

Outlier-Robust Approach for High-dimensional Gaussian Linear Structural Equation Model

Gunwoong Park

GW.PARK23@GMAIL.COM

*Department of Statistics
Seoul National University
Seoul, 08826, South Korea*

Editor:

Abstract

This study focuses on learning a Gaussian linear structural equation model (SEM) in high-dimensional and corrupted sample settings. First, it defines the considered outlier and bad samples in Gaussian linear SEMs, and formalizes the corrupted Gaussian linear SEM. Subsequently, it develops a new outlier-robust algorithm that consists of two steps: (1) element-wise ordering estimation, and (2) its parent estimation, where both problems are effectively addressed using truncated conditional variances and ℓ_1 -regularized least trimmed squares. It is proven that, under appropriate conditions, the proposed method can successfully recover the graph, even when all observations are outliers on some nodes, by providing its breakdown point. Furthermore, it shows that the number of trimmed samples $h = \Omega((d + |B|) \log p)$ and the number of samples for truncated conditional variances $h' = \Omega(d^2 \log p)$ are sufficient for the proposed algorithm to learn corrupted Gaussian linear SEMs, where p is the number of nodes, d is the maximum degree of the moralized graph, and $|B|$ is the maximum number of bad samples. It is demonstrated through various simulated data that the proposed algorithm is statistically consistent to learn the model in high-dimensional and corrupted sample settings. In addition, the numerical experiments verify that the proposed algorithm performs better, when compared to the state-of-the-art HLSM, HGSM, TD, US, and GDS algorithms in noisy data settings.

Keywords: Bayesian networks, directed acyclic graph, ℓ_1 -regularization, least trimmed square, linear structural equation model, structure learning

1. Introduction

Gaussian linear structural equation models (SEMs) have been applied as an effective method to study the causal or conditional independence structure among variables in many fields, such as meteorology, epidemiology, finance, genetics, neuroscience, sports science, and many others (e.g., Lauritzen, 1996; Spirtes et al., 2000; Peters and Bühlmann, 2014; Eigenmann et al., 2017; Wang et al., 2020). Nevertheless, learning the model from purely observational data is limited owing to the identifiability issue.

Recent studies in Peters and Bühlmann (2014); Ghoshal and Honorio (2018); Park and Kim (2020); Park (2020) provide a fully identifiable class of Gaussian linear SEMs by placing restrictions on the model. Particularly, Peters and Bühlmann (2014) first shows that the models are identifiable if error variances are the same or known. Furthermore, Ghoshal and Honorio (2018); Park and Kim (2020) relax the equal error variance condition using

both error variances and edge weights. Hence, they prove that Gaussian linear SEMs with different error variances can be identifiable. The detailed identifiability conditions are explained in Section 2.2.

Many studies have developed learning strategies in both low- and high-dimensional sample settings for these identifiable classes of Gaussian linear SEMs (e.g., Ghoshal and Honorio, 2018; Chen et al., 2019; Park, 2020; Park and Kim, 2021; Park et al., 2021). For instance, in terms of learning Gaussian linear SEMs with heterogeneous error variances, Park and Kim (2020) provides a consistent conditional variance and an independence test-combined algorithm for low dimensional settings. In high-dimensional settings, Ghoshal and Honorio (2018) develops a constrained ℓ_1 -minimization for inverse covariance matrix estimation (CLIME)-based algorithms with sample bound $n = \Omega(d^4 \log p)$, in which d is the maximum degree of the (moralized) graph. Park and Kim (2021) develops a graphical lasso and a conditional independence test-combined approach with sample bound $n = \Omega(d^2 \log p)$. Lastly, Park et al. (2021) introduces a ℓ_1 -regularized regression-based algorithm in which the sample bound is $n = \Omega(d^2 \log p)$.

Therefore, the existing algorithms, including those for the model with homogeneous error variances, successfully recover the underlying structure of the Gaussian linear SEM under appropriate conditions, when the samples are independent and identically distributed. In other words, the desirable consistency of the algorithms would be compromised in the corrupted sample settings, where some outliers exist. This challenge motivates the development of an outlier-robust structure learning algorithm for Gaussian linear SEMs that can cope with observations deviating from the model assumptions.

The main objective of this study is to develop a new outlier-robust algorithm for learning high-dimensional Gaussian linear SEMs, which allows many corrupted observations. Hence, this study first defines the considered outliers and bad samples in the Gaussian linear SEM setting, and formalizes the corrupted Gaussian linear SEM. Subsequently, it develops a consistent ℓ_1 -regularized least trimmed squares (LTS)-based algorithm that consists of two steps: (1) element-wise ordering estimation; and (2) its parent estimation, where both problems are effectively addressed using truncated conditional variances and ℓ_1 -regularized LTS.

This study provides the breakdown point of the proposed algorithm, which shows that the proposed method can consistently learn the graph even when *all* observations are outliers on some nodes. It also proves that the number of trimmed samples $h = \Omega(d + |B| \log p)$ and number of samples for truncated conditional variances $h' = \Omega(d^2 \log p)$ are sufficient for the proposed algorithm to learn a corrupted Gaussian linear SEM, where p is the number of nodes, d is the maximum degree of the moralized graph, and $|B|$ is the maximum number of bad samples. Therefore, the proposed algorithm consistently learns the model in high-dimensional and corrupted sample settings.

The theoretical findings of this study are verified through various numerical experiments, where the proposed algorithm has a large breakdown point, and consistently recovers the underlying graphs from corrupted samples with sample complexity $h = \Omega(d + |B| \log p)$ and $h' = \Omega(d^2 \log p)$. Furthermore, the proposed algorithm is compared to the state-of-the-art HLSE (Park et al., 2021), HGSM (Park and Kim, 2021), TD (Chen et al., 2019), uncertainty scoring (US) (Park, 2020), and GDS (Peters and Bühlmann, 2014) algorithms in various corrupted sample settings.

The remainder of this paper is structured as follows. Section 2.1 summarizes the necessary notations and explains basic concepts of a Gaussian linear SEM. Section 2.2 discusses the identifiability conditions and the existing learning algorithms. Section 3 defines the considered outlier and bad samples in Gaussian linear SEM settings, and then formalizes the corrupted Gaussian linear SEM. Section 4 introduces the outlier-robust algorithm for a high-dimensional Gaussian linear SEM. Sections 4.1 and 4.2 provide the breakdown point and theoretical guarantees of the proposed algorithm, respectively. Section 5 evaluates the proposed algorithm and state-of-the-art algorithms in various simulation settings. Lastly, Section 6 offers a discussion, and suggests possibilities for further study.

2. Preliminaries

First this study introduces some necessary notations and definitions for Gaussian linear SEMs. Subsequently, it provides the identifiability conditions and the algorithms for learning the model from i.i.d. samples.

2.1 Gaussian Linear Structural Equation Models

A directed acyclic graph $G = (V, E)$ consists of a set of nodes $V = \{1, 2, \dots, p\}$ and a set of directed edges $E \subset V \times V$ with no directed cycles. A directed edge from node j to k is denoted by (j, k) or $j \rightarrow k$. The set of *parents* of node k , denoted by $\text{Pa}(k)$, consists of all nodes j such that $(j, k) \in E$. In addition, the set of *children*, denoted by $\text{Ch}(j)$, consists of all nodes k such that $(j, k) \in E$. The set of *neighbors* of node j , denoted by $\text{Ne}(j)$, consists of all nodes k connected by an edge. If there is a directed path $j \rightarrow \dots \rightarrow k$, then k is called a *descendant* of j , and j is called an *ancestor* of k . The sets $\text{De}(k)$ and $\text{An}(k)$ denote the set of all descendants and ancestors, respectively, of node k . An important property of DAGs is the existence of an (possibly non-unique) *ordering* $\pi = (\pi_1, \pi_2, \dots, \pi_p)$ for a directed graph that represents the directions of edges such that for every directed edge $(j, k) \in E$, j comes before k in the ordering. Hence, learning a graph is equivalent to inferring the ordering and its parents.

Consider a set of random variables $X := (X_j)_{j \in V}$ with a probability distribution taking values in a sample space \mathcal{X}_V over the nodes in G . For any subset S of V , let $X_S := \{X_j : j \in S \subset V\}$ and $\mathcal{X}_S := \times_{j \in S} \mathcal{X}_j$ where \mathcal{X}_j is a sample space of X_j . For any node $j \in V$, $\Pr(X_j \mid X_S)$ denotes the conditional distribution of a variable X_j given a random vector X_S . Then, a DAG model has the following factorization joint probability density function:

$$\Pr(G) = \Pr(X_1, X_2, \dots, X_p) = \prod_{j=1}^p \Pr(X_j \mid X_{\text{Pa}(j)}), \quad (1)$$

where $\Pr(X_j \mid X_{\text{Pa}(j)})$ is the conditional distribution of X_j given its parents variables $X_{\text{Pa}(j)} := \{X_k : k \in \text{Pa}(j) \subset V\}$.

A Gaussian linear SEM is a special case of DAG models where each variable is defined by the following linear structural equation: For any node $j \in V$,

$$X_j = \beta_{0,j} + \sum_{k \in \text{Pa}(j)} \beta_{k,j} X_k + \epsilon_j, \quad (2)$$

where $(\epsilon_j)_{j \in V}$ are independent Gaussian distributions, $N(0, \sigma_j^2)$.

