}$ based on (i) the ranking of the absolute values of estimates of $\cup_{t=1}^{r}\{\eta_1^{(t)}, \ldots, \eta_p^{(t)}\}$ and (ii) the ranking of the corresponding p-values (in the reverse order). As we would expect, the methods based on the true DAG performs the best. Although it is unrealistic to assume that the true graph is known, we include it in our results to gain insight into the loss due to estimating the true CPDAG instead of the true DAG. We note that

TABLE 3
Empirical coverage probabilities (mean and median) and average length of the 95% confidence intervals for each of the groups, when the CPDAG is estimated.

p	n	$\lvert \max(\theta_{1j}^{(t)}, \text{aver}(\Theta_{jp}^{(t)}))\rvert$	Median Coverage	Mean Coverage	Length
252	500	Low	100 (0)	99.97 (0.24)	0.01
		Medium	99 (1.48)	96.16 (8.75)	0.04
		High	94 (2.97)	88.95 (15.16)	0.14
	1000	Low	100 (0)	99.96 (0.27)	0.01
		Medium	98 (2.97)	94.63 (11.62)	0.03
		High	94 (2.97)	87.32 (18.12)	0.10
	5000	Low	100 (0)	99.92 (0.47)	0.00
		Medium	96 (2.97)	91.22 (16.88)	0.01
		High	94 (2.97)	83.02 (24.94)	0.04
502	500	Low	100 (0)	99.97 (0.14)	0.01
		Medium	99.5 (0.74)	97.29 (6.68)	0.04
		High	94.5 (2.22)	90.56 (13.23)	0.14
	1000	Low	100 (0)	99.97 (0.18)	0.01
		Medium	98.5 (2.22)	96.19 (8.7)	0.03
		High	94.5 (2.22)	89.16 (16.2)	0.10
	5000	Low	100 (0)	99.93 (0.43)	0.00
		Medium	96.5 (2.97)	93.47 (12.41)	0.01
		High	94.5 (2.22)	86.17 (21.23)	0.04
1002	500	Low	100 (0)	99.98 (0.14)	0.01
		Medium	98.5 (2.22)	97.23 (5.09)	0.05
		High	94.5 (2.22)	92.89 (7.19)	0.15
	1000	Low	100 (0)	99.96 (0.22)	0.01
		Medium	97.5 (2.97)	96.12 (6.83)	0.04
		High	94.5 (2.22)	91.82 (9.88)	0.11
	5000	Low	100 (0)	99.86 (1.15)	0.00
		Medium	96 (2.22)	93.4 (11.34)	0.02
		High	94.5 (2.22)	88.79 (16.77)	0.05

the methods based on the estimated CPDAG and based on the true CPDAG perform equally well, and they outperform the naive method based on the empty graph. Finally, Figure 2 demonstrates that we can achieve substantial performance gain by using p-values instead of the raw estimates.

The harmonic mean of precision and recall is known as F-score and it is a popular way combining precision and recall into a single performance measure that ranges between 0 and 1. By adopting this notion of performance measure, we aim to choose a set of top mediators that maximizes F-score. However, since the p-value corresponding to the test $\text{aver}(\Theta_{jp})^{(t)} = 0$ has a stochastically larger distribution than Uniform$[0, 1]$ when both $\theta_{1j}^{(t)}$ and $\text{aver}(\Theta_{jp})^{(t)}$ are zero, identifying the true set of true mediators based on p-values using either false discovery rate (FDR) control or Bonferroni adjustment is not expected to perform well. Instead, we apply a heuristic

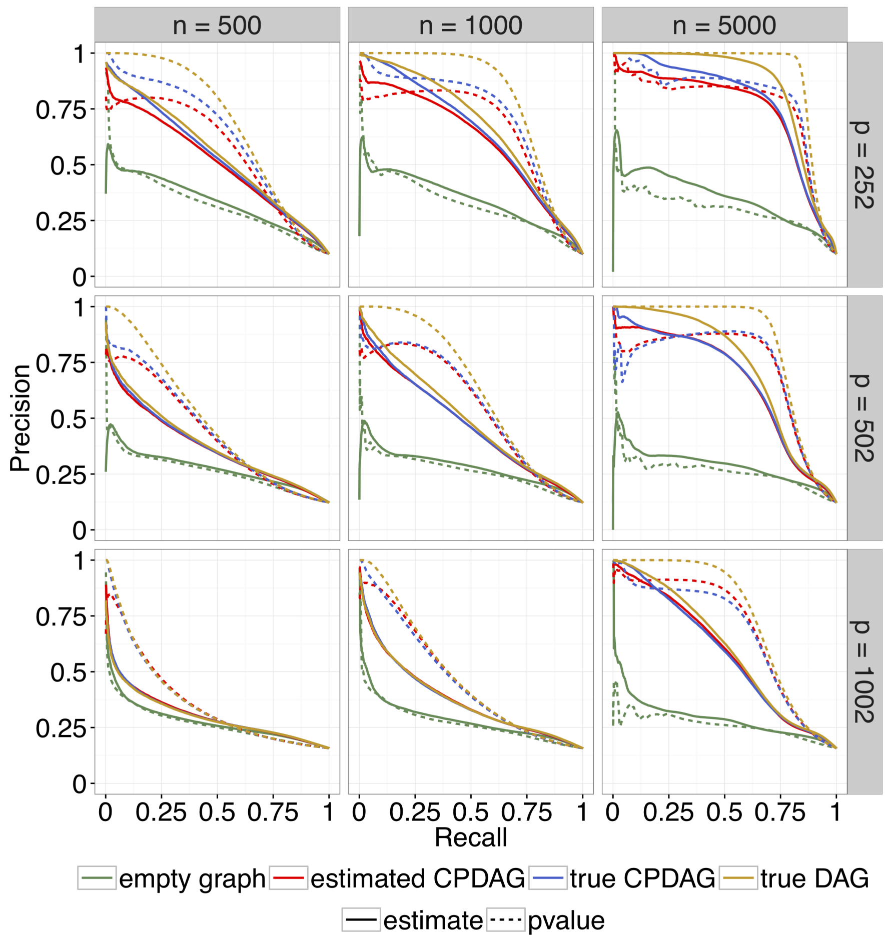

Fig 2: Precision-Recall curves for the target set with top k potential mediators according to (i) large $|\hat{\eta}_j^{(t)}|$ values (estimate) (solid line) and (ii) small corresponding p-values (dashed line), for $k = 1, \ldots, p \cdot r$. Note that $p \cdot r = 5000$ for each simulation setting.

screening where we select the potential mediators for which the total effect of the treatment X_1 on the mediator is non-zero, by testing $\theta_{1j}^{(t)} = 0$ at a given significance level. Although precision and recall cannot to be computed without knowing the true set of mediators, Table 4 demonstrates that we can achieve a nearly optimal F-score (i.e. the best achievable F-score of our method on a given dataset) by thresholding of p-values at level 0.01 for $n = 5000$ and at level 0.1 for $n \in \{500, 1000\}$. We acknowledge that this heuristic p-value thresholding technique does not possess any theoretical justification, but it seems to work surprisingly well for estimating the target set $\cup_{t=1}^{r} S^{(t)}$ in our simulation settings. We also provide discussions on some alternative approaches in Section 9 and in the Supplementary Material.

TABLE 4

Averaged (over 200 iterations) recall, precision and F-score for estimating the target set $\cup_{t=1}^{r} S^{(t)}$ based on p-value thresholding, where we used estimated CPDAGs for computing the value of our IDA-based estimators and the corresponding p-values. The numbers in the brackets are the corresponding standard deviations.

p	n	Target size	Estimated size	Recall	Precision	Achieved F-score	Optimal F-score
252	500	503	568 (18.1)	0.61 (0.02)	0.54 (0.02)	0.57 (0.01)	0.59 (0.01)
	1000		673.6 (19.3)	0.74 (0.02)	0.55 (0.01)	0.63 (0.01)	0.69 (0.01)
	5000		560.7 (11.4)	0.82 (0.01)	0.73 (0.01)	0.77 (0.01)	0.79 (0.01)
502	500	613	445.8 (17.2)	0.38 (0.02)	0.52 (0.02)	0.44 (0.02)	0.46 (0.01)
	1000		569.6 (17.8)	0.54 (0.02)	0.58 (0.02)	0.56 (0.01)	0.56 (0.01)
	5000		505.8 (10.5)	0.68 (0.01)	0.83 (0.01)	0.75 (0.01)	0.75 (0.01)
1002	500	788	390.7 (17.5)	0.24 (0.01)	0.48 (0.02)	0.32 (0.01)	0.38 (0.01)
	1000		502.6 (18.1)	0.35 (0.01)	0.55 (0.02)	0.43 (0.01)	0.45 (0.01)
	5000		455 (14.2)	0.5 (0.01)	0.87 (0.02)	0.63 (0.01)	0.67 (0.01)

8. Application. We demonstrate the performance of MIDA estimator in real data using a data set collected on 104 yeast segregants created by crossing of two genetically diverse strains, BY and RM [Brem and Kruglyak, 2005]. The data set includes the growth yields of each segregant grown in the presence of different chemicals or small molecule drugs [Perlstein et al., 2007]. These segregants have different genotypes that contribute to rich phenotypic diversity. One key question is to understand how genetic variants contribute to the phenotypic variability. One possible path is through regulation of gene expression variations. Besides genotype data, 6189 yeast genes are profiled in rich media and in the absence of any chemical or drug using expression arrays [Brem and Kruglyak, 2005]. We use the same data preprocessing steps as Chen et al. [2009] to create a list of candidate gene expression features based on their potential regulatory effects, including transcription factors, signalling molecules, chromatin factors and RNA factors and genes involved in vacuolar transport, endosome, endosome transport and vesicle-mediated transport. We further filter out genes with standard deviation (s.d.) ≤ 0.2 in expression level, resulting in a total of 813 genes in our analysis.

We are interested in identifying genes whose expression levels mediate the effect of genetic variants on yeast growth yield after being treated with hydrogen peroxide. In particular, the genetic variant M2_477206_486640 is highly associated with the yeast growth yield (p=0.00032). Our goal is to identify the gene expressions that mediate the effect of this genetic variant. At a nominal p-value of 0.05, MIDA identified six genes that may mediate the effects of the genetic variant M2_477206_486640 on yeast growth (see Table 5). Due to relatively small sample sizes, these genes are not significant after we adjust for multiple comparisons. However, the estimated signs of the mediation effects agree with known biology. Among these genes, overexpression of DRP8 leads to vegetative and decreased rate of growth. In contrast, lower expression of the GPA1 gene leads to decreased resistance to chemicals and decreased sporulation efficiency.

TABLE 5
Analysis of yeast growth yield in the presence of hydrogen peroxide. The gene expression mediators of genetic variant M2_477206_48664 identified by MIDA with a nominal p-value<0.05. Gene names, their estimated mediation effects and 95% confidence intervals are presented.

Gene ID	Gene name	Estimated effect	95% confidence interval	p-value
YNR047W	YNR047W	-0.0532 (0.0233)	(-0.0988, -0.0077)	0.022
YHR136C	SPL2	0.0257 (0.0116)	(0.0029, 0.0485)	0.027
YHR184W	SSP1	0.0468 (0.0229)	(0.0019, 0.0916)	0.041
YAL035W	FUN12	0.0757 (0.0375)	(0.0022, 0.1491)	0.043
YHR005C	GPA1	0.0766 (0.0381)	(0.0020, 0.1512)	0.044
YHR169W	DBP8	-0.1035 (0.0526)	(-0.2067, -0.0003)	0.049

9. Discussion.

In this paper, we have considered the problem of mediation analysis in the setting where we have high-dimensional and possibly interacting mediators. Directed acyclic graphs are used to characterize the possible interaction effects among the high dimensional mediators and to define the individual mediation effects based on linear structural equation models. We have developed an IDA-based procedure, MIDA, to estimate the individual mediation effects which takes into account the uncertainty of the estimated DAGs. We have also derived the asymptotic distributions of the estimates of both interventional effects as well as individual mediation effects, under the assumption of sub-Gaussian errors for the LSEMs, which facilitates inference based on these estimators, and to the best of our knowledge, are the first such results available in the literature. We have illustrated the methods in simulation studies with promising performance, as well as using a real data set on yeast in order to identify the possible gene expression mediators of a genetic variant for yeast growth in the presence of drugs. The methods can also be applied to the problem of identifying the important gene expression or methylation mediators that mediate the effects of genetic variants identified through genome-wide association studies on disease phenotypes.

Another crucial contribution of this work lies in the results of Section 5 on uniform ALEs and non-asymptotic control of error terms for linear regression estimators based on varying subsets of high dimensional covariates that serve as the backbone of our results on the asymptotic distribution of the MIDA estimators. These results are applicable quite generally in several other problems and should therefore be of independent interest. It is worth mentioning that some results on 'uniform-in-model' bounds, similar in flavor to those derived in Section 5, were also obtained independently in the recent work of Kuchibhotla et al. [2018], although their results are not directly comparable to ours and their analysis, in general, is far more involved. However, their results are targeted towards more general settings and are quite elaborate, albeit less tractable. Our approach on the other hand is simpler and the results are more explicit and ready-to-use for application purposes. Further, our results automatically account for data dependent centering of Y and $\mathbf{X}$, unlike theirs, and the bounds in our main result (Theorem 5.1) are also more flexible in the sense that they directly involve the cardinality of $\mathcal{S}$, as opposed to the bounds of Kuchibhotla et al. [2018] which generally aim at a worst case analysis with $\mathcal{S}$ assumed to include all subsets of $\mathbf{X}$ having cardinality bounded by some $k \leq p$. Our bounds are therefore adaptive in $|\mathcal{S}|$ and lead to sharper rates when $|\mathcal{S}|$ is not too large (or at least not growing as fast as the worst case) which can be quite useful in practice.

Our simulations have shown that the confidence intervals based on MIDA provide correct coverage when the mediation effects are not too small. In high dimensional settings with thousands of possible mediators, large sample sizes are needed in order to accurately estimate the individual mediation effects. In order to estimate the set of true mediators $\{X_j : \eta_j \neq 0\}$ as an alternative to our heuristic p-value based thresholding approach proposed in Section 7.2, one can also apply the Benjamini-Hochberg (BH) false discovery rate control procedure at some desired level α (e.g. $\alpha = 0.1$). However, the theoretical guarantee of the BH procedure does not apply in our case due to the fact that the true DAG is not identifiable. Instead, the BH procedure (asymptotically) guarantees to control FDR for estimating $\{X_j : \theta_{1j}\text{aver}(\Theta_{jp}) \neq 0\}$, especially when the true CPDAG is known. Furthermore, Table 4 suggests that it might be unreasonable and too optimistic to enforce a high precision level such as 0.9 (equivalently, FDR control at level $\alpha = 0.1$) in the challenging problem of estimating the set of true mediators in high-dimensional settings. For these reasons, we recommend the estimation of the target set by maximizing the F-score because of its adaptive capability of automatically adjusting to the best achievable precision level (or least FDR level) for the problem at hand.

We also empirically observed that the BH procedure can become very conservative for estimating $\{X_j : \theta_{1j}\text{aver}(\Theta_{jp}) \neq 0\}$, mainly because the p-value corresponding to the test $\text{aver}(\Theta_{jp}) = 0$ has a non-uniform and left-skewed distribution when both θ_{1j} and $\text{aver}(\Theta_{jp})$ are zero (see also the last paragraph of Section 6). One way to mitigate this issue is to apply a heuristic screening procedure. We explored this approach in our simulation studies to obtain a potential set of mediators for which the total effect of the treatment on the mediator is non-zero and to apply the BH procedure on the selected set. We provide further discussions on this approach in the Supplementary Material and the simulation results presented therein show that this strategy can significantly improve the FDR controlling compared to a direct application of the BH procedure based on the p-values.

We conclude by noting that although the theoretical results for MIDA have been obtained under the absence of any confounding variable (a.k.a. common cause) W between the treatment variable X_1 and $X_j \in \mathbf{X} \setminus \{X_1\}$, it is straightforward to extend our results in the presence of a finite set of known confounding variables $\mathbf{W} = \{W_1, \ldots, W_k\}$ by adding the set of regressors $\mathbf{W}$ to each linear regression involving X_1.

APPENDIX A: SUB-GAUSSIANS AND SUB-EXPONENTIALS

In this section, we formally define sub-Gaussian and sub-exponential random variables (and vectors) based on the general concept of (exponential) Orlicz norms. These norms are used in some of our main assumptions.

DEFINITION A.1 (The ψ_α-Orlicz norm, and sub-Gaussian and sub-exponential variables). *For any $\alpha > 0$, define the function $\psi_\alpha(u) := \exp(u^\alpha) - 1$ $\forall\, u \geq 0$. For any random variable X and any $\alpha > 0$, the ψ_α-*Orlicz norm *(the exponential Orlicz norm of order $\alpha > 0$) of X is then defined as:*

$$\|X\|_{\psi_\alpha} := \inf\{c > 0 : \mathbb{E}\{\psi_\alpha(|X|/c)\} \leq 1\},$$

where $\|X\|_{\psi_\alpha}$ is understood be ∞ if the infimum above is over an empty set. The special cases of $\alpha = 2$ and $\alpha = 1$ correspond to the well known sub-Gaussian and sub-exponential random variables. X is said to be sub-Gaussian if $\|X\|_{\psi_2} < \infty$ (and $\|X\|_{\psi_2}$ is also referred to as the 'sub-Gaussian norm' of X). X is said to be sub-exponential if $\|X\|_{\psi_1} < \infty$.