It can be restated in the following matrix form:

$$(X_1, X_2, \dots, X_p)^T = B_0 + B(X_1, X_2, \dots, X_p)^T + (\epsilon_1, \epsilon_2, \dots, \epsilon_p)^T \quad (3)$$

where $B \in \mathbb{R}^{p \times p}$ is an edge weight matrix or a weighted adjacency matrix with (j, k) -th element $B_{jk} = \beta_{k,j}$, in which $\beta_{k,j}$ is the linear weight of an edge from X_k to X_j . Hence, the distribution of a Gaussian linear SEM is as follows:

$$X \sim N((I_p - B)^{-1}B_0, (I_p - B)^{-1}\Sigma_\epsilon(I_p - B)^{-T}),$$

where $I_p \in \mathbb{R}^{p \times p}$ is the identity matrix, and $\Sigma_\epsilon = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_p^2)$ is a covariance matrix of the independent errors. Hence, its density can be parameterized by the inverse covariance $\Theta = (I_p - B)^T \Sigma_\epsilon^{-1} (I_p - B) \succ 0$, and many algorithms apply the inverse covariance matrix to learn Gaussian linear SEMs.

2.2 Model Identifiability and Existing Algorithms

This section reviews recent studies on the model identifiability and learning algorithms for Gaussian linear SEMs. As discussed, Peters and Bühlmann (2014); Ghoshal and Honorio (2018); Park and Kim (2020) provide the identifiability conditions for Gaussian linear SEMs, and Park (2020) summarizes their identifiability conditions using the conditional variances.

Lemma 1 (Identifiability Conditions in Theorem 4 of Park, 2020) *Consider a Gaussian linear SEM in Equation (2) with DAG G and true ordering π . Then, DAG G is uniquely identifiable if one of the following conditions is satisfied. For any $r \in \{1, 2, \dots, p\}$, $\pi_r = j$, $k \in \text{De}(j)$, $\ell \in \text{An}(j)$, $S = \{\pi_1, \dots, \pi_{r-1}\}$, and $T_j = \{\pi_1, \dots, \pi_{p+1-r}\} \setminus \{j\}$.*

- *Equal error variance condition (Peters and Bühlmann, 2014):*

$$\sigma_1^2 = \sigma_2^2 = \dots = \sigma_p^2.$$

- *Forward selection condition (Park and Kim, 2020):*

$$\text{Var}(X_j \mid X_S) = \Sigma_{j,j} - \Sigma_{j,S} \Sigma_{S,S}^{-1} \Sigma_{S,j} < \Sigma_{k,k} - \Sigma_{k,S} \Sigma_{S,S}^{-1} \Sigma_{S,k} = \text{Var}(X_k \mid X_S).$$

- *Backward selection condition (Ghoshal and Honorio, 2018):*

$$\frac{1}{\text{Var}(X_j \mid X_{T_j})} = \frac{1}{\Sigma_{j,j} - \Sigma_{j,T_j} \Sigma_{T_j,T_j}^{-1} \Sigma_{T_j,j}} < \frac{1}{\Sigma_{\ell,\ell} - \Sigma_{\ell,T_\ell} \Sigma_{T_\ell,T_\ell}^{-1} \Sigma_{T_\ell,\ell}} = \frac{1}{\text{Var}(X_\ell \mid X_{T_\ell})},$$

where Σ is the covariance matrix for X . In addition, $A_{S,S'}$ denotes the sub-matrix of A corresponding to set of nodes S and S' .

Based on these identifiability conditions, many Gaussian linear SEM learning algorithms have been developed, and most of them can be categorized into the following three groups: (i) a graph-wise learning algorithm, (ii) an inverse covariance matrix-based algorithm, and (iii) a node-wise regression-based algorithm. A popular graph-wise estimation approach

is the GDS algorithm proposed in Peters and Bühlmann (2014). It applies the following penalized maximum likelihood by assuming that all variables are centered and all error distributions are identically distributed:

$$\{\hat{B}, \hat{\sigma}^2\} = \arg \min_{B \in \mathbb{R}^{p \times p}, \sigma^2 \in \mathbb{R}^+} \frac{np}{2} \log(2\pi\sigma^2) + \frac{n}{2\sigma^2} \text{tr} \left\{ (I - B)^T (I - B) \hat{\Sigma} \right\} + \frac{\log(n)}{2} \|B\|_0,$$

where B is the weight matrix, and $\hat{\Sigma}$ is the sample covariance matrix for X . In addition, $\|B\|_0 = |\{(j, k) \mid B_{jk} \neq 0\}|$. Hence, the GDS algorithm calculates the likelihood of all possible directed acyclic graphs.

There are several inverse covariance matrix-based algorithms. For instance, the algorithms developed in Loh and Bühlmann (2014) and Ghoshal and Honorio (2018) first estimate the last element of the ordering, using the diagonal entries of the inverse covariance matrix (the reciprocal of conditional variances) estimated by graphical lasso and CLIME. The algorithms then determine its parents with non-zero entries on its row of the estimated inverse covariance matrix. After eliminating the last element of the ordering, the algorithm applies the same procedure until the graph is estimated completely. Loh and Bühlmann (2014); Ghoshal and Honorio (2018) prove that their algorithms consistently estimate the true graph with sample bound $n = \Omega(d^4 \log p)$, where d is the maximum degree of the moralized graph. In addition, Park and Kim (2021) applies the graphical lasso and a conditional independence test-combined method with sample bound $n = \Omega(d^2 \log p)$.

In terms of regression-based algorithms, Park and Kim (2020) develops an ordinary least square and a conditional independence test-combined method for a low-dimensional Gaussian linear SEM. Particularly, it estimates the ordering from its onset using the variance of residuals, and then, it infers the directed edges using conditional independence tests. Similarly, Chen et al. (2019) develops a ℓ_0 - and ℓ_1 -regression-combined TD algorithm that estimates the ordering from the first element, applying the best subsets regression. Subsequently, it infers the edges using a ℓ_1 -regularized approach. Chen et al. (2019) proves that sample complexity $n = \Omega(q^2 \log p)$ for ordering estimation, in which q is the predetermined upper bound of the maximum indegree. Park et al. (2021) provides the ℓ_1 -regression-based Gaussian linear SEM learning (LSEM) algorithm with sample sizes $n = \Omega(d^2 \log p)$. The LSEM algorithm learns the last element of the ordering using the conditional variance, in which the conditioning set is reduced by ℓ_1 -regularized regression. Furthermore, it determines the parents from the solution of the corresponding ℓ_1 -regularized regression. The algorithm iterates this procedure to estimate the remaining elements of the ordering and their parents.

However, the consistency of the existing algorithms is inherently vulnerable to even a few outliers, owing to the small breakdown point of maximum likelihood estimator, least square estimator, graphical lasso, CLIME, and lasso (see details in Donoho and Huber (1983); Alfons et al. (2013); Loh and Tan (2018)). As outliers are often present in real-world problems, it is necessary to develop an outlier-robust Gaussian linear SEM learning algorithm that can cope with samples deviating from the model assumptions. Therefore, this study develops an outlier-robust regression based-algorithm to recover high-dimensional Gaussian linear SEMs.

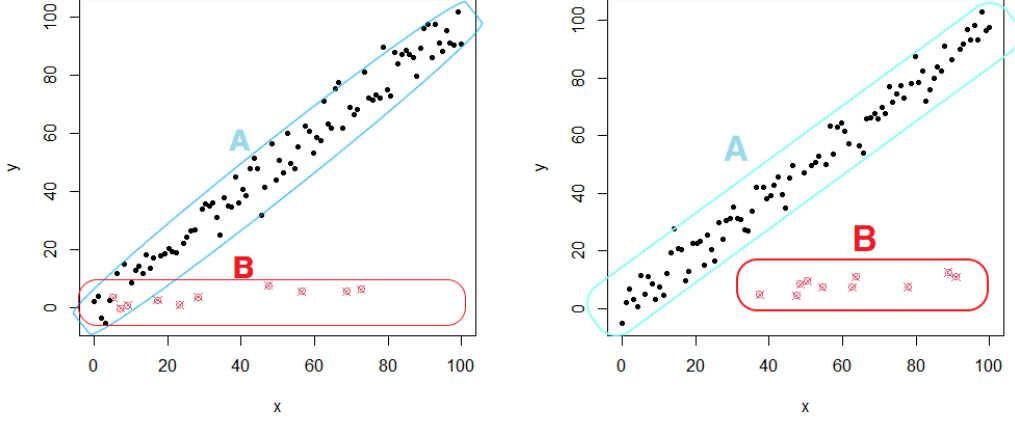


Figure 1: Two instances of outliers where non-outlier observations colored black are in A and outliers colored red are in B .

3. Corrupted Gaussian Linear SEMs

This section explains the considered outliers, and then formalizes the corrupted Gaussian linear SEMs by outliers. In a Gaussian linear SEM, an outlier should be defined on a node. Particularly, an outlier on a node is the observation that is not generated by the conditional distribution given its parents, whereas it still involves the distribution of its child as in Equation (2). Hence, the considered corrupted sample is not induced by the measurement error discussed in Zhang et al. (2017); Saeed et al. (2020), where an outlier on a node neither follows the conditional distribution nor influences its child distribution.

The motivation of this outlier is from the phenomenon where an observation vector rarely follows *all* causal relationships in Equation (2). For instance, most students, who are good at an elementary subject (A), are more likely to be good at an intermediate subject (B), and hence, they are good at an advanced subject (C). However, in reality, there are some students who are good at A and bad at B , which leads to poor performance in C . Similarly, there are some students who are good at A and B , while bad at C .

In addition, an outlier should be a data point that differs significantly from the non-outlier observations. For instance, suppose that the true model is $Y = X + N(0, \sigma^2)$ and consider outliers drawn from $\frac{1}{10}X + N(0, \sigma^2)$. Then, in Figure 1, the observations colored black in A are non-corrupted and the observations colored red in B are outliers. Although the outliers are not drawn from the true model, some outliers are close to the group of standard observations in Figure 1 (left), whereas all outliers are sufficiently far from the genuine observation group in Figure 1 (right). Throughout this study, the former case is not considered because these outliers cannot be eliminated, while they have a significant influence on estimating the conditional variance for model identifiability.

The corrupted Gaussian linear SEM by outliers can be defined as follows. Suppose that $(\varepsilon_j^{(i)})$ is a set of unspecified Gaussian outlier distributions that can be an arbitrary function

of their parents plus errors. Furthermore, suppose that $Y_j^{(i)}$ follows a Bernoulli distribution with $\Pr(Y_j^{(i)} = 1) = p_{ij}$, where p_{ij} is the fraction of outliers on X_j following $\varepsilon_j^{(i)}$. Then, for any $i \in \{1, 2, \dots, n\}$ and $j \in \{1, 2, \dots, p\}$, the distribution of the i -th observation of X_j is as follows:

$$(1 - Y_j^{(i)}) \left(\sum_{k \in \text{Pa}(j)} \beta_{k,j} X_k^{(i)} + \epsilon_j^{(i)} \right) + Y_j^{(i)} \varepsilon_j^{(i)}. \quad (4)$$

To provide intuition of the outlier and how outliers influence the covariance matrix, consider a three-node Gaussian linear SEM with outliers:

$$X_1 = \epsilon_1, \quad X_2 = \beta_1 X_1 + \epsilon_2, \quad \text{and} \quad X_3 = \beta_2 X_2 + \epsilon_3.$$

where $\epsilon_j \sim N(0, \sigma^2)$. Suppose that there is an outlier on X_j for $j \in \{1, 2, 3\}$, $X_j = \varepsilon_j$ where ε_j is an outlier distribution, $\varepsilon_j \sim N(0, \tau^2)$. Then, the covariance matrix for the true model is

$$\begin{bmatrix} \sigma^2 & \beta_1 \sigma^2 & \beta_1 \beta_2 \sigma^2 \\ \beta_1 \sigma^2 & (\beta_1^2 + 1) \sigma^2 & (\beta_1^2 \beta_2 + \beta_2) \sigma^2 \\ \beta_1 \beta_2 \sigma^2 & (\beta_1^2 \beta_2 + \beta_2) \sigma^2 & (\beta_1^2 \beta_2^2 + \beta_2^2 + 1) \sigma^2 \end{bmatrix}$$

Contrastingly, each corrupted model by an outlier on X_1, X_2, X_3 has the following covariance matrix, respectively.

$$\begin{bmatrix} \tau^2 & \beta_1 \tau^2 & \beta_1 \beta_2 \tau^2 \\ \beta_1 \tau^2 & \sigma^2 + \beta_1^2 \tau^2 & \beta_2 \sigma^2 + \beta_1^2 \beta_2 \tau^2 \\ \beta_1 \beta_2 \tau^2 & \beta_2 \sigma^2 + \beta_1^2 \beta_2 \tau^2 & (\beta_2^2 + 1) \sigma^2 + \beta_1^2 \beta_2^2 \tau^2 \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \tau^2 & \beta_2 \tau^2 \\ 0 & \beta_2 \tau^2 & \beta_2^2 \tau^2 + \sigma^2 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \beta_1 \sigma^2 & 0 \\ \beta_1 \sigma^2 & (\beta_1^2 + 1) \sigma^2 & 0 \\ 0 & 0 & \tau^2 \end{bmatrix}$$

As shown above, the covariance matrices of corrupted models are completely different to the true covariance matrix. Hence, the outlier in a Gaussian linear SEM can be understood as the *rowwise* contamination of data matrix rather than *cellwise* contamination (Alqallaf et al., 2009).

However, it should be pointed out that for the successful (inverse) covariance matrix estimation, only small fraction of rowwise contamination is allowed (Huber, 1992; Xue and Zou, 2012; Wegkamp and Zhao, 2016; Chen et al., 2018). Contrastingly, a large proportion of outliers can be allowed to recover the Gaussian linear SEM. For instance, suppose that there are three observations where each observation is an outlier on each node. Then, although all observations are contaminated, a majority of observations still follow the true linear relationships. This motivates a robust regression-based approach, rather than the robust inverse covariance matrix-based methods. The details on the new algorithm are provided in the next section.

4. Algorithm and Robust Identifiability Condition

This section introduces a new outlier-robust identifiability condition in terms of truncated conditional variances. In addition, it presents a robust regression-based algorithm for high-dimensional Gaussian linear SEMs. The proposed algorithm is the outlier-robust version of ℓ_1 -regularized regression based-algorithm (referred to as HLSM) developed in Park et al. (2021). Hence, the overall procedure of the proposed algorithm is the same as the HLSM

Algorithm 1: Robust Gaussian Linear SEM Learning Algorithm**Input** : n independent samples, $X^{1:n}$ **Output**: Estimated graph structure, $\hat{G} = (V, \hat{E})$ Set $\hat{\pi}_{p+1} = \emptyset$;**for** $r = \{1, 2, \dots, p-1\}$ **do** **for** $j \in V \setminus \{\hat{\pi}_{p+1}, \dots, \hat{\pi}_{p+2-r}\}$ **do** $S_j(r) = V \setminus (\{j\} \cup \{\hat{\pi}_{p+1}, \dots, \hat{\pi}_{p+2-r}\})$; Estimate $\hat{\theta}_j(r)$ for ℓ_1 -regularized LTS in Equation (5); Estimate truncated conditional variances $\widehat{\text{Var}}(X_j \mid X_{S_j(r)})$ using Equation (6); **end** Determine $\hat{\pi}_{p+1-r} = \arg \max_j \widehat{\text{Var}}(X_j \mid X_{S_j(r)})$; Determine $\widehat{\text{Pa}}(\hat{\pi}_{p+1-r}) = \{k \in S_j(r) : [\hat{\theta}_{\hat{\pi}_{p+1-r}}(r)]_k \neq 0\}$;**end****Return**: Estimate an edge set, $\hat{E} = \cup_{r \in \{1, 2, \dots, p-1\}} \{(k, \hat{\pi}_{p+1-r}) : k \in \widehat{\text{Pa}}(\hat{\pi}_{p+1-r})\}$

algorithm, where it estimates the last element of the ordering, and then determines its parents at the first iteration using ℓ_1 -regularized regression and truncated conditional variances. Subsequently, it estimates the next element of the ordering and its parents with the same method. It iterates this procedure until the complete graph structure is determined. However, unlike the HLSM algorithm, the proposed algorithm estimates the ordering and directed edges using ℓ_1 -regularized LTS and truncated conditional variances. Hence, it can successfully recover a high-dimensional Gaussian linear SEM from corrupted samples. The detailed process of the proposed algorithm is summarized in Algorithm 1.

Specifically, the r -th iteration of the algorithm estimates π_{p+1-r} and its parents, given the estimated $\hat{\pi}_{p+2-r}, \dots, \hat{\pi}_p$ if $r \geq 2$. Particularly, the r -th iteration of the algorithm is first conducted by the following ℓ_1 -regularized LTS: For each node $j \in V$,

$$(\hat{\theta}_j(r), \hat{H}_j(r)) := \arg \min_{\theta_{j0} \in \mathbb{R}, \theta_j \in \mathbb{R}^{|S_j(r)|}, H \in \mathbb{H}} \frac{1}{2h} \sum_{i \in H} \left(X_j^{(i)} - \theta_{j0} - \langle X_{S_j(r)}^{(i)}, \theta_j \rangle \right)^2 + \lambda_j(r) \|\theta_j\|_1, \quad (5)$$

where $S_j(r) = V \setminus (\{j\} \cup \{\hat{\pi}_{p+2-r}, \dots, \hat{\pi}_p\})$ if $r \geq 2$; otherwise, $S_j(r) = V \setminus \{j\}$. In addition, $\langle \cdot, \cdot \rangle$ represents the inner product and \mathbb{H} is a power set of $\{1, 2, \dots, n\}$ with predetermined trimmed sample size h .