DEFINITION A.2 (Sub-Gaussian norm(s) for random vectors). *A random vector $\mathbf{X} \in \mathbb{R}^d$ $(d \geq 1)$ is defined to be sub-Gaussian if and only if $\forall\, \mathbf{v} \in \mathbb{R}^d$,*

$\mathbf{v}^T\mathbf{X}$ *is sub-Gaussian, as in Definition A.1. For such random vectors, we define* two *sub-Gaussian norms as follows:*

$$\|\mathbf{X}\|_{\psi_2} := \max_{1\leq j\leq d} \left\|\mathbf{X}_{[j]}\right\|_{\psi_2} \quad \textit{and} \quad \|\mathbf{X}\|^*_{\psi_2} := \sup_{\|\mathbf{v}\|_2\leq 1} \left\|\mathbf{v}^T\mathbf{X}\right\|_{\psi_2}.$$

For a general $\alpha > 0$*, we also define, analogous to* $\|\mathbf{X}\|_{\psi_2}$*, the* ψ_α-Orlicz norm of a random vector $\mathbf{X} \in \mathbb{R}^d$ *as:* $\|\mathbf{X}\|_{\psi_\alpha} := \max_{1\leq j\leq d} \left\|\mathbf{X}_{[j]}\right\|_{\psi_\alpha}$.

For most of our analyses, we use $\|\cdot\|^*_{\psi_2}$ as the vector sub-Gaussian norm which has usually been the accepted definition [Vershynin, 2012, 2018, e.g.]. The corresponding extension of the $\|\mathbf{X}\|^*_{\psi_2}$ norm to a $\|\cdot\|^*_{\psi_\alpha}$ norm for a general $\alpha > 0$ is however not immediate, and certainly not standard in the literature.

ACKNOWLEDGEMENTS

The first author thanks Arun K. Kuchibhotla for helpful discussions.

SUPPLEMENTARY MATERIAL

Supplement to "Inference for Individual Mediation Effects and Interventional Effects in Sparse High- Dimensional Causal Graphical Models" (.pdf file). The supplement includes: (i) proofs of all theoretical results in the main paper, (ii) additional technical tools and supporting lemmas, and (iii) additional numerical results regarding FDR controlling.

REFERENCES

Baron, R. M. and Kenny, D. A. (1986). The moderatormediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* **51** 1173–1182.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188.

Boca, S. M., Sinha, R., Cross, A. J., Moore, S. C. and Sampson, J. N. (2014). Testing multiple biological mediators simultaneously. *Bioinformatics* **30** 214–220.

Brem, R. and Kruglyak, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc Natl Acad Sci USA* **102** 1572–1577.

Chen, B., Causton, H., Mancenido, D., Goddard, N., Perlstein, E. and Pe'er, D. (2009). Harnessing gene expression to identify the genetic basis of drug resistance. *Mol Syst Biol.* **5** 310.

Chickering, D. M. (2002a). Optimal structure identification with greedy search. *J. Mach. Learn. Res.* **3** 507–554.

Chickering, D. M. (2002b). Learning equivalence classes of Bayesian-network structures. *J. Mach. Learn. Res.* **2** 445–498.

Colombo, D. and Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.* **15** 3741-3782.

Harris, N. and Drton, M. (2013). PC algorithm for nonparanormal graphical models. *J. Mach. Learn. Res.* **14** 3365-3383.

Horn, R. A. and Johnson, C. R. (1990). *Matrix Analysis*. Cambridge University Press.

Huang, Y. T. and Pan, W. C. (2016). Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics* **72** 402–413.

James, L., Mulaik, S. and Brett, J. (1982). *Causal Analysis: Assumptions, Models, and Data*. Sage Publications.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research* **15** 2869–2909.

Judd, C. M. and Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Eval. Rev* **5** 602–619.

Kalisch, M. and Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.* **8** 613–636.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H. and Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *J. Statist. Software* **47** 1-26.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT Press.

Kuchibhotla, A. K., Brown, L. D., Buja, A., George, E. I. and Zhao, L. (2018). A Model Free Perspective for Linear Regression: Uniform-in-model Bounds for Post Selection Inference. *ArXiv preprint arXiv:1802.05801*.

Le, T. D., Zhang, J., Liu, L., Truong, B. M. T., Hu, S., Xu, T. and Li, J. (2017). Identifying microRNA Targets in Epithelial-mesenchymal Transition Using Joint-intervention Causal Inference. In *CSBio 2017* 34–41.

Maathuis, M. H., Kalisch, M. and Bühlmann, P. (2009). Estimating high-dimensional intervention effects from observational data. *Ann. Statist.* **37** 3133-3164.

Maathuis, M. H., Colombo, D., Kalisch, M. and Bühlmann, P. (2010). Predicting causal effects in large-scale systems from observational data. *Nature Methods* **7** 247–248.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G. and Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychol Methods.* **7** 83–104.

Nandy, P., Hauser, A. and Maathuis, M. H. (2018). High-dimensional consistency in score-based and hybrid structure learning. *Ann. Statist.* **46** 3151-3183.

Nandy, P., Maathuis, M. H. and Richardson, T. S. (2017). Estimating the effect of joint interventions from observational data in sparse high-dimensional settings. *Ann. Statist.* **45** 647-674.

Pearl, J. (2000). *Causality. Models, Reasoning, and Inference*. Cambridge University Press, Cambridge.

Pearl, J. (2009). Causal inference in statistics: An overview. *Stat. Surv.s* **3** 96-146.

Perlstein, E., Ruderfer, D., Roberts, D., Schreiber, S. and Kruglyak, L. (2007). Genetic basis of individual differences in the response to small?molecule drugs in yeast. *Nat Genet* **39** 496–502.

Pollard, D. (2015). A Few Good Inequalities Book Chapter, Department of Statistics, Yale University. (Available at www.stat.yale.edu/~pollard/Books/Mini/Basic.pdf).

Portnoy, S. (1984). Asymptotic behavior of M-Estimators of p Regression Parameters when p^2/n is Large. I. Consistency. *The Annals of Statistics* **12** 1298–1309.

Portnoy, S. (1985). Asymptotic Behavior of M Estimators of p Regression Parameters when p^2/n is Large. II. Normal Approximation. *The Annals of Statistics* **13** 1403–1417.

PORTNOY, S. (1986). On the Central Limit Theorem in $\mathbb{R}^p$ when $p \to \infty$. *Probability Thoery and Related Fields* **73** 571–583.

PORTNOY, S. (1988). Asymptotic Behavior of Likelihood Methods for Exponential Families when the Number of parameters Tends to Infinity. *The Annals of Statistics* **16** 356–366.

PREACHER, K. J. and HAYES, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behav Res Methods* **40** 879–891.

RIGOLLET, P. and HÜTTER, J.-C. (2017). Sub-Gaussian Random Variables. In *High Dimensional Statistics* 1 Massachusetts Institute of Technology OpenCourseWare Lecture notes. (Available at http://www-math.mit.edu/~rigollet/PDFs/RigNotes17.pdf).

SHEVTSOVA, I. (2011). On the Absolute Constants in the Berry-Esseen Type Inequalities for Identically Distributed Summands. *arXiv preprint arXiv:1111.6554v1*.

SOBEL, M. E. (1982). Asymptotic Confidence Intervals for Indirect Effects in Structural EQUATION MODELS. *Sociol. Methodol.* **13** 290–312.

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation, Prediction, and Search*, second ed. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge.

SPIRTES, P., RICHARDSON, T., MEEK, C., SCHEINES, R. and GLYMOUR, C. (1998). Using Path Diagrams as a Structural Equation Modeling Tool. *Socio. Meth. Res.* **27** 182-225.

STEKHOVEN, D. J., MORAES, I., SVEINBJÖRNSSON, G., HENNING, L., MAATHUIS, M. H. and BÜHLMANN, P. (2012). Causal stability ranking. *Bioinform.* **28** 2819-2823.

THOMPSON, R. C. (1972). Principal Submatrices IX: Interlacing Inequalities for Singular Values of Submatrices. *Linear Algebra and Its Applications* **5** 1–12.

TSAMARDINOS, I., BROWN, L. E. and ALIFERIS, C. F. (2006). The Max-min Hill-climbing Bayesian Network Structure Learning Algorithm. *Mach. Learn.* **65** 31-78.

VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models. *The Annals of Statistics* **42** 1166–1202.

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag, New York.

VANDERWEELE, T. J. and VANSTEELANDT, S. (2014). Mediation Analysis with Multiple Mediators. *Epidemiol Method.* **2** 95–115.

VERMA, T. and PEARL, J. (1990). Equivalence and synthesis of causal models. In *UAI 1990*.

VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing: Theory and Applications* 210-268. Cambridge University Press (ArXiv preprint available at arXiv:1011.3027v7).

VERSHYNIN, R. (2018). *High Dimensional Probability. An Introduction with Applications in Data Science*. Cambridge University Press (To appear). (Available at www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-book.pdf).

WAINWRIGHT, M. (2017). Basic Tail and Concentration Bounds. In *High-dimensional Statistics: A Non-asymptotic Viewpoint* (To appear) (Available at www.stat.berkeley.edu/~mjwain/stat210b/Chap2_TailBounds_Jan22_2015.pdf).

WRIGHT, S. (1921). Correlation and causation. *J. Agric. Res.* **20** 557–585.

YULE, G. U. (1907). On the Theory of Correlation for Any Number of Variables, Treated by a New System of Notation. *Proc. Roy. Soc. London, Ser. A* **79** 182–193.

ZHANG, F. (2005). *The Schur Complement and Its Applications*. *Numerical Methods and Algorithms* **4**. Springer, New York.

ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76** 217–242.

Zhang, H., Zheng, Y., Zhang, Z., Gao, T., Joyce, B., Yoon, G., Zhang, W., Schwartz, J., Just, A., Colicino, E., Vokonas, P., Zhao, L., Lv, J., Baccarelli, A., Hou, L. and Liu, L. (2016). Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics* **32** 3150–3154.

APPENDIX B: PROOFS OF ALL RESULTS

B.1. Proof of Proposition 2.1. By Theorem 3.1 of Nandy, Maathuis and Richardson [2017], we have,

$$\theta_{1p}^{(1,j)} = \theta_{1p} - \theta_{1j}\theta_{jp}.$$

Hence, $\eta_j = \theta_{1p} - \theta_{1p}^{(1,j)} = \theta_{1j}\theta_{jp}$. □

B.2. Proof of Lemma 3.1. The condition $(B_{\mathcal{G}_0})_{j1} = 0$ implies that $X_1 = \mu_1 + \epsilon_1$. Therefore, it follows from the independence of the error variables that

$$\mathbb{E}[\boldsymbol{\epsilon} \mid X_1] = \mathbb{E}[\boldsymbol{\epsilon} \mid \epsilon_1] = (\epsilon_1, 0, \ldots, 0)^T = \mathbf{e}_{1,p}^T\ (X_1 - \mu_1),$$

where $\mathbf{e}_{1,p}^T$ denotes the first column of a $p \times p$ identity matrix. Let A denote the $(p-2) \times p$ matrix such that $\mathbf{X'} = A\mathbf{X}$. Then, we have,

$$\mathbb{E}[\mathbf{X'} \mid X_1] = A\,\mathbb{E}[\mathbf{X} \mid X_1] = A(\mathrm{I} - B_{\mathcal{G}_0}^T)^{-1}\,\mathbb{E}[\boldsymbol{\epsilon} \mid X_1] = A(\mathrm{I} - B_{\mathcal{G}_0}^T)^{-1}\mathbf{e}_{1,p}^T\ X_1.$$

This completes the proof of the linearity property of the conditional expectation $\mathbb{E}[\mathbf{X'} \mid X_1]$.

Since $(B_{\mathcal{G}_0})pj = 0$ and $\mathbb{E}[\epsilon_1 \mid X_1] = 0$ for $j \in \{2, \ldots, p-1\}$, we have

$$\mathbb{E}[X_j \mid X_1] = (B_{\mathcal{G}_0})_{1j} X_1 + \sum_{k=2}^{p-1} (B_{\mathcal{G}_0})_{kj}\,\mathbb{E}[X_k \mid X_1].$$

This implies $\mathbf{X'} - \mathbb{E}[\mathbf{X'}] = B_{\mathcal{G}_0'}^T(\mathbf{X'} - \mathbb{E}[\mathbf{X'}]) + \boldsymbol{\epsilon}'$.

Finally, we show below that the faithfulness of the distribution of $\mathbf{X}^\dagger = \mathbf{X'} - \mathbb{E}[\mathbf{X'} \mid X_1]$ to $\mathcal{G}_0'$ follows from the faithfulness of the distribution of $\mathbf{X}$ to $\mathcal{G}_0$. Let $X_i^\dagger := X_i - \mathbb{E}[X_i \mid X_1]$ for $i = 2, \ldots, p-1$. Suppose $X_i^\dagger$ and $X_k^\dagger$ are conditionally independent given $\mathbf{X}_S^\dagger$ for some set $S \subseteq \{2, \ldots, p-1\} \setminus \{i, k\}$. In order to establish faithfulness of $\mathbf{X}^\dagger$ to $\mathcal{G}_0'$, we need to show that X_i and X_j are d-separated by $\mathbf{X}_S$ in $\mathcal{G}_0'$, that is $\mathbf{X}_S$ *blocks* every *path* between X_i

and X_k in $\mathcal{G}_0'$. A path in a graph is a sequence of distinct nodes such that all pairs of successive nodes in the sequence are adjacent in the graph, and S blocks a path in $\mathcal{G}_0'$ if the path contains a non-collider that is in $\mathbf{X}_S$, or the path contains a collider that has no descendant in $\mathbf{X}_S$, where (X_r, X_s, X_t) a collider in a graph $\mathcal{G}$ if $\{X_r, X_t\} \subseteq \mathbf{Pa}_{\mathcal{G}}(X_s)$.

Since $(B_{\mathcal{G}_0})_{j1} = 0$ for all j, X_1 cannot be collider on any path in $\mathcal{G}_0$. Further, since $(B_{\mathcal{G}_0})_{pj} = 0$ for all j, X_p cannot be non-collider on any path in $\mathcal{G}_0$. These imply X_i and X_k are d-separated by $\mathbf{X}_S$ in $\mathcal{G}_0'$ if and only if X_i and X_k are d-separated by $\mathbf{X}_S \cup \{X_1\}$ in $\mathcal{G}_0$, since all paths between X_i and X_k in $\mathcal{G}_0$ that are not present in $\mathcal{G}_0'$ must go through X_1 or X_p. Therefore, it is sufficient to show that X_i and X_k are d-separated by $\mathbf{X}_S \cup \{X_1\}$ in $\mathcal{G}_0$. This is equivalent to show that the partial correlation between X_i and X_k given $\mathbf{X}_S \cup \{X_1\}$, denoted by $\rho_{ik|S\cup\{1\}}$, is zero, as the distribution of $\mathbf{X}$ is generated from a LSEM and faithful to $\mathcal{G}_0$ (see Spirtes et al. [1998]; Nandy, Hauser and Maathuis [2018]).

Note that $X_i^\dagger$ and $X_k^\dagger$ are conditionally independent given $\mathbf{X}_S^\dagger$ implies that the partial correlation between $X_i^\dagger$ and $X_k^\dagger$ given $\mathbf{X}_S^\dagger$, denoted by $\rho_{ik|S}^\dagger$, is zero. Thus we complete the proof by showing that

$$1 - \rho_{ik|S}^{\dagger 2} = 1 - \rho_{ik|S\cup\{1\}}^2. \tag{B.1}$$

To this end, we note that the linearity of conditional expectation $\mathbb{E}[\mathbf{X}' \mid X_1]$ implies $\Sigma^\dagger = \Sigma_{(2,\ldots,p-1)(2,\ldots,p-1)} - \Sigma_{(2,\ldots,p-1)1}\Sigma_{11}^{-1}\Sigma_{1(2,\ldots,p-1)}$. Now recall that if $\sigma_{i|S}^\dagger := \Sigma_{ii}^\dagger - \Sigma_{iS}^\dagger \Sigma_{SS}^\dagger \Sigma_{Si}^\dagger$ is the variance of the residuals in the linear regression (based on $\Sigma^\dagger = \mathrm{Cov}(\mathbf{X}^\dagger)$) of $X_i^\dagger$ on $\mathbf{X}_S^\dagger$ and $\sigma_{i|S\cup\{k\}}^{\dagger 2}$ is the variance of the residuals in the linear regression of $X_i^\dagger$ on $\mathbf{X}_{S\cup\{k\}}^\dagger$, then $\sigma_{i|S\cup\{k\}}^{\dagger 2} = (1-\rho_{ik|S}^{\dagger 2})\sigma_{i|S}^{\dagger 2}$ [Yule, 1907]. Therefore, by applying the identity for expressing the Schur complement of a $(r-1)\times(r-1)$ principal submatrix of a $r \times r$ matrix as the ratio of determinants, we obtain

$$\begin{aligned} \sigma_{i|S}^{\dagger 2} &= \frac{|\Sigma_{(i,S)(i,S)}^\dagger|}{|\Sigma_{SS}^\dagger|} = \frac{|\Sigma_{(i,S)(i,S)} - \Sigma_{(i,S)1}\Sigma_{11}^{-1}\Sigma_{1(i,S)}|}{|\Sigma_{SS} - \Sigma_{S1}\Sigma_{11}^{-1}\Sigma_{1S}|} \\ &= \frac{|\Sigma_{(i,S,1)(i,S,1)}| \cdot |\Sigma_{11}|}{|\Sigma_{11}| \cdot |\Sigma_{(S,1)(S,1)}|} = \sigma_{i|S\cup\{1\}}^2, \end{aligned}$$

where $\sigma_{i|S\cup\{1\}}^2$ is the variance of the residuals in the linear regression (based on $\Sigma = \mathrm{Cov}(\mathbf{X})$) of X_i on $\{X_r : r \in S \cup \{1\}\}$. Similarly, we have $\sigma_{i|S\cup\{k\}}^{\dagger 2} =$

$\sigma_{i|S\cup\{k,1\}}$. Hence,

$$1-\rho^{\dagger 2}_{ik|S}=\frac{\sigma^{\dagger 2}_{i|S\cup\{k\}}}{\sigma^{\dagger 2}_{i|S}}=\frac{\sigma^{2}_{i|S\cup\{k\}}}{\sigma^{2}_{i|S}}=1-\rho^{2}_{ik|S\cup\{1\}}.$$

B.3. Proof of Theorem 4.1.

LEMMA B.1. *Let* $\mathbf{X}$ *be generated from a LSEM characterized by* $(B_{\mathcal{G}_0},\boldsymbol{\epsilon})$. *Then Assumptions 4.1, 4.3 and 4.6 imply that for any* $S\subseteq\{1,\dots,p\}$ *such that* $|S|\le\tilde{q}+3$,

$$||\mathbf{X}_S-\boldsymbol{\mu}_S||^*_{\psi_2}\le C_4,$$

where $C_4>0$ *is an absolute constant depending on* C_1, C_2 *and* C_3 *given by Assumptions 4.1 and 4.6, and* $||\cdot||^*_{\psi_2}$ *denotes the vector sub-Gaussian norm given by Definition A.2.*

PROOF. Fix $S\subseteq\{1,\dots,p\}$ such that $|S|\le\tilde{q}+3$. Let A be the $|S|\times p$ matrix such that $\mathbf{X}_S-\boldsymbol{\mu}_S=A(\mathbf{X}-\boldsymbol{\mu})$. Therefore, we have, $\mathbf{X}_S-\boldsymbol{\mu}_S=A(I-B^T_{G_0})^{-1}\boldsymbol{\epsilon}$. Hence,

$$\begin{aligned}||\mathbf{X}_S-\boldsymbol{\mu}_S||^*_{\psi_2} &= \sup_{||\mathbf{v}||_2=1}||\mathbf{v}^TA(I-B^T_{G_0})^{-1}\boldsymbol{\epsilon}||_{\psi_2}\\ &= \sup_{||\mathbf{v}||_2=1}||\mathbf{v}^TA(I-B^T_{G_0})^{-1}||_2\left|\left|\frac{\mathbf{v}^TA(I-B^T_{G_0})^{-1}\boldsymbol{\epsilon}}{||\mathbf{v}^TA(I-B^T_{G_0})^{-1}||_2}\right|\right|_{\psi_2}\\ &\le ||A(I-B^T_{G_0})^{-1}||_2\;||\boldsymbol{\epsilon}||^*_{\psi_2},\end{aligned}$$

where the last inequality follows from the definitions of spectral norm and $||\cdot||^*_{\psi_2}$ norm and the fact that $\sup_{\mathbf{x}} f(\mathbf{x})\,g(\mathbf{x})\le\sup_{\mathbf{x}} f(\mathbf{x})\sup_{\mathbf{x}} g(\mathbf{x})$.

Since $\epsilon_1,\epsilon_2,\dots,\epsilon_p$ are independent zero-mean sub-Gaussian random variables satisfying $\max_{1\le i\le p}||\epsilon_i||_{\psi_2}\le C_1$, it follows from Lemma 5.24 of Vershynin [2012] that $||\boldsymbol{\epsilon}||^*_{\psi_2}\le C_0C_1$, for some absolute constant C_0. Furthermore, since $(\Sigma_0)_S=A(I-B^T_{G_0})^{-1}D(I-B^T_{G_0})^{-T}A^T$ for $D:=\text{Cov}(\boldsymbol{\epsilon})$, it follows from the submultiplicity property of the spectral norm that

$$||A(I-B^T_{G_0})^{-1}||_2\le||(\Sigma_0)^{1/2}_S||_2\;||D^{-1/2}||_2\le\sqrt{C_3}\;||D^{-1/2}||_2,$$

where the last inequality follows from Assumption 4.6.

Thus it remains to show that $||D^{-1/2}||_2$ is bounded. To this end, note that

$$||D^{-1/2}||_2=\frac{1}{\sqrt{\min_{1\le i\le p}\text{Var}(\epsilon_i)}}.$$

Finally, from the interlacing property of eigenvalues of a Hermitian matrix A and the eigenvalues of the Schur complement of any principal sub-matrix of A (see, e.g., Corollary 2.3 of Zhang [2005]), it follows that

$$\begin{aligned}\mathrm{Var}(\epsilon_i) = &\ (\Sigma_0)_i - (\Sigma_0)_{i\mathbf{Pa}_{\mathcal{G}_0}(i)}(\Sigma_0)^{-1}_{\mathbf{Pa}_{\mathcal{G}_0}(i)\mathbf{Pa}_{\mathcal{G}_0}(i)}(\Sigma_0)_{\mathbf{Pa}_{\mathcal{G}_0(i)}i} \\ \geq &\ \lambda_{\min}\left((\Sigma_0)_{(i,\mathbf{Pa}_{\mathcal{G}_0}(i))(i,\mathbf{Pa}_{\mathcal{G}_0}(i))}\right) \\ \geq &\ C_2\end{aligned}$$

where $\mathbf{Pa}_{\mathcal{G}_0}(i) = \{r : X_r \in \mathbf{Pa}_{\mathcal{G}_0}(X_i)$. The last inequality follows from Assumption 4.6 since from Assumption 4.3 and the definition of $\tilde{q}$ we have $|\mathbf{Pa}_{\mathcal{G}_0}(i)| \leq q \leq \tilde{q}$. □

Proof of the First Part of Theorem 4.1. We recall that if $\mathbf{X}$ is generated from a linear SEM such that the distribution of $\mathbf{X}$ is faithful to a DAG $\mathcal{G}$, then for $i, k \in \{1, \ldots, p\}$ and $S \subseteq \{1, \ldots, p\} \setminus \{i, k\}$ X_i and X_k are d-separated by $X_S := \{X_k : k \in S\}$ in $\mathcal{G}$ if and only if the partial correlation between X_i and X_k given $X_\mathbf{S}$ is zero (see, for example, Section 6 of Nandy, Hauser and Maathuis [2018]). Therefore, the soundness of a δ-optimal oracle version of GES (Definition 5.3 of Nandy, Hauser and Maathuis [2018]) with $\mathrm{Cov}(\mathbf{X}' - \mathbb{E}[\mathbf{X}' \mid X_1]))$ follows from the proof of Theorem 5.1 of Nandy, Hauser and Maathuis [2018] and Lemma 3.1.

Next, we note that the following result follows directly from Lemma 5.1 of Nandy, Hauser and Maathuis [2018]. Let $\mathcal{H} = (\mathbf{X}', E)$ be a DAG such that $X_i' \in \mathbf{Nd}_\mathcal{H}(X_k') \setminus \mathbf{Pa}_\mathcal{H}(X_k')$. Let $\mathcal{H}' = (\mathbf{X}', E \cup \{X_i' \to X_k'\})$. Then

$$\mathcal{S}_\lambda(\mathcal{H}', \mathcal{D}_n) - \mathcal{S}_\lambda(\mathcal{H}, \mathcal{D}_n) = \frac{1}{2}\log\left(1 - \left(\hat{\rho}^\dagger_{ik|\mathbf{Pa}_\mathcal{H}(k)}\right)^2\right) + \lambda, \tag{B.2}$$

where $\mathcal{D}_n$ denote the data, $\hat{\rho}^\dagger_{ik|S}$ denotes the sample partial correlation between $\mathbf{r}_i$ and $\mathbf{r}_k$ given $\{\mathbf{r}_j : j \in S\}$ (see Algorithm 3.1). Then it follows from the proof of Lemma 3.1 that for any $i, k \in \{2, \ldots, p-1\}$ with $i \neq k$ and $S \subseteq \{2, \ldots, p-1\} \setminus \{i, k\}$, we have

$$1 - \hat{\rho}^{\dagger 2}_{ik|S} = 1 - \hat{\rho}^2_{ik|S\cup\{1\}}, \tag{B.3}$$

where $\hat{\rho}_{ik|S\cup\{1\}}$ denote the sample partial correlation between X_i and X_k given $\mathbf{X}_S \cup X_1$.

Therefore, (B.3) shows that the right hand side of the score difference formula given in (B.2) can be expressed as a function of sample partial correlation. Furthermore, Lemma B.1 enables us to use Corollary 5.50 of Vershynin [2012], as in the proof of Theorem 6.1 of Nandy, Hauser and

Maathuis [2018]. Hence, $\mathbb{P}(\hat{\mathcal{C}}'(\lambda_n) \neq \mathcal{C}_0') \longrightarrow 0$ can be obtained by applying the arguments given in the proof of Theorem 6.1 of Nandy, Hauser and Maathuis [2018]. □

Proof of the Second Part of Theorem 4.1. Let $A_n = \{\hat{\mathcal{C}}'(\lambda_n) = \mathcal{C}_0\}$. Since $\mathbb{P}(A_n) \longrightarrow 0$ (by the first part of Theorem 4.1) and $\text{aver}(\hat{\Theta}_{jp}(\hat{\mathcal{C}}'))$ $= \text{aver}(\hat{\Theta}_{jp}(\mathrm{C}_0'))$ on the set A_n, it is sufficient to show that for any $\delta > 0$,

$$\mathbb{P}\left(\max_{1<j<p} |\text{aver}(\hat{\Theta}_{jp}(\mathcal{C}_0')) - \text{aver}(\Theta_{jp})| > \delta,\ A_n\right) \longrightarrow 0.$$

For $j \in \{2, \ldots, p_n - 1\}$, we denote the distinct elements in the multi-set $\{\mathbf{Pa}_{\mathcal{G}}(X_j) \cup \{X_1\} : \mathcal{G} \in \text{MEC}(\mathcal{C}_0')\}$ by $\{\mathbf{X}_{S_{j1}}, \ldots, \mathbf{X}_{S_{jm_j}}\}$. By Assumption 4.3, we have $m_j \leq 2^q$ for all $j \in \{2, \ldots, p-1\}$. Therefore,

$$\begin{aligned}
&\mathbb{P}\left(\max_{1<j<p} |\text{aver}(\hat{\Theta}_{jp}(\mathcal{C}_0')) - \text{aver}(\Theta_{jp})| > \delta,\ A_n\right)\\
\leq\ &\mathbb{P}\left(\max_{1<j<p}\ \max_{1\leq r\leq m_j} |\hat{\beta}_{jp|\mathbf{X}_{S_{jr}}} - \beta_{jp|\mathbf{X}_{S_{jr}}}|\ > \delta,\ A_n\right)\\
\leq\ &(p-2)2^q \max_{1<j<p}\ \max_{1\leq r\leq m_j} \mathbb{P}\left(|\hat{\beta}_{jp|\mathbf{X}_{S_{jr}}} - \beta_{jp|\mathbf{X}_{S_{jr}}}|\ > \delta\right).
\end{aligned}$$

We complete the proof by showing that for all $j \in \{2, \ldots, p-1\}$ and $r \in \{1, \ldots, m_j\}$, $\mathbb{P}(|\hat{\beta}_{jp|\mathbf{X}_{S_{jr}}} - \beta_{jp|\mathbf{X}_{S_{jr}}}| > \delta) \leq 2\exp(-C_7 n\delta^2)$ for some absolute constant $C_7 > 0$. Note that this implies

$$\begin{aligned}
&(p-2)2^q \max_{1<j<p}\ \max_{1\leq r\leq m_{nj}} \mathbb{P}(|\hat{\beta}_{jp|\mathbf{X}_{S_{jr}}} - \beta_{jp|\mathbf{X}_{S_{jr}}}| > \delta)\\
&\leq\ O(\exp((\log(2) + \log(p))q - C_7 n\delta^2)) \longrightarrow 0,
\end{aligned}$$

since from Assumptions 4.2 and 4.3, we have $p = O(n^a)$ and $q = O(n^{1-b_1})$ for some $0 \leq a < \infty$ and $0 < b_1 \leq 1$.

Fix $j \in \{2, \ldots, p-1\}$ and $r \in \{1, \ldots, m_j\}$. Let Σ and Σ' denote the submatrices of $\Sigma_0 = \text{Cov}(\mathbf{X})$ that corresponds to $(X_j, \mathbf{X}_{S_{jr}}, X_p)$ and $(X_{nj}, \mathbf{X}_{S_{jr}})$ respectively. Then $\beta_{jp|\mathbf{X}_{S_{jr}}} = \mathbf{e}_1^T \Sigma'^{-1} \boldsymbol{\sigma}_p$, where $\mathbf{e}_1$ denote the first column of an identity matrix of appropriate order and $\boldsymbol{\sigma}_p$ denote the last column of Σ. Similarly, we define the corresponding sample covariance matrices $\hat{\Sigma}$ and $\hat{\Sigma}'$ to obtain $\hat{\beta}_{jp|\mathbf{X}_{S_{jr}}} = \mathbf{e}_1^T \hat{\Sigma}'^{-1} \hat{\boldsymbol{\sigma}}_p$, where $\hat{\boldsymbol{\sigma}}_p$ denote the last column of $\hat{\Sigma}$. We show below that

$$\begin{aligned}
|\mathbf{e}_1^T \hat{\Sigma}'^{-1} \hat{\boldsymbol{\sigma}}_p - \mathbf{e}_1^T \Sigma'^{-1} \boldsymbol{\sigma}_p| \ &\leq\ \frac{1}{C_2}||\hat{\Sigma} - \Sigma||_2 + C_3||\hat{\Sigma}'^{-1} - \Sigma'^{-1}||_2\\
&+\ ||\hat{\Sigma} - \Sigma||_2 ||\hat{\Sigma}'^{-1} - \Sigma'^{-1}||_2, \qquad \text{(B.4)}
\end{aligned}$$

where for a matrix A, $||A||_2$ denote its spectral norm, and C_3 and C_2 are given by Assumption 4.6. To this end, we first apply the inequality

$$|\mathbf{a}_1^T\mathbf{a}_2 - \mathbf{b}_1\mathbf{b}_2| \leq ||\mathbf{b}_1||_2||\mathbf{a}_2 - \mathbf{b}_2||_2 + ||\mathbf{b}_2||_2||\mathbf{a}_1 - \mathbf{b}_1||_2 + ||\mathbf{a}_1 - \mathbf{b}_1||_2||\mathbf{a}_2 - \mathbf{b}_2||_2,$$

with $\mathbf{a}_1 = \hat{\Sigma}_n'^{-1}\mathbf{e}_1$, $\mathbf{a}_2 = \hat{\boldsymbol{\sigma}}_p$, $\mathbf{b}_1 = \Sigma'^{-1}\mathbf{e}_1$ and $\mathbf{b}_2 = \boldsymbol{\sigma}_p$, where for a vector $\mathbf{a}$, $||\mathbf{a}||_2$ denote its ℓ_2 norm. Next, note that $||\Sigma'^{-1}\mathbf{e}_1||_2 \leq ||\Sigma'^{-1}||_2 \leq 1/C_2$, where the last inequality follows from Assumption 4.6 and Cauchy's interlacing theorem for eigenvalues of positive definite matrices, since $|S_{jr}| \leq q \leq \tilde{q}$. Similarly, we have $||\boldsymbol{\sigma}_p||_2 \leq C_3$. This completes the proof of (B.4).

From Lemma B.1, we have $||(X_j, \mathbf{X}_{S_{jr}}^T, X_p)^T||_{\psi_2} < C_4$ for some constant $C_4 > 0$. Therefore, for any $\delta \in (0,1)$ and sufficiently large n, we have

$$\mathbb{P}(||\hat{\Sigma} - \Sigma||_2 > \delta) \leq 2\exp(-C_5 n\delta^2), \tag{B.5}$$

for some absolute constant $C_5 > 0$ depending on C_4 (see Corollary 5.50 of Vershynin [2012]). Similarly, for any $\delta \in (0,1)$ and sufficiently large n, we have

$$\mathbb{P}(||\hat{\Sigma}' - \Sigma'||_2 > \delta) \leq 2\exp(-C_5 n\delta^2). \tag{B.6}$$

We show below that a similar result holds for $||\hat{\Sigma}'^{-1} - \Sigma'^{-1}||_2$. To this end, we consider $\delta \leq C_2/2$ and $||\hat{\Sigma} - \Sigma||_2 \leq \delta$. Using the sub-multiplicity property of the spectral norm, we obtain

$$||(\hat{\Sigma}' - \Sigma')\Sigma'^{-1}||_2 \leq ||\hat{\Sigma}' - \Sigma'||_2||\Sigma'^{-1}||_2 \leq \frac{\delta}{C_2} \leq 1/2 < 1.$$

This implies $(\hat{\Sigma}' - \Sigma')\Sigma'^{-1} + \mathrm{I}$ is invertible and the following inequality holds (see, for example, Section 5.8 of Horn and Johnson [1990]):

$$\begin{aligned} ||((\hat{\Sigma}' - \Sigma')\Sigma'^{-1} + \mathrm{I})^{-1} - \mathrm{I}||_2 &\leq \frac{||(\hat{\Sigma}' - \Sigma')\Sigma'^{-1}||_2}{1 - ||(\hat{\Sigma}' - \Sigma')\Sigma'^{-1}||_2} \\ &\leq 2\,||(\hat{\Sigma}' - \Sigma')\Sigma'^{-1}||_2 \quad \leq \frac{2\delta}{C_2}, \end{aligned} \tag{B.7}$$

where the second inequality follows from $||(\hat{\Sigma}' - \Sigma')\Sigma'^{-1}||_2 \leq 1/2$ and the third inequality follows from the sub-multiplicity property of the spectral norm, the assumption that $||\hat{\Sigma}' - \Sigma'||_2 \leq \delta$ and Assumption 4.6.