Then, the proposed algorithm determines π_{p+1-r} with the largest conditional variance, given $X_{S_j(r)}$ for all $j \in \{\pi_1, \dots, \pi_{p+1-r}\}$, according to the backward selection condition in Lemma 1. However, since the ℓ_1 -regularized LTS estimator is biased, the mean squared residuals does not represent the error variance. Hence, the proposed algorithm applies the

following re-weighted conditional variance.

$$\widehat{\text{Var}}(X_j | X_{S_j(r)}) := \frac{1}{\sum_i \widehat{w}_i} \sum_{i=1}^n \widehat{w}_i (X_j^{(i)} - \widehat{\alpha}_{j0}(r) - \langle X_{\widehat{Q}_j(r)}^{(i)}, \widehat{\alpha}_j(r) \rangle)^2, \quad (6)$$

$$\text{where } (\widehat{\alpha}_{j0}(r), \widehat{\alpha}_j(r)) := \arg \min_{\alpha_0 \in \mathbb{R}, \alpha_j \in \mathbb{R}} \frac{1}{h} \sum_{i \in \widehat{H}_j(r)} \left(X_j^{(i)} - \alpha_0 - \langle X_{\widehat{Q}_j(r)}^{(i)}, \alpha_j \rangle \right)^2,$$

$\widehat{Q}_j(r)$ is the support of $\widehat{\theta}_j(r)$ (i.e., $\widehat{Q}_j(r) := \{k \in S_j(r) \mid [\widehat{\theta}_j(r)]_k \neq 0\}$) in which $[\widehat{\theta}_j(r)]_k$ is an element of $\widehat{\theta}_j(r)$ corresponding to the random variable X_k . In addition, $(\widehat{\alpha}_{j0}(r), \widehat{\alpha}_j(r))$ is the ordinary least square estimator from $X^{\widehat{H}_j(r)}$ where X_j is a response variable and $X_{\widehat{Q}_j(r)}$ are explanatory variables. Lastly, w_i is a binary weight that is 1 for a good observation; otherwise 0.

The weights can be estimated using the residuals of OLS in Equation (6) because the index set of outliers are unknown and $\widehat{H}_j(r)$ may contain a lot of non-corrupted observations. Specifically, the weights are determined as

$$\widehat{w}_i = \begin{cases} 1 & \text{if } |X_j^{(i)} - \widehat{\alpha}_{j0}(r) - \langle X_{\widehat{Q}_j(r)}^{(i)}, \widehat{\alpha}_j(r) \rangle| < \eta \\ 0 & \text{if } |X_j^{(i)} - \widehat{\alpha}_{j0}(r) - \langle X_{\widehat{Q}_j(r)}^{(i)}, \widehat{\alpha}_j(r) \rangle| \geq \eta, \end{cases} \quad (7)$$

for some predetermined positive constants η . The choice of η is discussed in Sections 4.2.2 and 5, and it should be emphasized that η does not depend on r and j .

The considered conditional variance estimator in Equation (6) is clearly a *truncated* conditional variance because it rules out some good observations with large residuals. Hence, Algorithm 1 requires the following robust identifiability condition.

Theorem 2 (Robust Identifiability Condition) *Consider a corrupted Gaussian linear SEM (4) with DAG G and true ordering π . Suppose that the backward selection condition is satisfied. Then, for any $r \in \{1, 2, \dots, p-1\}$, $\pi_r = j$, $\ell \in \text{An}(j)$, $T_j(r) = \{\pi_1, \dots, \pi_{p+1-r}\} \setminus \{j\}$, and $\eta > 0$,*

$$\frac{1}{\text{Var}(X_j | X_{T_j(r)}, |X_j - \mathbb{E}(X_j | X_{T_j(r)})| < \eta)} < \frac{1}{\text{Var}(X_\ell | X_{T_\ell(r)}, |X_\ell - \mathbb{E}(X_\ell | X_{T_\ell(r)})| < \eta)}.$$

The detailed proof is provided in Section A. Theorem 2 claims that the conditional variance relationships between node j and its ancestor are maintained for the truncated conditional variances, even when the thresholding parameters are the same for any r and j . Hence, it supports the validity of the ordering estimation step in the proposed algorithm.

Finally, the parent estimation for π_{p+1-r} is direct from the solution of ℓ_1 -regularized regression because its support is the parent for π_{p+1-r} in the population (see Proposition 5). Hence, the parents of node $j = \pi_{p+1-r}$ are determined as $\widehat{\text{Pa}}(j) := \{k \in S_j(r) \mid [\widehat{\theta}_j(r)]_k \neq 0\}$, where $\widehat{\theta}_j(r)$ is the solution to Equation (5).

4.1 Breakdown Point

A popular measure for the robustness of an estimator is the replacement finite-sample breakdown point. Let $Z = (X, y)$ denote the data matrix. The breakdown point of regression estimator β is defined as

$$\varepsilon^*(\hat{\beta}; Z) = \min_{1 \leq m \leq n} \left\{ \frac{m}{n} : \sup_{\tilde{Z}} \|\hat{\beta}(Z')\|_2 = \infty \right\},$$

where Z' are corrupted samples obtained from Z by replacing m of the original n data points with arbitrary values. Subsequently, Alfons et al. (2013) provides the following breakdown point result.

Theorem 3 (Breakdown Point of ℓ_1 -regularized LTS in Alfons et al., 2013) *Let $\rho(x)$ be a convex and symmetric loss function with $\rho(0) = 0$ and $\rho(x) > 0$ for $x \neq 0$, and define $\rho(x) = (\rho(x_1), \rho(x_2), \dots, \rho(x_n))^T$. With subset size $h \leq n$, consider the regression estimator*

$$\hat{\beta} = \arg \min_{\beta} \frac{1}{h} \sum_{i=1}^h (\rho(y - X\beta))_{i:n} + \lambda \|\beta\|_1,$$

where $(\rho(y - X\beta))_{i:n} \leq \dots \leq (\rho(y - X\beta))_{n:n}$ are the order statistics of the regression loss. Then, the breakdown point of $\hat{\beta}$ is given by

$$\varepsilon^*(\hat{\beta}; X) = \frac{n - h + 1}{n}.$$

As discussed in Alfons et al., 2013, the breakdown point property of ℓ_1 -regularized LTS shown in Theorem 3 is satisfied regardless of the data distributions and the size of dimension p . Hence, the r -th iteration of Algorithm 1 in Equation (5) has the same breakdown point:

$$\varepsilon^*(\hat{\theta}; X) = \frac{n - h + 1}{n}.$$

However, the solution to ℓ_1 -regularized LTS in Equation (5) involves not only the parents of a node, but its children. Particularly, suppose that $(\theta_{j0}^*(r), \theta_j^*(r))$ denotes the solution to Equation (5) with $\lambda = 0$ and $S_j(r) = \{\pi_1, \dots, \pi_{p+1-r}\} \setminus \{j\}$. Then, it can be expressed as follows:

$$(\theta_{j0}^*(r), \theta_j^*(r)) := \arg \min_{\theta_{j0} \in \mathbb{R}, \theta_j \in \mathbb{R}^{|S_j(r)|}} \mathbb{E} \left[(X_j - \theta_{j0} - \langle X_{S_j(r)}, \theta_j \rangle)^2 \right]. \quad (8)$$

In the Gaussian linear SEM setting, $\theta_{j0}^* + \langle X_{S_j(r)}, \theta_j^*(r) \rangle$ represents the conditional expectation, $\mathbb{E}(X_j | X_{S_j(r)})$. Specifically, simple algebra yields $\Sigma_{S_j(r), S_j(r)} \theta_j^*(r) = \Sigma_{S_j(r), j}$ where $\Sigma_{S_j(r), S_j(r)}$ is a sub-matrix of the true covariance matrix Σ corresponding to variables $X_{S_j(r)}$. Hence, for any $k \in (\text{Pa}(j) \cup \text{Ch}(j)) \cap T_j(r)$, in which $T_j(r) = \{\pi_1, \dots, \pi_{p+1-r}\} \setminus \{j\}$,

$$\theta_{k,j}^*(r) = \frac{\frac{\beta_{k,j}}{\sigma_j^2} + \frac{\beta_{j,k}}{\sigma_k^2} - \sum_{\ell' \in (\text{Ch}(j) \cap \text{Ch}(k)) \cap T_j(r)} \frac{\beta_{j,\ell'} \beta_{k,\ell'}}{\sigma_{\ell'}^2}}{\frac{1}{\sigma_j^2} + \sum_{\ell \in \text{Ch}(j) \cap T_j(r)} \frac{\beta_{j,\ell}^2}{\sigma_{\ell}^2}}, \quad (9)$$

where $\theta_{k,j}^*(r)$ is the true coefficient corresponding to X_k .

This implies that an outlier on parent $k' \in \text{Pa}(j)$ does not influence $\theta_{k',j}^*(r)$ because $\theta_{k,j}^*(r)$ is not involved with its error variance $\sigma_{k'}^2$ and corresponding edge weight $\beta_{k',j}$ is not changed by the definition of the outlier. In contrast, an outlier on child $k \in \text{Ch}(j)$ clearly changes $\theta_{k,j}^*(r)$ because it is a function of $\beta_{j,k}$ and σ_k^2 . Hence, it is emphasized that the *bad samples* for the $\theta_j^*(r)$ estimation are the outliers on X_j , and the outliers on X_k for $k \in \text{Ch}(j) \cap T_j(r)$. Thus, the following main result is reached when combining with Theorem 3.

Corollary 4 *ℓ_1 -regularized LTS problems of Equation (5) in Algorithm 1 are robust to outliers even in high-dimensional settings if*

$$|B| = \max_{j \in V} |B_j| = \max_{j \in V} \left| \mathcal{I}_j \cup_{k \in \text{Ch}(j)} \mathcal{I}_k \right| \leq n - h,$$

where $\mathcal{I}_j \subset \{1, 2, \dots, n\}$ is the index set of the outliers on X_j , and B_j is the index set of the bad samples for $\theta_j^*(1)$ estimation.