Therefore, $||\hat{\Sigma}' - \Sigma'||_2 \leq \delta < C_2/2$ implies

(B.8)
$$||\hat{\Sigma}'^{-1} - \Sigma'^{-1}||_2 = ||\Sigma'^{-1}\{((\hat{\Sigma}' - \Sigma')\Sigma'^{-1} + \mathrm{I})^{-1} - \mathrm{I}\}||_2 \leq ||\Sigma'^{-1}||_2\frac{2\delta}{C_2} \leq \frac{2\delta}{C_2^2},$$

where the second inequality follows from the sub-multiplicity property of the spectral norm and (B.7), and the third inequality follows from Assumption 4.6. By combining (B.6) and (B.8), we have, for any $\delta \in (0, C_2/2)$

$$\mathbb{P}(||\hat{\Sigma}_n'^{-1} - \Sigma_n'^{-1}||_2 > \delta) \leq 2\exp(-C_6 n\delta^2), \tag{B.9}$$

where $C_6 > 0$ is an absolute constant depending on C_2 and C_5.

Finally, by combining, (B.4), (B.5) and (B.6), we obtain

$$\mathbb{P}\left(|\hat{\beta}_{jp|\mathbf{X}_{S_{jr}}} - \beta_{jp|\mathbf{X}_{S_{jr}}}| > \delta\right) \leq 2\exp(-C_7 n\delta^2), \tag{B.10}$$

for some absolute constant $C_7 > 0$ depending on C_3, C_2 and C_4. This completes the proof. □

Proof of the Third Part of Theorem 4.1. Recall that $\hat{\eta}_j(\lambda_n) = \hat{\beta}_{1j} \cdot \text{aver}(\hat{\Theta}_{jp}(\lambda_n))$, and $\theta_{1j} = \beta_{1j}$. Therefore, we have

$$\begin{aligned} |\hat{\eta}_j(\lambda_n) - \theta_{1j} \cdot \text{aver}(\Theta_{jp})| \;\leq\; & |\theta_{1j}| \cdot |\text{aver}(\hat{\Theta}_{jp}(\lambda_n) - \text{aver}(\Theta_{jp})| \\ & + \; |\text{aver}(\Theta_{jp})| \cdot |\hat{\beta}_{1j} - \beta_{1j}|. \end{aligned} \tag{B.11}$$

From the second part of Theorem 4.1, we have

$$\max_{1<j<p} \left|\text{aver}(\hat{\Theta}_{jp}(\lambda_n)) - \text{aver}(\Theta_{jp})\right| \xrightarrow{\mathbb{P}} 0. \tag{B.12}$$

Further, by similar argument as given in the proof of the second part of Theorem 4.1, we can show that

$$\mathbb{P}\left(\max_{1<j<p} |\hat{\beta}_{1j} - \beta_{1j}| > \delta\right) \leq 2(p-2)\exp(-C_7 n\delta^2) \longrightarrow 0, \tag{B.13}$$

where C_7 is as in (B.10).

Finally, note that for any $i, k \in \{2, \ldots, p\}$ and $S \subseteq \{1, \ldots, p\} \setminus \{i, k\}$ such that $|S| \leq \tilde{q}$, we have

$$|\beta_{ik|\mathbf{X}_S}| = |\mathbf{e}_1^T \Sigma'^{-1} \boldsymbol{\sigma}_k| \leq ||\Sigma'^{-1}\mathbf{e}_1||_2 ||\boldsymbol{\sigma}_k||_2 \leq \frac{C_3}{C_2}, \tag{B.14}$$

where Σ and Σ' denote the submatrices of $\Sigma_0 = \text{Cov}(\mathbf{X})$ corresponding to $(X_i, \mathbf{X}_S, X_k)$ and $(X_i, \mathbf{X}_S)$ respectively, $\boldsymbol{\sigma}_k$ is the last column of Σ Note that the first inequality in (B.14) is Cauchy-Schwarz and the second inequality in (B.14) follows from Assumption 4.6 and Cauchy's interlacing theorem for eigenvalues of positive definite matrices (we used similar arguments in the proof of the second part of Theorem 4.1).

Since (B.14) implies that $|\theta_{1j}|$ and $|\text{aver}(\Theta_{njp})|$ are bounded above by C_3/C_2 for all $j \in \{2, \ldots, p\}$, the the third part of Theorem 4.1 follows from (B.11), (B.12) and (B.13). □

B.4. Proof of Theorem 5.1. We first state two supporting lemmas that serve as essential ingredients in our proof of Theorem 5.1 and may also be of independent interest. Their proofs are given in Sections B.5 and B.6.

LEMMA B.2. *Suppose Assumption 5.1(i) holds for a given* $\mathcal{S} \subseteq \Omega_{\mathcal{J}}$ *with* $|\mathcal{S}| := L_n \equiv L_{n,\mathcal{S}}$ *and* $\sup_{S \in \mathcal{S}} |S| \leq q_n \equiv q_{n,\mathcal{S}} \leq \min(n, p_n)$*, and let* $r_n := (q_n + \log L_n)$*. Then, for any* $c > 0$*, the following bounds hold:*

(B.15)

$$(i)\ \mathbb{P}\left\{\sup_{S\in\mathcal{S}} \left\|\widetilde{\mathbf{\Sigma}}_S - \mathbf{\Sigma}_S\right\|_2 > (c+1)K_{\mathcal{S}}\left(\sqrt{\frac{r_n}{n}} + \frac{r_n}{n}\right)\right\} \leq\ 2\exp\left(-cr_n\right),$$

$$(ii)\ \mathbb{P}\left\{\sup_{S\in\mathcal{S}} \left\|\widehat{\mathbf{\Gamma}}_S\right\|_2 > 16(c+1)K_{\mathcal{S}}\left(\frac{\sqrt{r_n}}{n} + \frac{r_n}{n}\right) + \frac{\lambda_{\sup,\mathcal{S}}}{n}\right\} \leq\ 2\exp\left(-cr_n\right),$$

$$(iii)\ \mathbb{P}\left\{\sup_{S\in\mathcal{S}} \left\|\widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}_S\right\|_2 > (c+1)K_{\mathcal{S}}\left(\sqrt{\frac{r_n}{n}} + \frac{33r_n}{n}\right) + \frac{\lambda_{\sup,\mathcal{S}}}{n}\right\} \leq\ 4\exp\left(-cr_n\right).$$

Further, let $\widetilde{r}_n := (r_n + 1)$*, and suppose Assumption 5.1(ii) also holds. Then, for any* $c > 0$*, the following bounds hold:*

(B.16)

$$(i)\ \mathbb{P}\left\{\sup_{S\in\mathcal{S}} \left\|\widetilde{\mathbf{\Sigma}}_{S,Y} - \mathbf{\Sigma}_{S,Y}\right\|_2 > (c+1)\widetilde{K}_{\mathcal{S}}\left(\sqrt{\frac{\widetilde{r}_n}{n}} + \frac{\widetilde{r}_n}{n}\right)\right\} \leq\ 2\exp\left(-c\widetilde{r}_n\right),$$

$$(ii)\ \mathbb{P}\left\{\sup_{S\in\mathcal{S}} \left\|\widehat{\mathbf{\Gamma}}_{S,Y}\right\|_2 > 16(c+1)\widetilde{K}_{\mathcal{S}}\left(\frac{\sqrt{\widetilde{r}_n}}{n} + \frac{\widetilde{r}_n}{n}\right) + \frac{\widetilde{\lambda}_{\sup,\mathcal{S}}}{n}\right\} \leq\ 2\exp(-c\widetilde{r}_n).$$

Lastly, the constants $\lambda_{\sup,\mathcal{S}}$ *and* $\widetilde{\lambda}_{\sup,\mathcal{S}}$ *may be chosen such that* $\lambda_{\sup,\mathcal{S}} \leq 2\sigma^2_{\mathbf{X},\mathcal{S}}$ *and* $\widetilde{\lambda}_{\sup,\mathcal{S}} \leq 2\widetilde{\sigma}^2_{\mathbf{Z},\mathcal{S}} \equiv 2(\sigma_Y + \sigma_{\mathbf{X},\mathcal{S}})^2$*. Moreover,* $\sup_{s\in\mathcal{S}} \|\boldsymbol{\beta}_S\|_2^2 \leq \mathrm{Var}(Y)\lambda^{-1}_{\inf,\mathcal{S}} \leq 2\sigma_Y^2\lambda^{-1}_{\inf,\mathcal{S}}$*.*

LEMMA B.3. *Let* $\mathcal{S}$ *and* r_n *be as in Lemma B.2, and suppose Assumption 5.1(i) holds. Let* $c^* > 0$ *be any constant satisfying:*

$$(c^*+1)K_{\mathcal{S}}\left(\sqrt{\frac{r_n}{n}} + \frac{33r_n}{n}\right) + \frac{\lambda_{\sup,\mathcal{S}}}{n}\ \leq\ \frac{1}{2}\lambda_{\inf,\mathcal{S}}, \tag{B.17}$$

and let $K^*_{\mathcal{S}} := 2\lambda^{-2}_{\inf,\mathcal{S}}K_{\mathcal{S}}$*. Then, for any* $c^* > 0$ *as in (B.17) and for any*

$c > 0$, *and defining* $a_n(c, c^*, r_n) := \exp(-cr_n) + \exp(-c^* r_n)$, *we have*

$$\mathbb{P}\left\{\sup_{S\in\mathcal{S}}\left\|\widehat{\boldsymbol{\Sigma}}_S^{-1} - \boldsymbol{\Sigma}_S^{-1}\right\|_2 > (c+1)K_{\mathcal{S}}^*\left(\sqrt{\frac{r_n}{n}} + \frac{33r_n}{n}\right) + \frac{2}{n}\frac{\lambda_{\sup,\mathcal{S}}}{\lambda_{\inf,\mathcal{S}}^2}\right\} \\ \leq\; 4a_n(c, c^*, r_n) \;\equiv\; 4\exp(-cr_n) + 4\exp(-c^* r_n).$$

The proof of Theorem 5.1 essentially follows from carefully combining all the results established in Lemmas B.2 and B.3. To this end, first note that under Assumption 5.1 and using Lemmas B.2 and B.3, we have: for any $c > 0$ and any $c^* > 0$ satisfying condition (B.17),

$$\mathbb{P}\left\{\sup_{S\in\mathcal{S}}\left\|\widetilde{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}_S\right\|_2 > \epsilon_{n,1}(c, r_n)\right\} \leq 2\exp(-cr_n), \;\; \sup_{S\in\mathcal{S}}\|\boldsymbol{\beta}_S\|_2 \leq C_{\mathcal{S}}, \tag{B.18}$$

$$\mathbb{P}\left\{\sup_{S\in\mathcal{S}}\left\|\widetilde{\boldsymbol{\Sigma}}_{S,Y} - \boldsymbol{\Sigma}_{S,Y}\right\|_2 > \epsilon_{n,2}(c, r_n)\right\} \leq 2\exp(-c\widetilde{r}_n) \leq 2\exp(-cr_n), \tag{B.19}$$

$$\mathbb{P}\left\{\sup_{S\in\mathcal{S}}\left\|\widehat{\boldsymbol{\Sigma}}_S^{-1} - \boldsymbol{\Sigma}_S^{-1}\right\|_2 > \delta_n(c, r_n)\right\} \;\leq\; 4\exp(-cr_n) + 4\exp(-c^* r_n), \tag{B.20}$$

$$\mathbb{P}\left\{\sup_{S\in\mathcal{S}}\left\|\widehat{\boldsymbol{\Gamma}}_{S,Y}\right\|_2 > \eta_{n,2}(c, r_n)\right\} \;\leq\; 2\exp(-c\widetilde{r}_n) \;\leq\; 2\exp(-cr_n), \;\; \text{and} \tag{B.21}$$

$$\mathbb{P}\left\{\sup_{S\in\mathcal{S}}\left\|\widehat{\boldsymbol{\Gamma}}_S\right\|_2 > \eta_{n,1}(c, r_n)\right\} \;\leq\; 2\exp(-cr_n), \;\; \sup_{S\in\mathcal{S}}\left\|\boldsymbol{\Sigma}_S^{-1}\right\|_2 \leq \lambda_{\inf,\mathcal{S}}^{-1}, \tag{B.22}$$

where $\{\epsilon_{n,j}(c, r_n), \eta_{n,j}(c, r_n)\}_{j=1}^2$, $\delta_n(c, r_n)$ and $C_{\mathcal{S}}$ are all as defined in Theorem 5.1. Note that for (B.21) and (B.22), we also used $\sqrt{r_n} \leq r_n$, and for (B.19) and (B.21), we used $r_n \leq \widetilde{r}_n$.

Next, noting that $(\boldsymbol{\Sigma}_{S,Y} - \boldsymbol{\Sigma}_S\boldsymbol{\beta}_S) = \mathbf{0}$ for any $S \in \mathcal{S}$, due to (5.1), we have:

$$\begin{aligned} \sup_{S\in\mathcal{S}}\|\mathbf{T}_{n,S}\|_2 \;&\equiv\; \sup_{S\in\mathcal{S}}\left\|(\widehat{\boldsymbol{\Sigma}}_S^{-1} - \boldsymbol{\Sigma}_S^{-1})\{(\widetilde{\boldsymbol{\Sigma}}_{S,Y} - \boldsymbol{\Sigma}_{S,Y}) - (\widetilde{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}_S)\boldsymbol{\beta}_S\}\right\|_2 \\ &\leq\; \sup_{S\in\mathcal{S}}\left\|\widehat{\boldsymbol{\Sigma}}_S^{-1} - \boldsymbol{\Sigma}_S^{-1}\right\|_2 \sup_{S\in\mathcal{S}}\left\|\widetilde{\boldsymbol{\Sigma}}_{S,Y} - \boldsymbol{\Sigma}_{S,Y}\right\|_2 \qquad \text{(B.23)} \\ &\quad + \sup_{S\in\mathcal{S}}\left\|\widehat{\boldsymbol{\Sigma}}_S^{-1} - \boldsymbol{\Sigma}_S^{-1}\right\|_2 \sup_{S\in\mathcal{S}}\left\|\widetilde{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}_S\right\|_2 \sup_{S\in\mathcal{S}}\|\boldsymbol{\beta}_S\|_2, \end{aligned}$$

where the inequality in (B.23) follows from multiple applications of Lemma C.4 (i). Using (B.18), (B.19) and (B.20) in (B.23), along with the union bound, we have: for any $c > 0$,

$$\mathbb{P}\left[\sup_{S\in\mathcal{S}} \|\mathbf{T}_{n,S}\|_2 > \delta_n(c, r_n)\{\epsilon_{n,1}(c, r_n)C_{\mathcal{S}} + \epsilon_{n,2}(c, r_n)\}\right] \\ \leq\ 8\exp(-cr_n) + 4\exp(-c^* r_n).$$

This establishes the first of the two claims in result (i) of Theorem 5.1. $\square$

Next, recall $\mathbf{R}_{n,S} \equiv \mathbf{R}^{(1)}_{n,S} - \mathbf{R}^{(2)}_{n,S} \equiv \widehat{\boldsymbol{\Sigma}}_S^{-1}(\widehat{\boldsymbol{\Gamma}}_S\boldsymbol{\beta}_S - \widehat{\boldsymbol{\Gamma}}_{S,Y})$, and hence using Lemma C.4 (i),

$$\sup_{S\in\mathcal{S}} \|\mathbf{R}_{n,S}\|_2 \ \leq\ \sup_{S\in\mathcal{S}} \left\|\widehat{\boldsymbol{\Sigma}}_S^{-1}\right\|_2 \left(\sup_{S\in\mathcal{S}} \left|\left|\widehat{\boldsymbol{\Gamma}}_S\right|\right|_2 \sup_{S\in\mathcal{S}} \|\boldsymbol{\beta}_S\|_2 + \sup_{S\in\mathcal{S}} \left\|\widehat{\boldsymbol{\Gamma}}_{S,Y}\right\|_2\right). \tag{B.24}$$

Consequently, using (B.20), (B.21) and (B.22) in (B.24), along with the union bound, we have:

$$\mathbb{P}\left[\sup_{S\in\mathcal{S}} \|\mathbf{R}_{n,S}\|_2 > \{\delta_n(c, r_n) + \lambda^{-1}_{\text{inf},\mathcal{S}}\}\{\epsilon_{n,1}(c, r_n)C_{\mathcal{S}} + \epsilon_{n,2}(c, r_n)\}\right] \\ \leq\ 8\exp(-cr_n) + 4\exp(-c^* r_n) \quad \forall\ c > 0.$$

This establishes the second and final claim in result (i) of Theorem 5.1. $\square$

Finally, recall that $\boldsymbol{\Psi}_S(\mathbf{Z}) = \boldsymbol{\Sigma}_S^{-1}\boldsymbol{\psi}_S(\mathbf{Z})$ and $n^{-1}\sum_{i=1}^n \boldsymbol{\psi}_S(\mathbf{Z}_i) \equiv \widetilde{\boldsymbol{\Sigma}}_{S,Y} - \widetilde{\boldsymbol{\Sigma}}_S\boldsymbol{\beta}_S$. Further $(\boldsymbol{\Sigma}_{S,Y} - \boldsymbol{\Sigma}_S\boldsymbol{\beta}_S) = \mathbf{0}\ \forall\ S \in \mathcal{S}$, due to (5.1). Hence, using Lemma C.4 (i), we have:

$$\sup_{S\in\mathcal{S}} \left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{\Psi}_S(\mathbf{Z}_i)\right\|_2 \ \equiv\ \sup_{S\in\mathcal{S}} \left\|\boldsymbol{\Sigma}_S^{-1}\{(\widetilde{\boldsymbol{\Sigma}}_{S,Y} - \boldsymbol{\Sigma}_{S,Y}) - (\widetilde{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}_S)\boldsymbol{\beta}_S\}\right\|_2 \\ \leq\ \sup_{S\in\mathcal{S}} \left|\left|\boldsymbol{\Sigma}_S^{-1}\right|\right|_2 \left(\sup_{S\in\mathcal{S}} \left\|\widetilde{\boldsymbol{\Sigma}}_{S,Y} - \boldsymbol{\Sigma}_{S,Y}\right\|_2 + \sup_{S\in\mathcal{S}} \left|\left|\widetilde{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}_S\right|\right|_2 \sup_{S\in\mathcal{S}} \|\boldsymbol{\beta}_S\|_2\right).$$

Hence, we have: for any $c > 0$,

$$\mathbb{P}\left[\sup_{S\in\mathcal{S}} \left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{\Psi}_S(\mathbf{Z}_i)\right\|_2 > \lambda^{-1}_{\text{inf},\mathcal{S}}\{\epsilon_{n,1}(c, r_n)C_{\mathcal{S}} + \epsilon_{n,2}(c, r_n)\}\right] \\ \leq\ 4\exp(-cr_n) \quad \forall\ c > 0,$$

where the final probability bound follows from applying (B.18), (B.19) and (B.22), along with the union bound, to the preceding bound. This establishes the result (ii) of Theorem 5.1. $\square$

Finally, all the '$\lesssim$' type bounds claimed in results (i) and (ii) are quite straightforward and follow trivially from the definitions of $\{\epsilon_{n,j}(c, r_n)\}_{j=1}^2\}$, $\{\eta_{n,j}(c, r_n)\}_{j=1}^2$, $\delta_n(c, r_n)$ and $C_{\mathcal{S}}$. The details are thus skipped here for brevity. The proof of Theorem 5.1 is now complete. □

B.5. Proof of Lemma B.2. Applying Lemma C.6, under Assumption 5.1 (i), to the random vectors $\{\mathbf{X}_{S,i} - \boldsymbol{\mu}_S\}_{i=1}^n$ for any $S \in \mathcal{S}$, and recalling the definition of the constant $K_{\mathcal{S}} > 0$ in (5.4) along with the fact that $s \leq q_n$ $\forall\ S \in \mathcal{S}$, it follows that for any $\epsilon \geq 0$ and for each $S \in \mathcal{S}$,

$$\mathbb{P}\left\{\left|\left|\widetilde{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}_S\right|\right|_2 \ > \ K_{\mathcal{S}}\left(\sqrt{\frac{q_n + \epsilon}{n}} + \frac{q_n + \epsilon}{n}\right)\right\} \ \leq \ 2\exp\left(-\epsilon\right),$$

or equivalently, for any $\epsilon \geq 0$ and $S \in \mathcal{S}$,

$$\mathbb{P}\left\{\left|\left|\widetilde{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}_S\right|\right|_2 \ > \ K_{\mathcal{S}}\left(\sqrt{\epsilon} + \epsilon\right)\right\} \ \leq \ 2\exp\left(-n\epsilon + q_n\right). \tag{B.25}$$

Consequently, using (B.25) along with the union bound, we then have:

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{S \in \mathcal{S}}\left|\left|\widetilde{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}_S\right|\right|_2 \ > \ K_{\mathcal{S}}\left(\sqrt{\epsilon} + \epsilon\right)\right\} \\
&\quad \leq \ \sum_{s \in \mathcal{S}}\mathbb{P}\left\{\left|\left|\widetilde{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}_S\right|\right|_2 \ > \ K_{\mathcal{S}}\left(\sqrt{\epsilon} + \epsilon\right)\right\} \\
&\quad \leq \ 2L_n\exp\left(-n\epsilon + q_n\right) \ \equiv \ 2\exp\left(-n\epsilon + q_n + \log L_n\right) \quad \forall\ \epsilon \geq 0. \qquad \text{(B.26)}
\end{aligned}$$

Substituting ϵ in (B.26) above as: $\epsilon = (c+1)(q_n + \log L_n)/n \equiv (c+1)r_n/n$ for any $c \geq 0$, and noting that $\sqrt{c+1} \leq (c+1)$, we then have: $\forall\ c \geq 0$,

$$\mathbb{P}\left\{\sup_{S \in \mathcal{S}}\left|\left|\widetilde{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}_S\right|\right|_2 \ > \ (c+1)K_{\mathcal{S}}\left(\sqrt{\frac{r_n}{n}} + \frac{r_n}{n}\right)\right\} \ \leq \ 2\exp(-cr_n). \tag{B.27}$$

This therefore establishes the first claim (i) in (B.15). □

Next, using Lemma C.3, along with Lemma C.2 (i) and the definition of $\|\cdot\|_{\psi_2}^*$ in A.2, it follows, under Assumption 5.1 (i), that for any $S \in \mathcal{S}$,

$$\begin{aligned}
&\left\|\mathbf{v}^T(\overline{\mathbf{X}}_S - \boldsymbol{\mu}_S)\right\|_{\psi_2} \ \leq \ (4\sigma_{\mathbf{X},\mathcal{S}}/\sqrt{n})\left\|\mathbf{v}\right\|_2 \quad \text{for any } \mathbf{v} \in \mathbb{R}^s, \quad \text{and thus,} \\
&\sup_{S \in \mathcal{S}}\left\|\overline{\mathbf{X}}_S - \boldsymbol{\mu}_S\right\|_{\psi_2}^* \ \leq \ (4\sigma_{\mathbf{X},\mathcal{S}}/\sqrt{n}).
\end{aligned}$$

Further, $\overline{\boldsymbol{\Sigma}}_{n,S} := \mathbb{E}\{\overline{\mathbf{X}}_S - \boldsymbol{\mu}_S)(\overline{\mathbf{X}}_S - \boldsymbol{\mu}_S)^T \equiv \mathrm{Cov}(\overline{\mathbf{X}}_S - \boldsymbol{\mu}_S) = n^{-1}\boldsymbol{\Sigma}_S$, so that $\left|\left|\bar{\boldsymbol{\Sigma}}_{n,S}\right|\right|_2 \equiv n^{-1}\lambda_{\max}(\boldsymbol{\Sigma}_S) \leq n^{-1}\lambda_{\sup,\mathcal{S}}$. Hence, using Lemma C.6 again, this time applied to (a single observation of) $\overline{\mathbf{X}}_S - \boldsymbol{\mu}_S$ for any $S \in \mathcal{S}$, we have: for any $\epsilon \geq 0$ and any $S \in \mathcal{S}$,

$$\mathbb{P}\left\{ \left|\left|\widehat{\boldsymbol{\Gamma}}_S\right|\right|_2 \; > \; \frac{\lambda_{\sup,\mathcal{S}}}{n} + \frac{16}{n} K_{\mathcal{S}} \left(\sqrt{q_n + \epsilon} + q_n + \epsilon\right) \right\} \; \leq \; 2\exp(-\epsilon),$$

or equivalently, for any $\epsilon \geq 0$ and $S \in \mathcal{S}$,

$$\text{(B.28)} \quad \mathbb{P}\left\{ \left|\left|\widehat{\boldsymbol{\Gamma}}_S\right|\right|_2 \; > \; \frac{\lambda_{\sup,\mathcal{S}}}{n} + \frac{16}{n} K_{\mathcal{S}} \left(\sqrt{\epsilon} + \epsilon\right) \right\} \; \leq \; 2\exp(-\epsilon + q_n).$$

Consequently, using (B.28) along with the union bound, similar to the arguments used earlier for obtaining (B.26), we have: for any $\epsilon \geq 0$,

(B.29)

$$\mathbb{P}\left\{ \sup_{S \in \mathcal{S}} \left|\left|\widehat{\boldsymbol{\Gamma}}_S\right|\right|_2 \; > \; \frac{\lambda_{\sup,\mathcal{S}}}{n} + \frac{16}{n} K_{\mathcal{S}} \left(\sqrt{\epsilon} + \epsilon\right) \right\} \; \leq 2\exp\left(-\epsilon + q_n + \log L_n\right).$$

Substituting ϵ in (B.29) above as: $\epsilon = (c+1)(q_n + \log L_n) \equiv (c+1)r_n$ for any $c \geq 0$, and noting that $\sqrt{c+1} \leq (c+1)$, we then have: $\forall \; c \geq 0$,

(B.30)

$$\mathbb{P}\left\{ \sup_{S \in \mathcal{S}} \left|\left|\widehat{\boldsymbol{\Gamma}}_S\right|\right|_2 \; > \; 16(c+1) K_{\mathcal{S}} \left(\frac{\sqrt{r_n}}{n} + \frac{r_n}{n}\right) + \frac{\lambda_{\sup,\mathcal{S}}}{n} \right\} \; \leq \; 2\exp(-c r_n).$$

This establishes the second claim (ii) in (B.15). □

Finally, the third claim in (B.15) follows from a simple application of the triangle inequality, along with combination (via the union bound) of the bounds in (B.27) and (B.30) with a slight adjustment applied to (B.30). Specifically, since $\widehat{\boldsymbol{\Sigma}}_S \equiv \widetilde{\boldsymbol{\Sigma}}_S - \widehat{\boldsymbol{\Gamma}}_S \; \forall \; S \in \mathcal{S}$, we have:

$$\begin{aligned} \sup_{S \in \mathcal{S}} \left|\left|\widehat{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}_S\right|\right|_2 \; &\equiv \; \sup_{S \in \mathcal{S}} \left|\left|(\widetilde{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}_S) + \widehat{\boldsymbol{\Gamma}}_S\right|\right|_2 \\ &\leq \; \sup_{S \in \mathcal{S}} \left|\left|\widetilde{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}_S\right|\right|_2 + \sup_{S \in \mathcal{S}} \left|\left|\widehat{\boldsymbol{\Gamma}}_S\right|\right|_2 . \end{aligned}$$

Hence, combining (B.27) and (B.30) through the union bound, and simplifying the resulting bound further by noting that $\sqrt{r_n}/n \leq r_n/n$, we then have:

$$\begin{aligned} \text{(B.31)} \quad & \mathbb{P}\left\{ \sup_{S \in \mathcal{S}} \left|\left|\widehat{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}_S\right|\right|_2 \; > \; (c+1) K_{\mathcal{S}} \left(\sqrt{\frac{r_n}{n}} + \frac{33 r_n}{n}\right) + \frac{\lambda_{\sup,\mathcal{S}}}{n} \right\} \\ & \qquad \leq \; 4\exp(-c r_n) \quad \text{for any } c \geq 0. \end{aligned}$$

This now establishes the third and final claim (iii) in (B.15). □

To establish the claims (i) and (ii) in (B.16), we first recall the definitions of $\mathbf{Z}_S$, $\boldsymbol{\nu}_S$ and $\boldsymbol{\Xi}_S$ from Assumption 5.1 (ii) and further, with $\{\mathbf{Z}_{S,i}\}_{i=1}^n := \{(Y_i, \mathbf{X}_{S,i})\}_{i=1}^n$, we define: $\forall\, S \in \mathcal{S}$,

$$\widetilde{\boldsymbol{\Xi}}_S := \frac{1}{n}\sum_{i=1}^n (\mathbf{Z}_{S,i} - \boldsymbol{\nu}_S)(\mathbf{Z}_{S,i} - \boldsymbol{\nu}_S)^T, \quad \overline{\mathbf{Z}}_S := \frac{1}{n}\sum_{i=1}^n \mathbf{Z}_{S,i} \quad \text{and}$$
$$\widehat{\boldsymbol{\Upsilon}}_S := (\overline{\mathbf{Z}}_S - \boldsymbol{\nu}_S)(\overline{\mathbf{Z}}_S - \boldsymbol{\nu}_S)^T.$$

Then, note that the vectors $(\widetilde{\boldsymbol{\Sigma}}_{S,Y} - \boldsymbol{\Sigma}_{S,Y})$ and $\widehat{\boldsymbol{\Gamma}}_{S,Y}$ are simply $s \times 1$ sub-matrices of the matrices $\widetilde{\boldsymbol{\Xi}}_S$ and $\widehat{\boldsymbol{\Upsilon}}_S$ respectively. Hence, using Lemma C.4 (iii), we deterministically have: for any $S \in \mathcal{S}$,

$$\text{(B.32)} \qquad \left\|\widetilde{\boldsymbol{\Sigma}}_{S,Y} - \boldsymbol{\Sigma}_{S,Y}\right\|_2 \le \left\|\widetilde{\boldsymbol{\Xi}}_S - \boldsymbol{\Xi}_S\right\|_2 \quad \text{and} \quad \left\|\widehat{\boldsymbol{\Gamma}}_{S,Y}\right\|_2 \le \left\|\widehat{\boldsymbol{\Upsilon}}_S\right\|_2.$$

Next, under Assumption 5.1 (i), note that for any $\mathbf{u} \equiv (a, \mathbf{v}) \in \mathbb{R}^{s+1}$ with $a \in \mathbb{R}, \mathbf{v} \in \mathbb{R}^s$, and for any $S \in \mathcal{S}$, we have: $\left\|\mathbf{u}^T(\mathbf{Z}_S - \boldsymbol{\nu}_S)\right\|_{\psi_2} \equiv \left\|a\widetilde{Y} + \mathbf{v}^T\widetilde{\mathbf{X}}_S\right\|_{\psi_2} \le |a|\sigma_Y + \|\mathbf{v}\|_2 \|\widetilde{\mathbf{X}}_S\|_{\psi_2}^* \le \|\mathbf{u}\|_2 (\sigma_Y + \sigma_{\mathbf{X},\mathcal{S}})$, where the steps follow through repeated use of Lemma C.1 (i), along with use of Lemma C.2 (i), Assumption 5.1 (i) and the definition of $\|\cdot\|_{\psi_2}^*$ in A.2. Using Definition A.2 again, we therefore have: $\sup_{S\in\mathcal{S}} \|\mathbf{Z}_S - \boldsymbol{\nu}_S\|_{\psi_2}^* \le \widetilde{\sigma}_{\mathbf{Z},\mathcal{S}} \equiv (\sigma_Y + \sigma_{\mathbf{X},\mathcal{S}})$.

Hence, similar to (B.25), applying Lemma C.6 to the random vectors $\{\mathbf{Z}_{S,i} - \boldsymbol{\nu}_S\}_{i=1}^n$ for any $S \in \mathcal{S}$, and recalling the definition of the constant $\widetilde{K}_{\mathcal{S}} > 0$ in Assumption 5.1 (ii) along with the fact that $\dim(\mathbf{Z}_S) \le \widetilde{q}_n := q_n + 1\ \forall\, S \in \mathcal{S}$, it follows that for any $\epsilon \ge 0$,

$$\mathbb{P}\left\{\left\|\widetilde{\boldsymbol{\Xi}}_S - \boldsymbol{\Xi}_S\right\|_2 > \widetilde{K}_{\mathcal{S}}\left(\sqrt{\epsilon} + \epsilon\right)\right\} \le 2\exp\left(-n\epsilon + \widetilde{q}_n\right) \quad \forall\, S \in \mathcal{S},$$

and therefore, for any $\epsilon \ge 0$,

$$\text{(B.33)} \quad \mathbb{P}\left\{\sup_{S\in\mathcal{S}}\left\|\widetilde{\boldsymbol{\Xi}}_S - \boldsymbol{\Xi}_S\right\|_2 > \widetilde{K}_{\mathcal{S}}\left(\sqrt{\epsilon} + \epsilon\right)\right\} \le 2\exp\left(-n\epsilon + \widetilde{q}_n + \log L_n\right),$$

where the last bound follows from using the union bound, similar to (B.26). Consequently, for any $\epsilon \ge 0$,

$$\mathbb{P}\left\{\sup_{S\in\mathcal{S}}\left\|\widetilde{\boldsymbol{\Sigma}}_{S,Y} - \boldsymbol{\Sigma}_{S,Y}\right\|_2 > \widetilde{K}_{\mathcal{S}}\left(\sqrt{\epsilon} + \epsilon\right)\right\} \le 2\exp\left(-n\epsilon + \widetilde{q}_n + \log L_n\right),$$

and hence, for any $c \ge 0$,

$$\text{(B.34)} \quad \mathbb{P}\left\{\sup_{S\in\mathcal{S}}\left\|\widetilde{\boldsymbol{\Sigma}}_{S,Y} - \boldsymbol{\Sigma}_{S,Y}\right\|_2 > (c+1)\widetilde{K}_{\mathcal{S}}\left(\sqrt{\frac{\widetilde{r}_n}{n}} + \frac{\widetilde{r}_n}{n}\right)\right\} \le 2\exp(-c\widetilde{r}_n),$$

where the first bound follows from using (B.32) and (B.33), and the second bound follows from substituting ϵ as: $\epsilon = (c+1)(\widetilde{q}_n + \log L_n)/n \equiv (c+1)\widetilde{r}_n/n$ for any $c \geq 0$, and noting that $\sqrt{c+1} \leq (c+1)$. This therefore establishes the first claim (i) in (B.16). □

Next, similar to arguments used to prove claim (ii) in (B.15), it follows using Lemma C.3, Lemma C.2 (i) and the definition of $\|\cdot\|_{\psi_2}^*$ in A.2, that

$$\left\|\mathbf{u}^T(\overline{\mathbf{Z}}_S - \boldsymbol{\nu}_S)\right\|_{\psi_2} \leq (4\widetilde{\sigma}_{\mathbf{Z},\mathcal{S}}/\sqrt{n})\,\|\mathbf{u}\|_2\,, \quad \text{for any } S \in \mathcal{S} \text{ and any } \mathbf{u} \in \mathbb{R}^{s+1},$$

and thus,

$$\sup_{S\in\mathcal{S}} \left\|\overline{\mathbf{Z}}_S - \boldsymbol{\nu}_S\right\|_{\psi_2}^* \leq (4\widetilde{\sigma}_{\mathbf{Z},\mathcal{S}}/\sqrt{n}).$$