Corollary 4 shows that the smaller the value of h , the higher the breakdown point. Hence, it is possible to have a breakdown point larger than 50% for the $\theta_j(r)^*$ estimation. In addition, if a graph has a large maximum outdegree and $(\mathcal{I}_j)_{j \in V}$ are disjoint, then Algorithm 1 is more likely to be sensitive to outliers. Contrastingly, if a graph is sparse or $|\cup_{j \in V} \mathcal{I}_j|$ is small, the proposed algorithm is highly resilient to the massive number of outliers. Hence, in terms of a data matrix point of view, under appropriate conditions, Algorithm 1 can recover the underlying graph, even when all samples are rowwise outliers. The theoretical results on robustness are also reflected in the numerical experiments of Section 5.

4.2 Theoretical Guarantees

This section provides the statistical guarantees on Algorithm 1 for learning high-dimensional corrupted Gaussian linear SEMs. Specifically, the required assumptions and the theoretical results are provided for the consistent estimation of the directed edges, and ordering in Sections 4.2.1 and 4.2.2, respectively. The main results are expressed in terms of quadruple $(h, |B|, p, d)$, where h is the number of trimmed samples, $|B|$ is the maximum number of bad samples, p is the number of nodes, and d is the maximum degree of the moralized graph.

The consistency of Algorithm 1 is not surprising because the two key components of the proposed algorithm, ℓ_1 -regularized LTS and the ℓ_1 -regularized regression-based algorithm, are consistent. Hence, the theoretical results for the sign recovery of Equation (5) is analogous to the prior work in Yang et al., 2018, and proves the consistency of high-dimensional ℓ_1 -regularized LTS. In addition, the theoretical guarantees for the ordering estimation is based on the results in Park et al. (2021) that establishes the consistency of the ℓ_1 -regularized regression-based Gaussian linear SEM learning algorithm. However, the aforementioned works are not for the estimation of corrupted Gaussian linear SEM, and hence, this section presents the detailed statistical guarantees on Algorithm 1.

Before discussing the consistency of Algorithm 1, it begins by justifying the sparsity of the model (5). As shown in Equation (9), when j is π_{p+1-r} , $\theta_j^*(r)$ corresponds exactly to

the set of parents; that is, $[\theta_j^*(r)]_k = \beta_{k,j}$ if $k \in \text{Pa}(j)$; otherwise, $[\theta_j^*(r)]_k = 0$. However, when j is not π_{p+1-r} , the support of $\theta_j^*(r)$ is not the same as the parents of j . Hence, the following proposition is necessary to guarantee that the problem in terms of $\theta_j^*(r)$ is still a sparse regression problem under the bounded degree of the moralized graph condition. Thus, the proof is omitted (see details in Proposition 2 of Park et al., 2021).

Proposition 5 (Proposition 2 of Park et al., 2021) *For any $r \in \{1, 2, \dots, p-1\}$ -th iteration and $j \in \{\pi_1, \dots, \pi_{p+1-r}\}$, the support of the solution $\theta_j^*(r)$ is a subset of the neighborhood of j in the moralized graph:*

$$\text{Supp}(\theta_j^*(r)) \subset \text{Ne}(j).$$

In addition, for $j = \pi_{p+1-r}$, they are the parents of j :

$$\text{Supp}(\theta_j^*(r)) = \text{Pa}(j).$$

4.2.1 SIGN RECOVERY OF ℓ_1 -REGULARIZED LTS

The first key step of the proposed algorithm is the sign recovery of ℓ_1 -regularized LTS in Equation (5) that can be restated as the following weighted regularized problem: For any $r \in \{1, 2, \dots, p-1\}$ and $j \in \{\pi_1, \dots, \pi_{p+1-r}\}$,

$$(\hat{\theta}_{j0}, \hat{\theta}_j(r)) := \arg \min_{w \in \mathbb{W}, \theta_{j0} \in \mathbb{R}, \theta_j \in \rho \mathbb{B}_1} \frac{1}{2h} \sum_{i=1}^n w_i \left(X_j^{(i)} - \theta_{j0} - \langle X_{S_j(r)}^{(i)}, \theta_j \rangle \right)^2 + \lambda_j(r) \|\theta_j\|_1, \quad (10)$$

where $\mathbb{W} = \{w \in \{0, 1\}^n \mid \sum_{i=1}^n w_i = h\}$, \mathbb{B}_1 is the ℓ_1 -norm ball, and constraint $\theta_j \in \rho \mathbb{B}_1$ is $\|\theta\|_1 \leq \rho$. In addition, $S_j(r) = V \setminus (\{j\} \cup \{\hat{\pi}_{p+2-r}, \dots, \hat{\pi}_p\})$ if $r \geq 2$; otherwise, $S_j(r) = V \setminus \{j\}$.

For the consistency of the ℓ_1 -regularized LTS problem in Equation (10), the following assumptions are required. In principle, these assumptions help the commonly assumed restricted strong convexity on $\theta_j(r)$ and structural incoherence conditions to hold (Negahban et al., 2012; Yang et al., 2018 and Section B).

Assumption 6 *Suppose that B_j is the index set of bad samples for the $\theta_j^*(1)$ estimation in Equation (5); that is, $B_j = \mathcal{I}_j \cup_{k \in \text{Ch}(j)} \mathcal{I}_k$. There exists a positive constant $\alpha_{\min} \in (0, 1]$ such that*

$$\max_{j \in V} |B_j| \leq \min \{(1 - \alpha_{\min})h, n - h\}.$$

Assumption 7 *Suppose that $X_{S_j(r)}^{B_j}$ is the sub-design matrix in $\mathbb{R}^{|B_j| \times p-1}$ corresponding to bad samples for $\theta_j^*(1)$. Then, for any $j \in V$ and any subset $K_j \subset V \setminus \{j\}$ with $|K_j| = |B_j|$, there exists a positive constant $c_{\max} > 0$ such that*

$$\max_j \max_{K_j} \|X_{K_j}^{B_j}\| \leq c_{\max} \sqrt{|B_j| \log p}$$

where $\|A\| = \sup_{\|v\| \leq 1} \|Av\|_2$ is a spectral norm.

Assumption 8 For any $r \in \{1, 2, \dots, p-1\}$ -th iteration and $j \in \{\pi_1, \dots, \pi_{p+1-r}\}$, there exists a positive constant $c_1 > 0$ such that

$$\theta_{\min} < \min_{j,r} \|\theta_j^*(r)\|_{\min} \leq \max_{j,r} \|\theta_j^*(r)\|_1 \leq \rho \leq \frac{c_1}{2} \sqrt{\frac{h}{\log p}}$$

where ρ is the tuning parameter in Equation (10).

Assumption 6 implies that the maximum number of bad samples is strictly less than the number of trimmed samples. In addition, this condition reflects the breakdown point result in Section 4.1, that is, $|B| = \max_j |B_j| \leq n - h$. It is quite acceptable because the proposed algorithm aims for models that fit the majority of the data, that is, $h \geq \frac{n}{2}$ and $\max_{j \in V} |B_j| < \frac{n}{2}$. Assumption 7 is a mild restriction on outlier distributions because it is satisfied with high probability, if the outlier distributions are (sub-)Gaussian. As discussed in Vershynin (2010), for any positive constant $t > 0$, there exist positive constants c_2 and c_3 such that $\|X_{K_j}^{B_j}\| \leq ((1 + c_2)\sqrt{|B_j|} + t) \|(\Sigma_{K_j}^{B_j})^{1/2}\|$ with probability $1 - 2\exp(-c_3 t^2)$, where $\Sigma_{K_j}^{B_j}$ is the true covariance matrix for $X_{K_j}^{B_j}$.

Assumption 8 puts a constraint on the number of trimmed samples such that $h \geq \left(\frac{2 \max_{j,r} \|\theta_j^*(r)\|_1}{c_1}\right)^2 \log p$. In addition, this assumption forces that the coefficients of $\|\theta_j(r)^*\|_\infty$ are bounded by $\frac{c_1}{2} \sqrt{\frac{h}{d^2 \log p}}$. Lastly, Assumption 8 ensures that each non-zero coefficient of $\theta_j^*(r)$ is sufficiently far away from zero.

With these assumptions, the main results show that each objective problem in Equation (10) achieves the sign consistency.

Theorem 9 (Sign Recovery) Consider ℓ_1 -regularized LTS in Equation (10) and $\hat{\theta}_j(r)$ is its solution. Suppose that Assumptions 6, 7, and 8 are satisfied. Then, for any positive constant $\epsilon > 0$, there exists positive constants c_ϵ and κ_1 , such that if a regularization parameter is $\lambda_j(r) = c_\epsilon \sqrt{\frac{\log p}{h}}$ and the number of trimmed samples is $h \geq \left(\frac{c_\epsilon}{\kappa_1 \alpha_{\min} \theta_{\min}^2}\right)^2 \left(\frac{9}{4}d + 4|B_j|\right) \log p$,

$$\Pr\left(\text{sign}\left(\hat{\theta}_j(r)\right) = \text{sign}\left(\theta_j^*(r)\right)\right) \geq 1 - \frac{\epsilon}{p^2}.$$

The detailed proof is provided in Appendix B. Theorem 9 shows that for the fixed number of trimmed samples, ℓ_1 -regularized LTS performs better as portions of genuine samples α_{\min} and minimum edge weight θ_{\min} increase. Theorem 9 also indicates that if the number of trimmed samples is $h = \Omega((d + |B_j|) \log p)$, the support of $\theta_j^*(r)$ for the problem in Equation (10) can be successfully recovered with high probability. Hence, the performance of the proposed ℓ_1 -regularized LTS problem depends on the sparsity level of a graph and the fraction of corruptions.