Further, $\overline{\boldsymbol{\Xi}}_{n,S} := \mathbb{E}\{\overline{\mathbf{Z}}_S - \boldsymbol{\nu}_S)(\overline{\mathbf{Z}}_S - \boldsymbol{\nu}_S)^T \equiv \text{Cov}(\overline{\mathbf{Z}}_S - \boldsymbol{\nu}_S) = n^{-1}\boldsymbol{\Xi}_S$, so that $\left|\left|\bar{\boldsymbol{\Xi}}_{n,S}\right|\right|_2 \equiv n^{-1}\lambda_{\max}(\boldsymbol{\Xi}_S) \leq n^{-1}\widetilde{\lambda}_{\text{sup},\mathcal{S}}$. Hence, similar to (B.28), applying Lemma C.6 to (a single observation of) $\overline{\mathbf{Z}}_S - \boldsymbol{\mu}_S$ for any $S \in \mathcal{S}$, we have: for any $\epsilon \geq 0$ and any $S \in \mathcal{S}$,

$$\mathbb{P}\left\{\left|\left|\widehat{\boldsymbol{\Upsilon}}_S\right|\right|_2 > \frac{\widetilde{\lambda}_{\text{sup},\mathcal{S}}}{n} + \frac{16}{n}\widetilde{K}_{\mathcal{S}}\left(\sqrt{\epsilon} + \epsilon\right)\right\} \leq 2\exp(-\epsilon + \widetilde{q}_n),$$

and therefore, for any $\epsilon \geq 0$,

(B.35)

$$\mathbb{P}\left\{\sup_{S\in\mathcal{S}}\left|\left|\widehat{\boldsymbol{\Upsilon}}_S\right|\right|_2 > \frac{\widetilde{\lambda}_{\text{sup},\mathcal{S}}}{n} + \frac{16}{n}\widetilde{K}_{\mathcal{S}}\left(\sqrt{\epsilon} + \epsilon\right)\right\} \leq 2\exp\left(-\epsilon + \widetilde{q}_n + \log L_n\right),$$

where the last bound follows from using the union bound, similar to (B.29). Consequently, for any $\epsilon \geq 0$,

$$\mathbb{P}\left\{\sup_{S\in\mathcal{S}}\left\|\widehat{\boldsymbol{\Gamma}}_{S,Y}\right\|_2 > \frac{\widetilde{\lambda}_{\text{sup},\mathcal{S}}}{n} + \frac{16}{n}\widetilde{K}_{\mathcal{S}}\left(\sqrt{\epsilon} + \epsilon\right)\right\} \leq 2\exp\left(-\epsilon + \widetilde{q}_n + \log L_n\right),$$

$c \geq 0$, and hence, for any

(B.36)

$$\mathbb{P}\left\{\sup_{S\in\mathcal{S}}\left\|\widehat{\boldsymbol{\Gamma}}_{S,Y}\right\|_2 > 16(c+1)\widetilde{K}_{\mathcal{S}}\left(\frac{\sqrt{\widetilde{r}_n}}{n} + \frac{\widetilde{r}_n}{n}\right) + \frac{\widetilde{\lambda}_{\text{sup},\mathcal{S}}}{n}\right\} \leq 2\exp(-c\widetilde{r}_n),$$

where the bounds follow from using (B.32) and (B.35) and the second bound follows from substituting ϵ as: $\epsilon = (c+1)(\widetilde{q}_n + \log L_n) \equiv (c+1)\widetilde{r}_n$ for any

$c \geq 0$, and noting that $\sqrt{c+1} \leq (c+1)$. This therefore establishes the second claim (ii) in (B.16). □

The remaining claims at the end of Lemma B.2 are quite straightforward. Using Lemma C.2 (i) and Lemma C.1 (iii), we first note that under Assumption 5.1, for any $S \in \mathcal{S}$, $\mathbf{v} \in \mathbb{R}^s$ and $\mathbf{u} \in \mathbb{R}^{s+1}$, $\mathbb{E}\{(\mathbf{v}^T \widetilde{\mathbf{X}}_S)^2\} \leq 2\sigma^2_{\mathbf{X},\mathcal{S}} \|\mathbf{v}\|_2^2$ and $\mathbb{E}\{(\mathbf{u}^T \widetilde{\mathbf{Z}}_S)^2\} \leq 2\widetilde{\sigma}^2_{\mathbf{Z},\mathcal{S}}$. Further, $\lambda_{\max}(\mathbf{\Sigma}_S) \equiv \sup_{\|\mathbf{v}\|_2 \leq 1} \mathbb{E}\{(\mathbf{v}^T \widetilde{\mathbf{X}}_S)^2\}$ and $\lambda_{\max}(\mathbf{\Xi}_S) \equiv \sup_{\|\mathbf{u}\|_2 \leq 1} \mathbb{E}\{(\mathbf{u}^T \widetilde{\mathbf{Z}}_S)^2\}$ for each $S \in \mathcal{S}$. Hence, we have: $\sup_{S \in \mathcal{S}} \lambda_{\max}(\mathbf{\Sigma}_S) \leq 2\sigma^2_{\mathbf{X},\mathcal{S}}$, and $\sup_{S \in \mathcal{S}} \lambda_{\max}(\mathbf{\Xi}_S) \leq 2\widetilde{\sigma}^2_{\mathbf{Z},\mathcal{S}}$. This justifies the claimed choices for the constants $\lambda_{\sup,\mathcal{S}}$ and $\widetilde{\lambda}_{\sup,\mathcal{S}}$ in Assumption 5.1. □

Lastly, owing to the very definition of $\boldsymbol{\beta}_S$ in (5.1) and the estimating equation satisfied by $\boldsymbol{\beta}_S$ therein, we have: $\forall\ S \in \mathcal{S}$, $\mathbb{E}\{\widetilde{\mathbf{X}}_S(\widetilde{Y} - \widetilde{\mathbf{X}}_S^T \boldsymbol{\beta}_S)\} = \mathbf{0}$ and $\mathbb{E}\{(\widetilde{Y} - \widetilde{\mathbf{X}}_S^T \boldsymbol{\beta}_S)(\widetilde{\mathbf{X}}_S^T \boldsymbol{\beta}_S)\} = 0$, so that $\mathbb{E}(\widetilde{Y}^2) = \mathbb{E}\{(\widetilde{Y} - \widetilde{\mathbf{X}}_S^T \boldsymbol{\beta}_S)^2\} + \mathbb{E}\{(\widetilde{\mathbf{X}}_S^T \boldsymbol{\beta}_S)^2\}$ and hence, for each $S \in \mathcal{S}$,

$$\mathbb{E}(\widetilde{Y}^2) \ \equiv \ \mathrm{Var}(Y) \ \geq \ \mathbb{E}\{(\widetilde{\mathbf{X}}_S^T \boldsymbol{\beta}_S)^2\} \ \equiv \ \boldsymbol{\beta}_S^T \mathbf{\Sigma}_S \boldsymbol{\beta}_S \ \geq \ \|\boldsymbol{\beta}_S\|_2^2 \, \lambda_{\min}(\mathbf{\Sigma}_S).$$

Using (5.4), we therefore have: $\sup_{S \in \mathcal{S}} \|\boldsymbol{\beta}_S\|_2^2 \leq \lambda^{-1}_{\inf,\mathcal{S}} \mathrm{Var}(Y)$. Further, due to Lemma C.1 (iii), $\mathrm{Var}(Y) \leq 2\sigma_Y^2$ and thus $\sup_{S \in \mathcal{S}} \|\boldsymbol{\beta}_S\|_2^2 \leq \lambda^{-1}_{\inf,\mathcal{S}} \mathrm{Var}(Y) \leq 2\lambda^{-1}_{\inf,\mathcal{S}} \sigma_Y^2$. This establishes the final claim in Lemma B.2. The proof of Lemma B.2 is now complete. □

B.6. Proof of Lemma B.3. For for any $c > 0$ and any constant $c^* > 0$ satisfying (B.17), let us define the events:

(B.37)
$$\mathcal{A}_{n,S}(c) \ := \ \left\{ \sup_{S \in \mathcal{S}} \left\| \widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}_S \right\|_2 > \ (c+1) K_{\mathcal{S}} \left(\sqrt{\frac{r_n}{n}} + \frac{33 r_n}{n} \right) + \frac{\lambda_{\sup,\mathcal{S}}}{n} \right\},$$
$$\mathcal{A}_{n,S}(c^*) := \ \left\{ \sup_{S \in \mathcal{S}} \left\| \widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}_S \right\|_2 > \ (c^*+1) K_{\mathcal{S}} \left(\sqrt{\frac{r_n}{n}} + \frac{33 r_n}{n} \right) + \frac{\lambda_{\sup,\mathcal{S}}}{n} \right\},$$

$$\mathcal{B}_{n,S}(c) := \left\{ \sup_{S \in \mathcal{S}} \left\| \widehat{\mathbf{\Sigma}}_S^{-1} - \mathbf{\Sigma}_S^{-1} \right\|_2 > (c+1) K_{\mathcal{S}}^* \left(\sqrt{\frac{r_n}{n}} + \frac{33 r_n}{n} \right) + \frac{2}{n} \frac{\lambda_{\sup,\mathcal{S}}}{\lambda^2_{\inf,\mathcal{S}}} \right\}$$

and let $\mathcal{A}^c_{n,S}(c^*)$ denote the complement event of $\mathcal{A}_{n,S}(c^*)$. Then, for any $c > 0$ and for any $c^* > 0$ satisfying (B.17), we first note that

(B.38) $\quad \mathbb{P}\{\mathcal{A}_{n,S}(c)\} \ \leq \ 4\exp(-c r_n), \ \text{ and } \ \mathbb{P}\{\mathcal{A}_{n,S}(c^*)\} \ \leq \ 4\exp(-c^* r_n),$

where both bounds are direct consequences of Lemma B.2 (iii), which applies under Assumption 5.1 (i). Further, for the events defined in (B.37), the following inclusions hold:

$$\begin{aligned}\mathcal{B}_{n,S}(c) \cap \mathcal{A}^c_{n,S}(c^*) \ &\subseteq_{(a)} \ \mathcal{B}_{n,S}(c) \cap \left\{ \sup_{S \in \mathcal{S}} \left|\left| \widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}_S \right|\right|_2 \ \leq \ \frac{\lambda_{\inf,\mathcal{S}}}{2} \right\} \\ &\subseteq_{(b)} \ \mathcal{A}_{n,S}(c) \cap \left\{ \sup_{S \in \mathcal{S}} \left|\left| \widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}_S \right|\right|_2 \ \leq \ \frac{\lambda_{\inf,\mathcal{S}}}{2} \right\} \quad \subseteq \ \mathcal{A}_{n,S}(c), \end{aligned} \tag{B.39}$$

and hence, $\mathbb{P}\{\mathcal{B}_{n,S}(c) \cap \mathcal{A}^c_{n,S}(c^*)\} \leq \mathbb{P}\{\mathcal{A}_{n,S}(c)\}$. The inclusion (a) in (B.39) above follows since c^* satisfies the condition (B.17) in Lemma B.3, while the inclusion (b) follows from an application of Lemma C.5 and from noting the definitions of the constants $K^*_{\mathcal{S}}$ and $\lambda_{\inf,\mathcal{S}}$.

Hence, for any $c > 0$ and for any $c^* > 0$ satisfying (B.17), we then have:

$$\begin{aligned} &\mathbb{P}\left\{ \sup_{S \in \mathcal{S}} \left|\left| \widehat{\mathbf{\Sigma}}^{-1}_S - \mathbf{\Sigma}^{-1}_S \right|\right|_2 \ > \ (c+1)K^*_{\mathcal{S}} \left(\sqrt{\frac{r_n}{n}} + \frac{33 r_n}{n} \right) + \frac{2}{n} \frac{\lambda_{\sup,\mathcal{S}}}{\lambda^2_{\inf,\mathcal{S}}} \right\} \\ &\equiv \ \mathbb{P}\left\{\mathcal{B}_{n,S}(c)\right\} \ = \ \mathbb{P}\left\{\mathcal{B}_{n,S}(c) \cap \mathcal{A}^c_{n,S}(c^*)\right\} + \mathbb{P}\left\{\mathcal{B}_{n,S}(c) \cap \mathcal{A}_{n,S}(c^*)\right\} \\ &\leq_{(a)} \ \mathbb{P}\left\{\mathcal{A}_{n,S}(c)\right\} + \mathbb{P}\left\{\mathcal{A}_{n,S}(c^*)\right\} \\ &\leq_{(b)} \ 4\exp(-c r_n) + 4\exp(-c^* r_n) \ \equiv \ 4 a_n(c, c^*, r_n), \end{aligned}$$

where the inequalities (a) and (b) follow from using (B.39) and (B.38) respectively. This establishes the claim in Lemma B.3 and completes the proof. □

B.7. Proof of Theorem 6.1.

Proof of the first part of Theorem 6.1. First, note that

$$\text{aver}(\hat{\Theta}_{jp}(\hat{\mathcal{C}}'_0)) - \text{aver}(\Theta_{jp}(\mathcal{C}'_0)) - E_{n,jp}(\hat{\mathcal{C}}'_0, \mathcal{C}'_0) = \text{aver}(\hat{\Theta}_{jp}(\mathcal{C}'_0)) - \text{aver}(\Theta_{jp}(\mathcal{C}'_0)).$$

Next, let $\mathbf{X}_{S'_{j1}}, \ldots, \mathbf{X}_{S'_{jL_{\text{distinct,j}}}} \subseteq \mathbf{Adj}_{\mathcal{C}'_0}(X_j)$ be the distinct parent sets of X_j in $\text{MEC}(\mathcal{C}'_0)$ with multiplicities $m_{j1}, \ldots, m_{jL_{\text{distinct},j}}$ respectively, so that

$$\text{aver}(\hat{\Theta}_{jp}(\mathcal{C}'_0)) = \frac{1}{L} \sum_{r=1}^{L_{\text{distinct,j}}} m_r \ \hat{\beta}_{jp|\mathbf{X}_{S'_{jr}} \cup \{X_1\}} \quad \text{and}$$

$$\text{aver}(\Theta_{jp}(\mathcal{C}'_0)) = \frac{1}{L} \sum_{r=1}^{L_{\text{distinct,j}}} m_r \ \beta_{jp|\mathbf{X}_{S'_{jr}} \cup \{X_1\}}, \quad \text{where} \ \ L \ := \ \sum_r m_r.$$

These now enable us to apply the results obtained in Section 5.1. To see this, note that the assumptions of Theorem 5.1 for

$$Y = X_p \text{ and } \mathcal{S} = \{S_{jr} = (j, 1, S'_{jr}) : r \in \{1, \ldots, L_{\text{distinct},j}\}\}$$

follow from Assumptions 4.1, 4.2, 4.3 and 6.2 and Lemma B.1. Thus, it follows directly from the first part of Remark 5.2 with $\mathbf{a}^T_{S_{jr}} = (m_r/L, 0, \ldots, 0)$ that

$$\text{aver}(\hat{\Theta}_{jp}(\mathcal{C}'_0)) - \text{aver}(\Theta_{jp}(\mathcal{C}'_0)) = \frac{1}{n}\sum_{r=1}^{n} Z^{(r)}_{jp} + O_{\mathbb{P}}\left(\frac{q_j + \log(L_{\text{distinct},j})}{n}\right),$$

since

$$\sum_{r=1}^{L_{\text{distinct,j}}} ||\mathbf{a}_{S_{jr}}||_2 = \sum_{r=1}^{L_{\text{distinct,j}}} \frac{m_r}{L} = 1. \quad \square$$

Proof of the second part of Theorem 6.1. Recall that

$$\hat{\eta}(\hat{\mathcal{C}}'_0) - \eta(\mathcal{C}'_0) = \hat{\theta}_{ij} \ \text{aver}(\hat{\Theta}_{jp}(\hat{\mathcal{C}}'_0)) - \theta_{1j} \ \text{aver}(\Theta_{jp}(\mathcal{C}'_0)).$$

Therefore, it is straightforward to obtain the result by applying the identity $a_n b_n - ab = a(b_n - b) + b(a_n - a) + (a_n - a)(b_n - b)$ with $a_n = \hat{\theta}_{ij}$, $b_n = \text{aver}(\hat{\Theta}_{jp}(\hat{\mathcal{C}}'_0))$, $a = \theta_{1j}$, and $b = \text{aver}(\Theta_{jp}(\mathcal{C}'_0))$, and using the first part of Theorem 6.1 and the following well-known result from the asymptotic theory for simple linear regression

$$\hat{\theta}_{1j} - \theta_{1j} = \hat{\beta}_{1j} - \beta_{1j} = \frac{1}{n}\sum_{r=1}^{n} Z^{(r)}_{1j} \ + \ O_{\mathbb{P}}\left(\frac{1}{n}\right). \quad \square$$

B.8. Proofs of Corollary 6.1 and Corollary 6.2.

Proof of Corollary 6.1. Recall that

$$\begin{aligned} &\text{aver}(\hat{\Theta}_{jp}(\hat{\mathcal{C}}'_0)) - \text{aver}(\Theta_{jp}(\mathcal{C}'_0)) \\ &= E_{n,jp}(\hat{\mathcal{C}}'_0, \mathcal{C}'_0) \ + \ \text{aver}(\hat{\Theta}_{jp}(\mathcal{C}'_0)) - \text{aver}(\Theta_{jp}(\mathcal{C}'_0)). \end{aligned}$$

Since from the discussion before Theorem 6.1 (see Section 6) we have

$$E_{n,jp}(\hat{\mathcal{C}}'_0, \mathcal{C}'_0) = o_{\mathbb{P}}(1/\sqrt{n}),$$

it is sufficient to show $\mathbb{E}[Z^2_{jp}] = \Omega(1)$ and

$$\text{(B.40)} \qquad \frac{\sqrt{n}\left\{\text{aver}(\hat{\Theta}_{jp}(\mathcal{C}'_0)) - \text{aver}(\Theta_{jp}(\mathcal{C}'_0))\right\}}{\sqrt{\mathbb{E}[Z^2_{jp}]}} \ \xrightarrow{d} \ \mathcal{N}(0, 1).$$

As discussed in the proof of Theorem 6.1, $\text{aver}(\hat{\Theta}_{jp}(\mathcal{C}_0')) - \text{aver}(\Theta_{jp}(\mathcal{C}_0'))$ can be written as $\sum_{S \in \mathcal{S}} \mathbf{a}_S^T(\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S)$ for some set $\{\mathbf{a}_S : S \in \mathcal{S}\}$ satisfying $\sum_S ||\mathbf{a}_S||_2 = O(1)$, where $\boldsymbol{\beta}_S$ denotes the vector of regression coefficients in the regression of $Y := X_p$ on $\mathbf{X}_S$ and and $\hat{\boldsymbol{\beta}}_S$ denotes its sample version (i.e. the corresponding OLS estimator) respectively, and $\mathcal{S} = \{S_{jr} = (j, 1, S'_{jr}) : r \in \{1, \ldots, L_{\text{distinct},j}\}\}$ is as in the proof of Theorem 6.1.

Therefore, given the stronger sparsity assumption

$$n^{-1/2}\{q_j + \log(L_{\text{distinct},j})\} \longrightarrow 0,$$

(B.40) follows from the second part of Remark 5.2 as long as the second moment of the influence function in the asymptotic linear expansion of $\text{aver}(\hat{\Theta}_{jp}(\mathcal{C}_0')) - \text{aver}(\Theta_{jp}(\mathcal{C}_0'))$ is bounded below, i.e. $\mathbb{E}[Z_{jp}^2] = \Omega(1)$.

We prove $\mathbb{E}[Z_{jp}^2] = \Omega(1)$ by verifying the sufficient conditions given in the last paragraph of Section 5.1: there exist constants $c_1 > 0$ and $c_2 > 0$ such that

(B.41)

$$\text{Var}(Y \mid \cup_{S \in \mathcal{S}} \mathbf{X}_S) > c_1 \text{ and } \mathbb{E}\left[\left(\sum_{S \in \mathcal{S}} \mathbf{a}_S^T \left(\Sigma_S\right)^{-1} \left(\mathbf{X}_S - \boldsymbol{\mu}_S\right)\right)^2\right] > c_2.$$

The first part of Assumption 6.4 and the first part of (B.41) with $c_1 = v$ are identical, since $\cup_{S \in \mathcal{S}} \mathbf{X}_S = \mathbf{Adj}_{\mathcal{C}_0'}(X_j) \cup \{X_1, X_j\}$. Thus we complete the proof by showing that the second part of (B.41) follows from Assumptions 6.2 and 6.4.

To this end, following the notation in the proof of Theorem 6.1, we write

$$\begin{aligned} &\sum_{S \in \mathcal{S}} \mathbf{a}_S^T \Sigma_S^{-1} (\mathbf{X}_S - \boldsymbol{\mu}_S) \\ = &\sum_{r=1}^{L_{\text{distinct},j}} \frac{m_r}{L}\, e^T_{1,|S'_{jr}|+2} \left((\Sigma_0)_{(j,1,S'_{jr})}\right)^{-1} \left(\mathbf{X}_{(j,1,S'_{jr})} - \boldsymbol{\mu}_{(j,1,S'_{jr})}\right), \end{aligned}$$

where $e^T_{1,|S'_{jr}|+2}$ denote the first row of an $(|S'_{jr}| + 2) \times (|S'_{jr}| + 2)$ identity matrix.

By partitioning $(\Sigma_0)_{(j,1,S'_{jr})}$ as

$$(\Sigma_0)_{(j,1,S'_{jr})} = \begin{pmatrix} (\Sigma_0)_j & (\Sigma_0)_{j(1,S'_{jr})} \\ (\Sigma_0)_{(1,S'_{jr})j} & (\Sigma_0)_{(1,S'_{jr})} \end{pmatrix}$$

and applying the well-known formula for the inverse of a partitioned matrix, we obtain

$$
\begin{aligned}
& e^T_{1,|S'_{jr}|+2} \left((\Sigma_0)_{(j,1,S'_{jr})} \right)^{-1} \left(\mathbf{X}_{(j,1,S'_{jr})} - \boldsymbol{\mu}_{(j,1,S'_{jr})} \right) \\
= & \frac{(X_j - \mu_j) - (\Sigma_0)_{j(1,S'_{jr})} \left((\Sigma_0)_{(1,S'_{jr})} \right)^{-1} \left(\mathbf{X}_{(1,S'_{rj})} - \boldsymbol{\mu}_{(1,S'_{rj})} \right)}{(\Sigma_0)_j - (\Sigma_0)_{j(1,S'_{jr})} \left((\Sigma_0)_{(1,S'_{jr})} \right)^{-1} (\Sigma_0)_{(1,S'_{jr})j}}.
\end{aligned}
$$

We define

$$
\begin{aligned}
& \boldsymbol{\beta}^T_{j|(1,S'_{jr})} := (\Sigma_0)_{j(1,S'_{jr})} \left((\Sigma_0)_{(1,S'_{jr})} \right)^{-1} \text{ and} \\
& \sigma^2_{j|(1,S'_{jr})} := (\Sigma_0)_j - (\Sigma_0)_{j(1,S'_{jr})} \left((\Sigma_0)_{(1,S'_{jr})} \right)^{-1} (\Sigma_0)_{(1,S'_{jr})j}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
& \mathbb{E} \left[\left(\sum_{S \in \mathcal{S}} \mathbf{a}^T_S \, (\Sigma_S)^{-1} \, (\mathbf{X}_S - \boldsymbol{\mu}_S) \right)^2 \right] \\
= \; & \mathbb{E} \left[\left(\sum_{r=1}^{L_{\text{distinct},j}} \frac{m_r}{L} \; e^T_{1,|S'_{jr}|+2} \left((\Sigma_0)_{(j,1,S'_{jr})} \right)^{-1} \left(\mathbf{X}_{(j,1,S'_{jr})} - \boldsymbol{\mu}_{(j,1,S'_{jr})} \right) \right)^2 \right] \\
= \; & \mathbb{E} \left[\left(\sum_{r=1}^{L_{\text{distinct},j}} \frac{m_r}{L \; \sigma^2_{j|(1,S'_{jr})}} \; \left((X_j - \mu_j) - \boldsymbol{\beta}^T_{j|(1,S'_{jr})} (\mathbf{X}_{(1,S'_{rj})} - \boldsymbol{\mu}_{(1,S'_{rj})}) \right) \right)^2 \right] \\
= \; & \operatorname{Var} \left(\sum_{r=1}^{L_{\text{distinct},j}} \frac{m_r \; \left(X_j - \boldsymbol{\beta}^T_{j|(1,S'_{jr})} \mathbf{X}_{(1,S'_{rj})} \right)}{L \; \sigma^2_{j|(1,S'_{jr})}} \right) \\
\geq \; & \mathbb{E} \left[\operatorname{Var} \left(\left\{ \sum_{r=1}^{L_{\text{distinct},j}} \frac{m_r \; \left(X_j - \boldsymbol{\beta}^T_{j|(1,S'_{jr})} \mathbf{X}_{(1,S'_{rj})} \right)}{L \; \sigma^2_{j|(1,S'_{jr})}} \right\} \mid \mathbf{Adj}_{\mathcal{C}'_0}(X_j) \cup \{X_1\} \right) \right] \\
= \; & \mathbb{E} \left[\operatorname{Var} \left(\left\{ \sum_{r=1}^{L_{\text{distinct},j}} \frac{m_r X_j}{L \; \sigma^2_{j|(1,S'_{jr})}} \right\} \mid \mathbf{Adj}_{\mathcal{C}'_0}(X_j) \cup \{X_1\} \right) \right] \\
= \; & \left\{ \sum_{r=1}^{L_{\text{distinct},j}} \frac{m_r}{L \; \sigma^2_{j|(1,S'_{jr})}} \right\}^2 \mathbb{E} \left[\operatorname{Var} \left(X_j \mid \mathbf{Adj}_{\mathcal{C}'_0}(X_j) \cup \{X_1\} \right) \right],
\end{aligned}
$$

where the second last equality follows from the fact that $\mathbf{X}_{S'_{jr}} \subseteq \mathbf{Adj}_{\mathcal{C}'_0}(X_j)$ for all $r \in \{1, \ldots, L_{\text{distinct},j}\}$.

Now since $\sum_r m_r = L$ and $\sigma^2_{j|(1,S'_{jr})} \leq (\Sigma_0)_j$ for all $r \in \{1, \ldots, L_{\text{distinct},j}\}$, by Assumption 6.2, we have

$$\left\{ \sum_{r=1}^{L_{\text{distinct},j}} \frac{m_r}{L \ \sigma^2_{j|(1,S'_{jr})}} \right\}^2 \geq \frac{1}{C_3^2}.$$

Finally, by Assumption 6.4, we have

$$\mathbb{E}\left[\text{Var}\left(X_j \mid \mathbf{Adj}_{\mathcal{C}'_0}(X_j) \cup \{X_1\} \right) \right] \geq v.$$

This completes the proof. □

Proof of Corollary 6.2. It is easy to see that Theorem 6.1, Corollary 6.1 and Assumption 6.3 imply

$$\begin{aligned} \hat{\eta}(\hat{\mathcal{C}}'_0) - \eta(\mathcal{C}'_0) \ &= \ \hat{\theta}_{1j} \ E_{n,jp}(\hat{\mathcal{C}}'_0, \mathcal{C}'_0) \ + \\ &+ \ \frac{1}{n} \sum_{r=1}^{n} \left\{ \theta_{1j} Z_{jp}^{(r)} \ + \ \text{aver}(\Theta_{jp}(\mathcal{C}'_0)) Z_{1j}^{(r)} \right\} \ + \ o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}} \right). \end{aligned}$$

Thus the result follows from the same arguments given in the proof of Corollary 6.1. □

APPENDIX C: TECHNICAL TOOLS - SUPPORTING LEMMAS

In this section, we collect some key technical lemmas that would be useful throughout in the proofs of all our main theoretical results.

C.1. Properties of Orlicz Norms and Concentration Bounds. We next enlist, through a sequence of lemmas, some useful general properties of Orlicz norms, as well as a few specific ones for sub-Gaussians and sub-exponentials. These are all quite well known and routinely used. Their statements (possibly with slightly different constants) and proofs can be found in several relevant references, including Van der Vaart and Wellner [1996]; Pollard [2015]; Vershynin [2012, 2018]; Wainwright [2017] and Rigollet and Hütter [2017], among others. The proofs are thus skipped for brevity.

Lemma C.1 (General properties of Orlicz norms, sub-Gaussians and sub–exponentials). *In the following, $X, Y \in \mathbb{R}$ denote generic random variables and μ denotes $\mathbb{E}(X) \in \mathbb{R}$.*

(i) (Basic properties). *For* $\alpha \geq 1$, $\|\cdot\|_{\psi_\alpha}$ *is a norm (and a quasinorm if* $\alpha < 1$*) satisfying: (a)* $\|X\|_{\psi_\alpha} \geq 0$ *and* $\|X\|_{\psi_\alpha} = 0 \Leftrightarrow X = 0$ *almost surely (a.s.), (b)* $\|cX\|_{\psi_\alpha} = |c| \, \|X\|_{\psi_2}$ $\forall\, c \in \mathbb{R}$ *and* $\||X|\|_{\psi_\alpha} = \|X\|_{\psi_\alpha}$, *and (c)* $\|X+Y\|_{\psi_\alpha} \leq \|X\|_{\psi_\alpha} + \|Y\|_{\psi_\alpha}$.

(ii) (Tail bounds and equivalences). *(a) If* $\|X\|_{\psi_\alpha} \leq \sigma$ *for some* $(\alpha, \sigma) > 0$*, then* $\forall\ \epsilon \geq 0$, $\mathbb{P}(|X| > \epsilon) \leq 2\exp(-\epsilon^\alpha/\sigma^\alpha)$. *(b) Conversely, if* $\mathbb{P}(|X| > \epsilon) \leq C\exp(-\epsilon^\alpha/\sigma^\alpha)$ $\forall\ \epsilon \geq 0$*, for some* $(C, \sigma, \alpha) > 0$*, then* $\|X\|_{\psi_\alpha} \leq \sigma(1 + C/2)^{1/\alpha}$.

(iii) (Moment bounds). *If* $\|X\|_{\psi_\alpha} \leq \sigma$ *for some* $(\alpha, \sigma) > 0$*, then* $\mathbb{E}(|X|^m) \leq C_\alpha^m \sigma^m m^{m/\alpha}$ $\forall\ m \geq 1$*, for some constant* C_α *depending only on* α*. (A converse of this result also holds, although not explicitly presented here). For* $\alpha = 1$ *and* 2 *in particular, we have:*

(a) If $\|X\|_{\psi_1} \leq \sigma$*, then for each* $m \geq 1$, $\mathbb{E}(|X|^m) \leq \sigma^m m! \leq \sigma^m m^m$.

(b) If $\|X\|_{\psi_2} \leq \sigma$*, then* $\mathbb{E}(|X|^m) \leq 2\sigma^m \Gamma(m/2 + 1)$ *for each* $m \geq 1$*, where* $\Gamma(a) := \int_0^\infty x^{a-1} exp(-x)dx$ $\forall\ a > 0$ *denotes the Gamma function. Hence,* $\mathbb{E}(|X|) \leq \sigma\sqrt{\pi}$ *and* $\mathbb{E}(|X|^m) \leq 2\sigma^m (m/2)^{m/2}$ *for any* $m \geq 2$.

(iv) (Hölder-type inequality for the Orlicz norm of products). *For any* $\alpha, \beta > 0$*, let* $\gamma := (\alpha^{-1} + \beta^{-1})^{-1}$*. Then, for any two random variables* X *and* Y *with* $\|X\|_{\psi_\alpha} < \infty$ *and* $\|Y\|_{\psi_\beta} < \infty$, $\|XY\|_{\psi_\gamma} < \infty$ *and* $\|XY\|_{\psi_\gamma} \leq \|X\|_{\psi_\alpha} \|Y\|_{\psi_\beta}$*. In particular, for any two sub-Gaussians* X *and* Y, XY *is sub-exponential and* $\|XY\|_{\psi_1} \leq \|X\|_{\psi_2} \|Y\|_{\psi_2}$*. Moreover, if* $Y \leq M$ *a.s. and* $\|X\|_{\psi_\alpha} < \infty$*, then* $\|XY\|_{\psi_\alpha} \leq M \|X\|_{\psi_\alpha}$.

(v) (MGF related properties of sub-Gaussians). *Let* $\mathbb{E}[\exp\{t(X-\mu)\}]$ *denote the moment generating function (MGF) of* $X - \mu$ *at* $t \in \mathbb{R}$*. Then:*

(a) If $\|X - \mu\|_{\psi_2} < \sigma$ *for some* $\sigma > 0$*, then* $\mathbb{E}[\exp\{t(X-\mu)\}] \leq \exp(2\sigma^2 t^2)$ $\forall\, t \in \mathbb{R}$.

(b) Conversely, if $\mathbb{E}[\exp\{t(X-\mu)\}] \leq \exp(\sigma^2 t^2)$ $\forall\ t \in \mathbb{R}$ *for some* $\sigma \geq 0$*, then for any* $\epsilon \geq 0$, $\mathbb{P}(|X - \mu| > \epsilon) \leq 2\exp(-\epsilon^2/4\sigma^2)$ *and hence,* $\|X - \mu\|_{\psi_2} \leq 2\sqrt{2}\sigma$.

Lemma C.2 (Properties of sub-Gaussian random vectors). *Let* $\mathbf{X} \in \mathbb{R}^d$ $(d \geq 1)$ *be any random vector, and let* $\mathbf{v} \in \mathbb{R}^d$, $\mathbf{M} \in \mathbb{R}^{d \times d}$ *be generic (fixed) vectors and matrices. Then,*

(i) For any $\mathbf{v} \in \mathbb{R}^d$, $\left\|\mathbf{v}^T\mathbf{X}\right\|_{\psi_2} \leq \|\mathbf{v}\|_2 \|\mathbf{X}\|_{\psi_2}^*$ *and* $\left\|\mathbf{v}^T\mathbf{X}\right\|_{\psi_2} \leq \|\mathbf{v}\|_1 \|\mathbf{X}\|_{\psi_2} \leq \sqrt{d}\, \|\mathbf{v}\|_2 \|\mathbf{X}\|_{\psi_2}$*. Hence,* $\|\mathbf{X}\|_{\psi_2} \leq \|\mathbf{X}\|_{\psi_2}^* \leq \sqrt{d}\, \|\mathbf{X}\|_{\psi_2}$*. Further, for*

any $\mathbf{M} \in \mathbb{R}^{d\times d}$, $\|\mathbf{M}\mathbf{X}\|_{\psi_2} \leq ||\mathbf{M}||_\infty \|\mathbf{X}\|_{\psi_2} \leq \sqrt{d}\,||\mathbf{M}||_2 \|\mathbf{X}\|_{\psi_2}$ *and* $\|\mathbf{M}\mathbf{X}\|^*_{\psi_2} \leq ||\mathbf{M}||_2 \|\mathbf{X}\|^*_{\psi_2} \leq \sqrt{d}\,||\mathbf{M}||_2 \|\mathbf{X}\|_{\psi_2}$.