Applying the union bound, for any $r \in \{1, 2, \dots, p-1\}$ -th iteration and $j \in \{\pi_1, \dots, \pi_{p+1-r}\}$, ℓ_1 -regularized LTS successfully recovers the support of $\theta_j(r)$ with high probability, if $h = \Omega(d + |B| \log p)$. Furthermore, a combination of Proposition 5 and Theorem 9 implies that the proposed algorithm accurately learns the parents of each node with high probability when the ordering is provided.

Corollary 10 (Consistency of the Parents Estimation) *Consider a corrupted Gaussian linear SEM (4). Suppose that Assumptions 6, 7, and 8 are satisfied, and an appropriate regularization parameter is chosen. In addition, suppose that a true ordering is provided. Then, Algorithm 1 recovers the parents of each node with high probability, if the number of trimmed sample size $h = \Omega(d + \max_j |\mathcal{I}_j| \log p)$.*

4.2.2 ORDERING RECOVERY

The second key step of the proposed algorithm is the ordering estimation using truncated conditional variances. Hence, given the support of $\theta_j^*(r)$, the ordering estimation step of Algorithm 1 is equivalent to solving the following $p - 1$ iterative problems: For $r \in \{1, 2, \dots, p - 1\}$,

$$\hat{\pi}_r := \arg \max_{j \in V \setminus \{\hat{\pi}_{p+1-r}, \dots, \hat{\pi}_p\}} [\hat{\Sigma}_{S'_j(r)}^{\hat{G}_j(r)}]_{j,j}^{-1},$$

where $\hat{\Sigma}_{S'_j(r)}^{\hat{G}_j(r)}$ is a sample covariance matrix for $X_{S'_j(r)}^{\hat{G}_j(r)} \in \mathbb{R}^{|\hat{G}_j(r)| \times |S'_j(r)|}$ corresponding to genuine observations, where $S'_j(r) := \{j\} \cup \text{Supp}(\theta_j^*(r))$ and $\hat{G}_j(r)$ is the estimated index set of good observations estimated by the magnitude of residuals using Equation (7). In addition, $[A]_{j,j}$ denotes the diagonal entry of A corresponding to variable X_j .

It begins by discussing the required condition to eliminate the effects of outliers for the conditional variance estimation. This guarantees the large distance between the bad samples and true expected value.

Assumption 11 *For any $r \in \{1, 2, \dots, p - 1\}$ -th iteration, $j \in \{\pi_1, \dots, \pi_{p+1-r}\}$, and $T_j(r) = \{\pi_1, \dots, \pi_{p+1-r}\} \setminus \{j\}$, there exists a positive constant $\eta_{\min} > 0$ such that*

$$\min_{i \in B_j(r)} |X_j^{(i)} - \mathbb{E}(X_j^{(i)} | X_{T_j(r)}^{(i)})| > \eta_{\min},$$

where $B_j(r)$ is an index set of bad samples for X_j for the r -th iteration, that is, $B_j(r) = \mathcal{I}_j \cup_{k \in \text{Ch}(j) \cap T_j(r)} \mathcal{I}_k$.

Assumption 11 ensures that, in population, $G_j(r)$ excludes all bad samples for $\theta_j^*(r)$ with an appropriate choice of η in Equation (6). In order to eliminate all bad samples, η should be smaller than η_{\min} . Hence, this study recommends small η such that $h' = \min_{j,r} |G_j(r)| \geq h$, if h is sufficiently larger than the number of outliers. It should be pointed out that Assumption 11 does not help prevent the falsely detected outliers in $G_j(r)$.

The next required assumption is the sample version of the robust identifiability condition in terms of *truncated* conditional variances. Similar assumptions regarding the conditional variance can be found in the previous studies for learning Gaussian linear SEMs (e.g., Park, 2020; Park et al., 2021).

Assumption 12 *For any $r \in \{1, 2, \dots, p - 1\}$ -th iteration, $j \in \{\pi_1, \dots, \pi_{p+1-r}\}$, $T_j(r) = \{\pi_1, \dots, \pi_{p+1-r}\} \setminus \{j\}$, and $\eta \in (0, \eta_{\min})$, there exists a positive constant $\tau_{\min} > 0$ such that*

$$\text{Var}\left(X_j | X_{T_j(r)}, |X_j - \mathbb{E}(X_j | X_{T_j(r)})| < \eta\right) - \text{Var}\left(X_\ell | X_{T_j(r)}, |X_\ell - \mathbb{E}(X_\ell | X_{T_j(r)})| < \eta\right) > \tau_{\min}$$

Armed with Assumptions 11 and 12, the main result shows that Algorithm 1 estimates the ordering with high probability.

Theorem 13 (Recovery of the Ordering) *Consider a corrupted Gaussian linear SEM (4). Suppose that Assumptions 11 and 12 are satisfied. In addition, suppose that, for any $r \in \{1, 2, \dots, p-1\}$ and $j \in \{\pi_1, \dots, \pi_{p+1-r}\}$, the supports of $(\theta_j^*(r))$ are provided. Lastly, suppose that $\eta \in (0, \eta_{\min})$. Then, for any positive constant $\epsilon > 0$, there exists a positive constant C'_ϵ such that if the number of samples for the truncated conditional variance estimation $h' = \min_{j,r} |\widehat{G}_j(r)| \geq C'_\epsilon d^2 \log p$,*

$$\Pr(\widehat{\pi} \in \Pi^*) \geq 1 - \epsilon.$$

where Π^* is a set of the true orderings of a graph.

The detailed proof is provided in Section F. The main idea of the proof is to show the consistency of the truncated conditional variance estimator and the truncated conditional variance comparisons based on the robust identifiability condition in Theorem 2. The provided sample complexity $h' = \min_{j,r} |\widehat{G}_j(r)| = \Omega(d^2 \log p)$ intuitively makes sense, because the ordering estimation step involves $O(p^2)$ truncated conditional variance estimations, with given variables that size up to maximum degree d . Although the proof involves the truncated conditional variance of Gaussian distribution, the procedure is similar to the proof for Theorem 8 of Park et al. (2021), where the conditional variance of (sub-)Gaussian distribution is considered. Hence, the proof is constructed upon the one in Park et al. (2021).

Finally, by combining Theorems 9, 13, and Corollary 10, the final result is reached that Algorithm 1 successfully recovers the true structure of a corrupted Gaussian linear SEM with high probability.

Corollary 14 (Consistency of Algorithm 1) *Consider a corrupted Gaussian linear SEM (4). Suppose that Assumptions 6, 7, 8, and 12 are satisfied. In addition, suppose that appropriate regularization and conditional variance threshold parameters are chosen. Then, Algorithm 1 recovers the true graph with high probability if $h = \Omega((d + |B|) \log p)$ and $h' = \Omega(d^2 \log p)$.*

Corollary 14 indicates that the performance of Algorithm depends on the number of trimmed samples h , the maximum number of bad samples $|B|$, the number of nodes p , and the thresholding parameter for truncated conditional variances η . The proven sample complexity is empirically verified in the numerical experiments of Section 5.

5. Numerical Experiments

This section presents the empirical performance of Algorithm 1 and supports the theoretical results of robustness to outliers in Sections 4.1, and the sample complexity result of the proposed algorithm $h = \Omega((d + |B|) \log p)$ and $h' = \Omega(d^2 \log p)$ discussed in Section 4.2. Hence, the proposed algorithms are evaluated by varying number of samples with the following settings: (i) various portions of trimmed samples α , (ii) various maximum number of outliers $|B|$, (iii) various number of nodes p , and (iv) various truncated conditional

variance threshold η that determines the h' . Furthermore, a comparison of Algorithm 1 shows the standard ℓ_1 -regularized regression-based HLSM (Park et al., 2021), the graphical lasso-based HGSM (Park and Kim, 2021), the ℓ_0 - and ℓ_1 -regression-based TD (Chen et al., 2019), the conditional independence-based US (Park and Kim, 2020), and the likelihood-based GDS (Peters and Bühlmann, 2014) algorithms in terms of accuracy.

In terms of the evaluation measure, the empirical probability of successfully recovering all edges is applied; that is, $\Pr(E = \hat{E})$ to validate our theoretical findings from Corollaries 4 and 14. In addition, the proposed algorithm is compared with the US, HGSM, HLSM, LISTEN, and TD algorithms in terms of the average precision ($\frac{\# \text{ of correctly estimated edges}}{\# \text{ of estimated edges}}$) and recall ($\frac{\# \text{ of correctly estimated edges}}{\# \text{ of true edges}}$). In terms of empirical probability, precision and recall, the bigger the better.

The regularization parameters for the proposed algorithm were set to $\sqrt{\frac{\log p}{h}}$ as provided in Theorem 9. In addition, for the HGSM, HLSM, and TD algorithms, the regularization parameters were set to $\sqrt{\frac{\log p}{n}}$. For the TD algorithm, the predetermined parameter q was set to the true maximum indegree of a graph. For the US algorithm, the backward selection condition and Fisher's independence test are exploited with the significance level $\alpha = 1 - \Phi(0.5n^{1/3})$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Lastly, the GDS algorithm applies a random graph to the initial graph, just as in Peters and Bühlmann (2014).

5.1 Random Graph with Outliers on a Node

The first conducted simulations were 100 realizations of p -node Gaussian linear SEMs (4) with randomly generated underlying DAG structures for node size $p \in \{5, 10, \dots, 25\}$ where the maximum degree is $d \leq 4$ and the minimum indegree is 1. Hence, the considered graphs are sparse and there are no isolated nodes. The set of non-zero parameters, $\beta_{jk} \in \mathbb{R}$ in Equation (4), was uniformly generated, at random, in the range $\beta_{k,j} \in (-1, -0.75) \cup (0.75, 1)$. In addition, all noise variances were set to $\sigma_j^2 = 0.75$. Lastly, the various number of outliers on the first element of the ordering, $|\mathcal{I}_{\pi_1}| \in \{1, 30, 60, 90\}$, were considered and they were generated from $N(100, \sigma_j^2)$. In this setting, Assumption 11 is well satisfied and the maximum number of bad samples are well restricted, that is, $|B| = \max_{j \in V} |\mathcal{I}_j \cup_{k \in \text{Ch}(j)} \mathcal{I}_k| = |\mathcal{I}_{\pi_1}|$.

Figure 2 compares the empirical probability of successful 10-node corrupted Gaussian linear SEMs recovery using Algorithm 1 with various fractions of the sample size $\alpha \in \{0.5, 0.6, \dots, 0.9\}$ by varying sample size $n \in \{50, 100, \dots, 500\}$ for different number of bad samples $|B| \in \{1, 30, 60, 90\}$, respectively. For simplicity, only the results of the proposed algorithm with $\eta = 2$ are presented.

Figure 2 heuristically confirms that the breakdown point result of the proposed algorithm discussed in Section 4.1. The proposed algorithm can successfully recover a graph, if $|B| \leq n - h = (1 - \alpha) \times n$. Specifically, the proposed method with $\alpha = 0.9$ can recover a graph when $|B| = 1$, even when $n = 50$, because it can ignore the effect of $0.1 \times n = 5$ outliers. However, it fails to recover any graph when $|B| = 30$ and $n \leq 250$ because the trimmed samples inevitably contain some outliers owing to $|B| = 30 > 0.1 \times 250 = 25$. It cannot recover a graph when $|B| \in \{60, 90\}$ due to the same reason. For other α cases, the same phenomenon can be seen that the proposed algorithm fails to recover a graph if

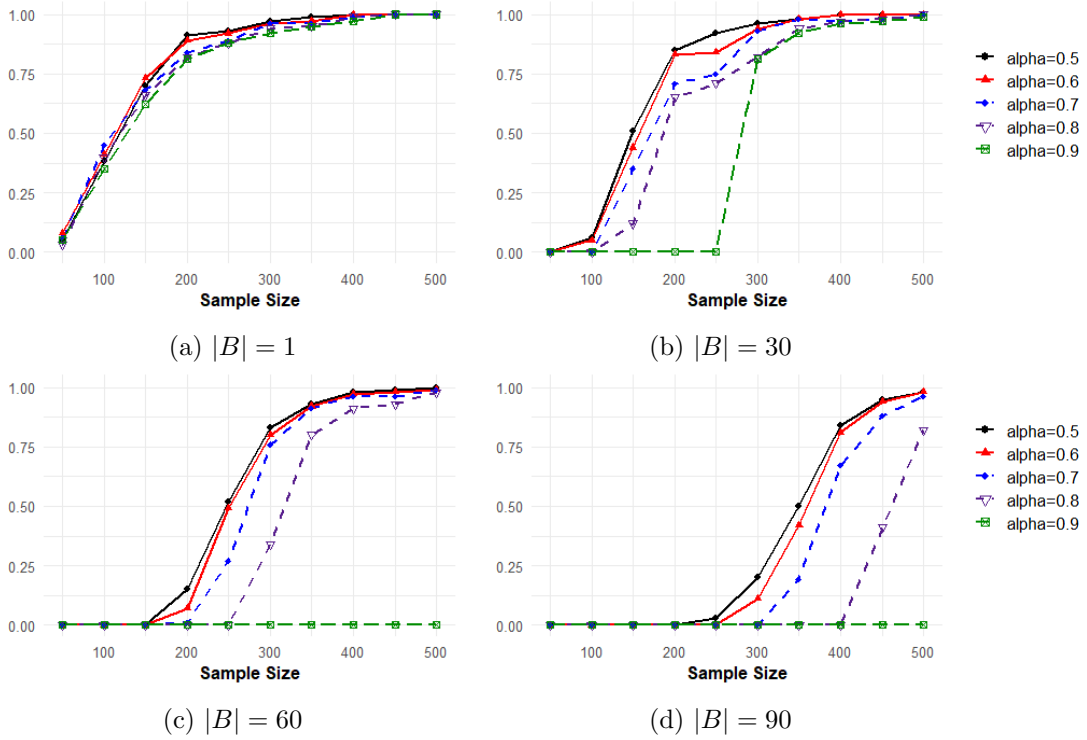


Figure 2: Performance of the proposed algorithm with various fractions of the sample size, $\alpha \in \{0.5, 0.6, \dots, 0.9\}$, for learning 10-node corrupted Gaussian linear SEMs with the different maximum number of bad samples ($|B| \in \{1, 30, 60, 90\}$) on the first element of the ordering. The empirical probability of successful graph recovery is shown versus sample size.

$|B| < (1 - \alpha) \times n$. Contrastingly, Figure 2 highlights that the proposed algorithm consistently learns the model from corrupted observations if $|B| < (1 - \alpha) \times n$.

Furthermore, as shown in Figure 2, the probability of graph recovery decreases as the maximum number of bad samples increase. This implicitly supports one of the main results that the sample bound of the proposed algorithm involves the maximum number of bad samples, as proven in Corollary 14. In order to precisely validate the sample complexity $h = \Omega((d + |B|) \log p)$ shown in Corollary 14, Figures 3 and 4 compare the performance of the proposed algorithm with different $|B|$ and p , respectively.

Figure 3 (a) compares the empirical probability of successful 10-node corrupted Gaussian linear SEMs estimation via Algorithm 1, when the maximum number of bad samples are $|B| \in \{30, 60, \dots, 150\}$, by varying sample size $n \in \{50, 100, \dots, 500\}$. In addition, Figure 3 (b) illustrates the empirical probability against re-scaled trimmed sample size $C = h/|B|$. For a simple comparison, the results with number of trimmed samples $h = \frac{n}{2}$ and $\eta = 2$ are only presented. Figure 3 shows that the empirical curves for different number of outliers more closely align with the re-scaled sample size on the horizontal axis. This confirms Corollary 14 that trimmed sample size h required for a successful graph structure recovery

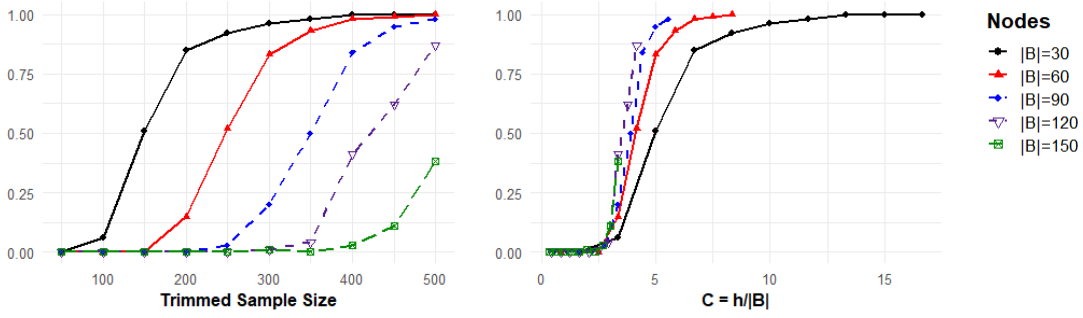


Figure 3: Performance of the proposed algorithm for 10-node corrupted Gaussian linear SEMs with various number of bad samples ($|B| \in \{30, 60, \dots, 150\}$). The empirical probability of successful directed graph recovery is shown versus trimmed sample size h (left) and versus re-scaled sample size $C = h/|B|$ (right).

scales linearly with the maximum number of bad samples $|B|$, given the fixed number of nodes.

However, it should be pointed out that given a fixed re-scaled sample size, as the maximum number of bad samples increases, the probability of successful graph recovery appears to increase. It is mainly because the sample complexity of the proposed algorithm is determined by the support recovery of ℓ_1 -regularized LTS, and the number of samples for truncated conditional variance, $h' = \Omega(d^2 \log p)$ as shown in Corollary 14. In other words, given a fixed re-scaled sample size, $h/|B| = c$, the corresponding sample size n increases as the maximum number of bad samples $|B|$ increases. Hence, as $|B|$ increases, h' increases, which leads to the more accurate ordering estimation.

Figures 4 (a) and (c) show the empirical probability of successful DAG recovery using Algorithm 1 with $\alpha = 0.5$ and $\eta = 2$, by varying sample size $n \in \{50, 100, \dots, 500\}$ when $|B| = 1$ and 30, respectively. Figures 4 (b) and (d) plot the empirical probability against re-scaled sample size $C = h/\log p$. This confirms Corollary 14 that the trimmed sample size required for a successful graph structure recovery scales logarithmically, with the number of nodes given the fixed number of bad samples. Hence, the empirical curves for different number of nodes align closely with this re-scaled sample size on the horizontal axis as expected, and this is clearly provided in Figures 4 (b) and (d).