(ii) Suppose $\mathbb{E}(\mathbf{X}) = \mathbf{0}$, $\|\mathbf{X}\|_{\psi_2} \leq \sigma$ *and assume further that the coordinates* $\{\mathbf{X}_{[j]}\}_{j=1}^d$ *of* $\mathbf{X}$ *are independent. Then for any* $\mathbf{v} \in \mathbb{R}^d$, $\left\|\mathbf{v}^T\mathbf{X}\right\|_{\psi_2} \leq 2\sqrt{2}\sigma \|\mathbf{v}\|_2$. *Thus, under these additional assumptions on* $\mathbf{X}$, *it holds that* $\|\mathbf{X}\|_{\psi_2} \leq \|\mathbf{X}\|^*_{\psi_2} \leq 2\sqrt{2}\,\|\mathbf{X}\|_{\psi_2}$. *Further, for any* $\mathbf{M} \in \mathbb{R}^{d\times d}$, $\|\mathbf{M}\mathbf{X}\|_{\psi_2} \leq \|\mathbf{M}\mathbf{X}\|^*_{\psi_2} \leq ||\mathbf{M}||_2 \|\mathbf{X}\|^*_{\psi_2} \leq 2\sqrt{2}\,||\mathbf{M}||_2 \|\mathbf{X}\|_{\psi_2}$.

Lemma C.3 (Concentration bounds for sums of independent sub-Gaussian variables). *For any* $n \geq 1$, *let* $\{X_i\}_{i=1}^n$ *be independent (not necessarily i.i.d.) random variables with means* $\{\mu_i\}_{i=1}^n$ *and* $\max_{1\leq i\leq n} \|X_i - \mu_i\|_{\psi_2} \leq \sigma$ *for some constant* $\sigma \geq 0$. *Then, for any collection of real numbers* $\{a_i\}_{i=1}^n$ *and letting* $\mathbf{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n$, *we have:*

$$\mathbb{E}\left[\exp\left\{t\sum_{i=1}^n a_i(X_i - \mu_i)\right\}\right] \leq \exp\left(2\sigma^2 t^2 \|\mathbf{a}\|_2^2\right) \quad \forall\, t \in \mathbb{R}, \quad \textit{and}$$

$$\mathbb{P}\left\{\left|\sum_{i=1}^n a_i(X_i - \mu_i)\right| > \epsilon\right\} \leq 2\exp\left\{-\epsilon^2/\left(8\sigma^2\|\mathbf{a}\|_2^2\right)\right\} \quad \forall\, \epsilon \geq 0.$$

In particular, when $a_i = 1/n$, *we have:* $\left\|\frac{1}{n}\sum_{i=1}^n (X_i - \mu_i)\right\|_{\psi_2} \leq (4\sigma)/\sqrt{n}$ *and* $\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n (X_i - \mu_i)\right| > \epsilon\right\} \leq 2\exp\left\{-n\epsilon^2/\left(8\sigma^2\right)\right\} \forall\, \epsilon \geq 0$.

C.2. Basic Matrix Inequalities and Deviation Bounds for Random Matrices under the Spectral Norm. We provide here a sequence of lemmas collecting some useful and fairly well known inequalities regarding matrix norms and spectral properties of matrices and their submatrices and inverses. The lemmas also include some important results such as deterministic inequalities relating spectral distance between inverses of two p.d. matrices to that between the original matrices, as well as exact concentration bounds for deviations (under the spectral norm) of covariance-type random matrices defined by sub-gaussian random vectors.

Lemma C.4 (Basic inequalities on matrix norms and spectral properties of submatrices). *Let* $\mathbf{M} \in \mathbb{R}^{d\times d}$ $(d \geq 1)$ *denote any generic square matrix. Then,* $||\mathbf{M}||_\infty \leq \sqrt{d}\,||\mathbf{M}||_2 \leq d\,||\mathbf{M}||_\infty$ *and* $||\mathbf{M}||_{\max} \leq ||\mathbf{M}||_2 \leq d\,||\mathbf{M}||_{\max}$. *Further, the following results hold.*

(i) $\|\mathbf{M}\mathbf{v}\|_2 \leq ||\mathbf{M}||_2 \|\mathbf{v}\|_2$ *and* $\|\mathbf{M}\mathbf{v}\|_\infty \leq ||\mathbf{M}||_\infty \|\mathbf{v}\|_\infty$ *for any* $\mathbf{v} \in \mathbb{R}^d$. *Further, for any* $\mathbf{M}_1, \mathbf{M}_2 \in \mathbb{R}^{d\times d}$, $||\mathbf{M}_1\mathbf{M}_2||_2 \leq ||\mathbf{M}_1||_2 ||\mathbf{M}_2||_2$ *and* $||\mathbf{M}_1\mathbf{M}_2||_\infty \leq ||\mathbf{M}_1||_\infty ||\mathbf{M}_2||_\infty$.

(ii) *Let* $\mathbf{M}$ *be symmetric and let* $\mathbf{M}_k$ *denote any principal submatrix of* $\mathbf{M}$ *of order* $k \leq d$. *Let* $\lambda_1 \geq \ldots \geq \lambda_d$ *and* $\mu_1 \geq \ldots \geq \mu_k$ *respectively denote the ordered eigenvalues of* $\mathbf{M}$ *and* $\mathbf{M}_k$. *Then, these are 'interlaced' as:* $\lambda_{d-k+1} \leq \mu_i \leq \lambda_i \ \forall \ 1 \leq i \leq k$.

(iii) *For any* $\mathbf{M} \in \mathbb{R}^{d \times d}$ *(not necessarily symmetric) and any square submatrix (not necessarily principal)* $\mathbf{M}_k$ *of* $\mathbf{M}$ *of order* $k \leq d$, *let* $\lambda_1 \geq \ldots \geq \lambda_d$ *and* $\mu_1 \geq \ldots \geq \mu_k$ *respectively denote the ordered singular values of* $\mathbf{M}$ *and* $\mathbf{M}_k$. *Then, we have the 'upper' and 'lower' interlacing(s):* $\mu_i \leq \lambda_i \ \forall \ 1 \leq i \leq k$, *and* $\mu_i \geq \lambda_{2d-2k+i} \ \forall \ 1 \leq i \leq (2k-d)$.

A few remarks regarding Lemma C.4 (ii)-(iii) are in order. The interlacing inequalities in (ii) are special cases of the well known Poincare Separation Theorem (and more generally, the Courant-Fisher Min-Max Theorem). The particular case of $k = d-1$ is also known as the Cauchy Interlacing Theorem (see Thompson [1972] for further details). Note that these inequalities are only for the eigenvalues (not singular values) of symmetric matrices and their principal submatrices (for n.n.d. matrices however, these two coincide). The inequalities in (iii) are adopted from Thompson [1972] (they also apply more generally to non-square matrices). Notably, they apply directly to singular values (not eigenvalues) of matrices and submatrices of arbitrary nature and order. Among other implications, they also establish that $||\mathbf{M}^*||_2 \leq ||\mathbf{M}||_2$ for arbitrary matrices $\mathbf{M}$ and submatrices $\mathbf{M}^*$ of $\mathbf{M}$.

Lemma C.5 (Inequalities relating spectral deviations of p.d. matrices and their inverses). *Let* $\mathbf{M}_0 \in \mathbb{R}^{d \times d}$ *be any symmetric positive definite matrix with inverse* $\mathbf{M}_0^{-1}$ *and minimal eigenvalue (also singular value)* $\lambda_{\min}(\mathbf{M}_0) \equiv \left|\left|\mathbf{M}_0^{-1}\right|\right|_2^{-1} > 0$. *Let* $\mathbf{M} \in \mathbb{R}^{d \times d}$ *be any matrix such that* $||\mathbf{M} - \mathbf{M}_0||_2 \leq \lambda_{\min}(\mathbf{M}_0)$. *Then,* $\left|\left|(\mathbf{M} - \mathbf{M}_0)\mathbf{M}_0^{-1}\right|\right|_2 < 1$, *and* $\{I + (\mathbf{M} - \mathbf{M}_0)\mathbf{M}_0^{-1}\}$ *and* $\mathbf{M}$ *are both invertible. Further,*

$$\begin{aligned}\left|\left|\mathbf{M}^{-1} - \mathbf{M}_0^{-1}\right|\right|_2 \ &\leq \ \frac{\lambda_{\min}^{-2}(\mathbf{M}_0)}{1 - ||\mathbf{M} - \mathbf{M}_0||_2 \, \lambda_{\min}^{-1}(\mathbf{M}_0)} \, ||\mathbf{M} - \mathbf{M}_0||_2 \\ &\leq \ 2\lambda_{\min}^{-2}(\mathbf{M}_0) \, ||\mathbf{M} - \mathbf{M}_0||_2 \quad \text{if} \ \ ||\mathbf{M} - \mathbf{M}_0||_2 \ \leq \ \frac{1}{2}\lambda_{\min}(\mathbf{M}_0).\end{aligned}$$

Lemma C.6 (Deviation bounds under the spectral norm for covariance–type matrices). *Let* $\mathbf{X} \in \mathbb{R}^d$ *be any random vector with* $\mathbb{E}(\mathbf{X}) = \mathbf{0}$ *and* $\|\mathbf{X}\|_{\psi_2}^* \leq \sigma_*$ *for some* $\sigma_* \geq 0$. *Let* $\mathbf{\Sigma} := \mathbb{E}(\mathbf{X}\mathbf{X}^T)$ *which is assumed to be positive definite with minimum and maximum eigenvalues* $\lambda_{\min}(\mathbf{\Sigma}) > 0$ *and* $\lambda_{\max}(\mathbf{\Sigma}) \equiv ||\mathbf{\Sigma}||_2 \geq \lambda_{\min}(\mathbf{\Sigma}) > 0$ *respectively. Consider a collection* $\{\mathbf{X}_i\}_{i=1}^n$

of $n \geq 1$ independent realizations of $\mathbf{X}$. Then, for any $\epsilon \geq 0$, we have:

$$\mathbb{P}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^T - \mathbf{\Sigma}\right\|_2 > C_1 K_{\mathbf{X}}^2\left(\sqrt{\frac{d+\epsilon}{n}} + \frac{d+\epsilon}{n}\right)\right\} \leq 2\exp(-\epsilon)$$

$$\text{and } \mathbb{E}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^T - \mathbf{\Sigma}\right\|_2\right) \leq C_2 K_{\mathbf{X}}^2\left(\sqrt{\frac{d}{n}} + \frac{d}{n}\right), \quad \text{where}$$

$K_{\mathbf{X}}^2 := \frac{\sigma_*^2 \lambda_{\max}(\mathbf{\Sigma})}{\lambda_{\min}(\mathbf{\Sigma})}$, and $C_1, C_2 > 0$ are absolute constants that do not depend on any other quantities introduced above. Specifically, choosing $\epsilon = cd$ for any $c > 0$ and noting that $\sqrt{c+1} \leq c+1$, we have: for any $c > 0$,

$$\mathbb{P}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^T - \mathbf{\Sigma}\right\|_2 > C_1 K_{\mathbf{X}}^2 (c+1)\left(\sqrt{\frac{d}{n}} + \frac{d}{n}\right)\right\} \leq 2\exp(-cd).$$

Lemma C.5 is adopted from (the proof of) Lemma 5 in Harris and Drton [2013]. Lemma C.6 is obtained using Theorem 4.7.1 (more fundamentally, Theorem 4.6.1) of Vershynin [2018], in conjunction with Exercise 4.7.3 therein, along with appropriate modifications of his notations and assumptions to adapt to our setting. Simialr results, though slightly more involved and with less explicit constants, may also be obtained using Theorem 5.39 of Vershynin [2012], along with equation (5.26) in Remark 5.40 therein.

APPENDIX D: FALSE DISCOVERY RATE CONTROL

As an additional validation to our asymptotic results, we discuss some FDR control results for estimating $\cup_{t=1}^r S^{(t)} = \cup_{t=1}^r \{X_j^{(t)} : \eta^{(t)} \neq 0,\ j = 2, \ldots, p-1\}$ (`Target`) and $\cup_{t=1}^r S^{*(t)} = \cup_{t=1}^r \{X_j^{(t)} : \theta_{1j}^{(t)} \text{aver}(\Theta_{jp}^{(t)}) \neq 0,\ j = 2, \ldots, p-1\}$ (`Target_CPDAG`) when the true CPDAG is known as well as when the CPDAG is estimated. The BH procedure at level α (asymptotically) guarantees to control FDR at level $\alpha m_0/m$ for estimating `Target_CPDAG`, where for each simulation setting m_0 denotes the total number of true hypotheses $|\cup_{t=1}^r S^{*(t)}|$ among the $m = r*(p-2) = 5000$ hypotheses. Since $\cup_{t=1}^r S^{*(t)} \subseteq \cup_{t=1}^r S^{(t)}$, it is expected that the empirical FDR level would be higher when it is measured with respect to `Target`.

Figure 3 shows that the BH procedure becomes very conservative for estimating $\cup_{t=1}^r S^{*(t)}$ even though we ignore the additional adjustment suggested by Benjamini and Yekutieli [2001] in order to correct for possible dependencies among hypotheses. The conservativeness of BH procedure can be attributed to the fact that the p-value corresponding to the test $\theta_{1j}^{(t)} \text{aver}(\Theta_{jp}^{(t)}) = 0$ has a stochastically larger distribution than Uniform$[0,1]$

when both $\theta_{1j}^{(t)}$ and aver$(\Theta_{jp}^{(t)})$ are zero. In order to mitigate this issue, we apply a heuristic screening where we select the potential mediators for which the total effect of the treatment X_1 on the mediator is non-zero, by testing $\theta_{1j}^{(t)} = 0$ at the significance level 0.01. Then we apply the BH procedure on the selected set. Figure 4 shows that the heuristic screening method is indeed effective in reducing the conservativeness of BH procedure the empirical FDR of the estimated sets based on the BH procedure.

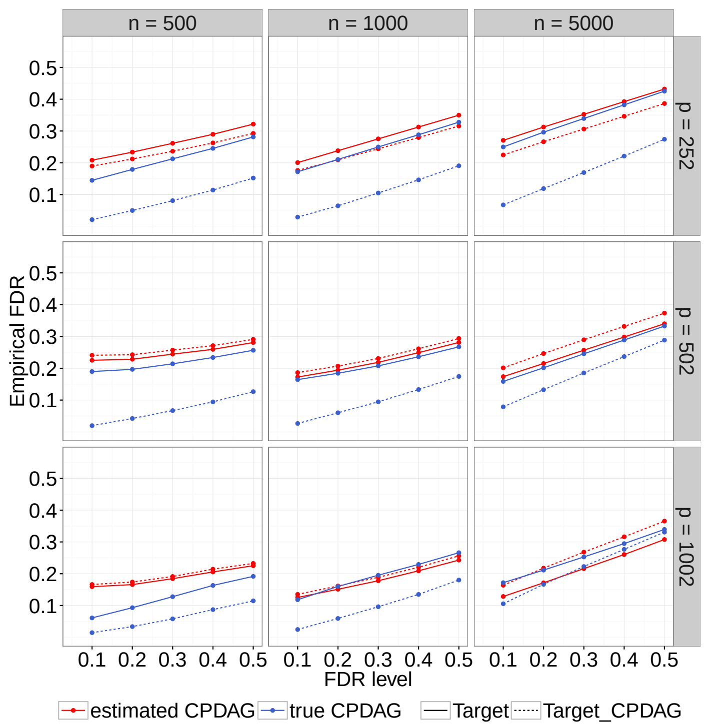

Fig 3: Empirical FDR of the estimated sets based on the BH procedure without p-value screening for estimating `Target` and `Target_CPDAG` when the true CPDAG is known as well as when the CPDAG is estimated

Abhishek Chakrabortty
Department of Statistics, The Wharton School
University of Pennsylvania
Philadelphia, PA 19104, USA.
E-mail: abhich@upenn.edu

Preetam Nandy
LinkedIn Corp.
Mountain View, CA 94043, USA.
E-mail: preetamnandy@gmail.com

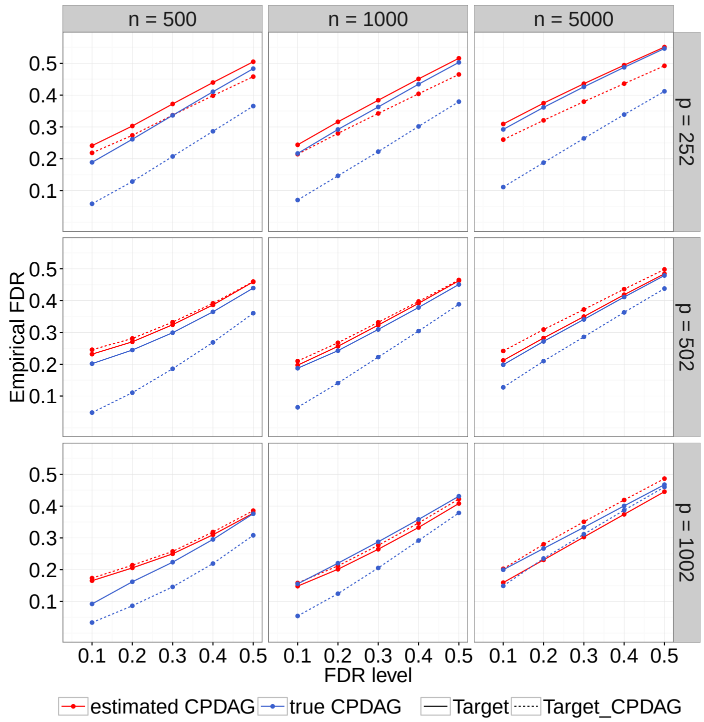

Fig 4: Empirical FDR of the estimated sets based on the BH procedure with p-value screening for estimating `Target` and `Target_CPDAG` when the true CPDAG is known as well as when the CPDAG is estimated.

Hongzhe Li
Department of Biostatistics, Epidemiology and Informatics (DBEI)
University of Pennsylvania
Philadelphia, PA 19104, USA.
E-mail: hongzhe@pennmedicine.upenn.edu