Figure 5 provides the empirical probability of successful 10-node corrupted Gaussian linear SEMs' learning via Algorithm 1 with various threshold for truncated conditional variances $\eta \in \{1, 1.5, 2, 3, 100\}$, by varying sample size $n \in \{50, 100, \dots, 500\}$ when the numbers of outliers $|B| \in \{1, 30\}$, respectively. For simplicity, the results with $h = \frac{n}{2}$ are only presented.

This numerical result heuristically supports Theorem 13 and Corollary 14 that the sample bound for the Algorithm 1 involves the number of samples for the truncated conditional variance estimations. In other words, the performance of Algorithm 1 depends on the truncated conditional variance threshold. Specifically, the algorithm with $\eta = 100$ performs the worst because it violates Assumption 11, and hence, fails to eliminate the outliers, when estimating the conditional variance. Hence this leads to incorrect ordering estimation. In

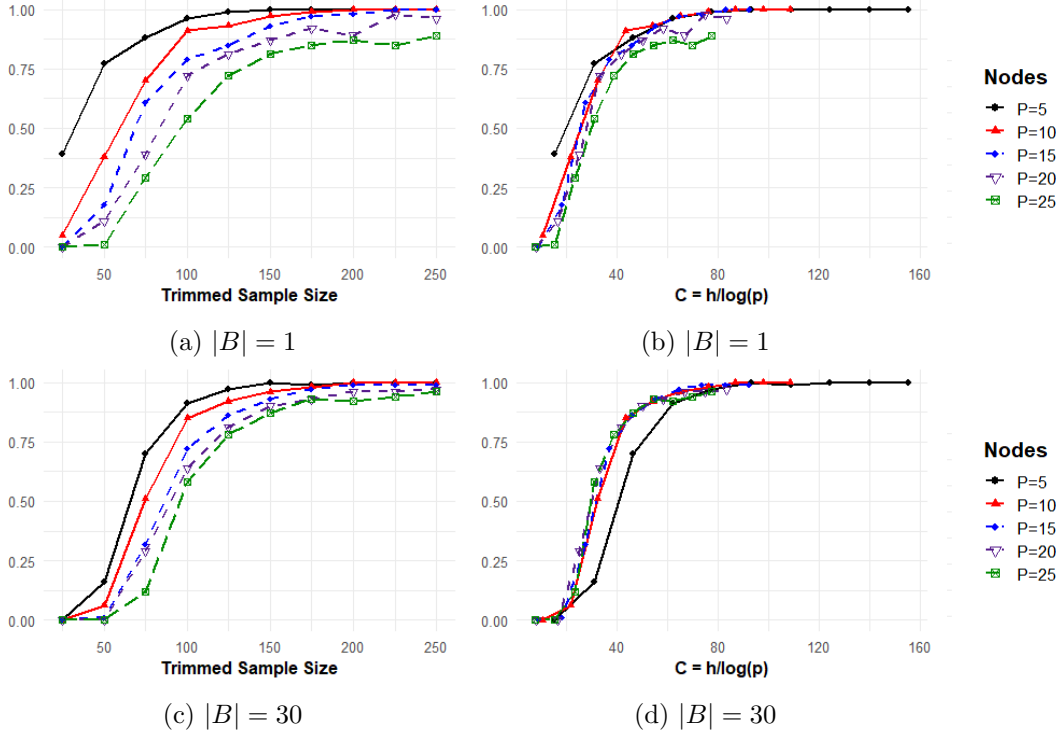


Figure 4: Performance of the proposed algorithm for learning p -node corrupted Gaussian linear SEMs with $\{1, 30\}$ outliers on the first element of the ordering. The empirical probability of successful directed graph recovery is shown versus trimmed sample size h (left) and versus re-scaled sample size $C = h/(\log p)$ (right).

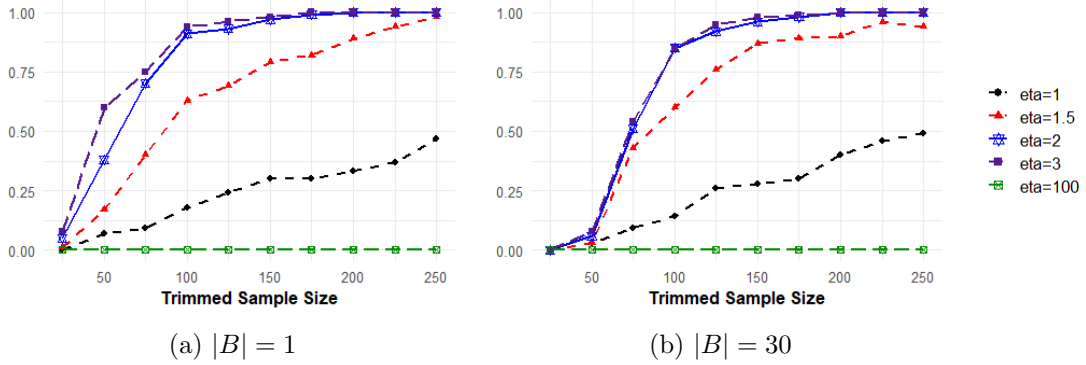


Figure 5: Performance of the proposed algorithm for learning 10-node corrupted Gaussian linear SEMs with various threshold for truncated conditional variances ($\eta \in \{1, 1.5, 2, 3, 100\}$). The empirical probability of successful directed graph recovery is shown versus trimmed sample size.

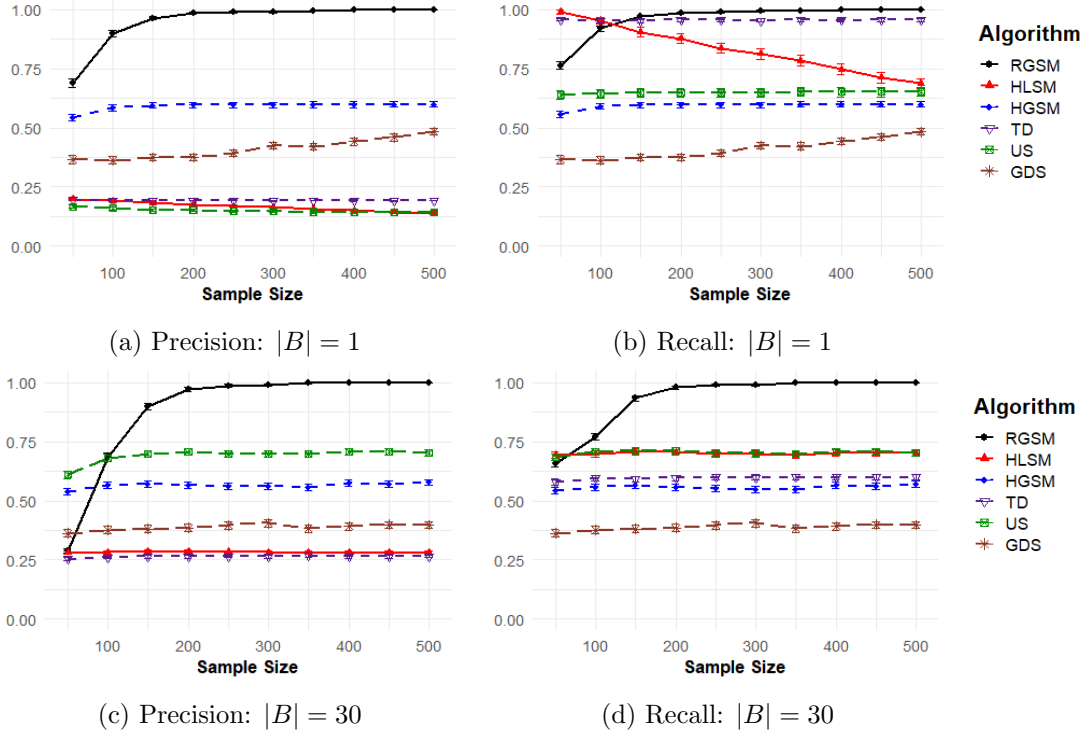


Figure 6: Comparison of the proposed algorithm (RGSM) against the HLSM, HGSM, TD, and US algorithms in terms of the average precision and recall for learning 10-node corrupted Gaussian linear SEMs with $|B| \in \{1, 30\}$.

addition, when $\eta = 1$, the number of samples for the conditional variance estimation is too small, which also leads to the low accuracy. Similarly, the proposed algorithm with $\eta = 3$ performs better than the method with $\eta = 1.5$ and 2 in this setting. Although the performance of the proposed algorithm relies on η , the simulation results still highlight that the proposed algorithm consistently learns the corrupted Gaussian linear SEMs, as long as η is sufficiently small.

Figure 6 evaluates the proposed algorithm (RGSM) and the state-of-the-art HLSM, HGSM, TD, US, and GDS algorithms in terms of the average precision and recall for recovering 10-node graphs by varying sample size $n \in \{50, 100, \dots, 500\}$. Again, for a simple presentation, the proposed algorithm with $\alpha = 0.5$ and $\eta = 2$ are only presented. As shown in Figure 6, the average precision and recall for the proposed algorithm converges to 1, which verifies the consistency of the proposed algorithm. Contrastingly, the comparison methods appear to be inconsistent even when $|B| = 1$. This phenomenon is expected because the comparison methods are designed for learning (Gaussian) linear SEMs from standard i.i.d. samples. Furthermore, Figure 6 shows that comparison methods can be vulnerable to a single outlier, owing to the small breakdown point. Hence, this comparison result highlights the advantages of the proposed algorithm.

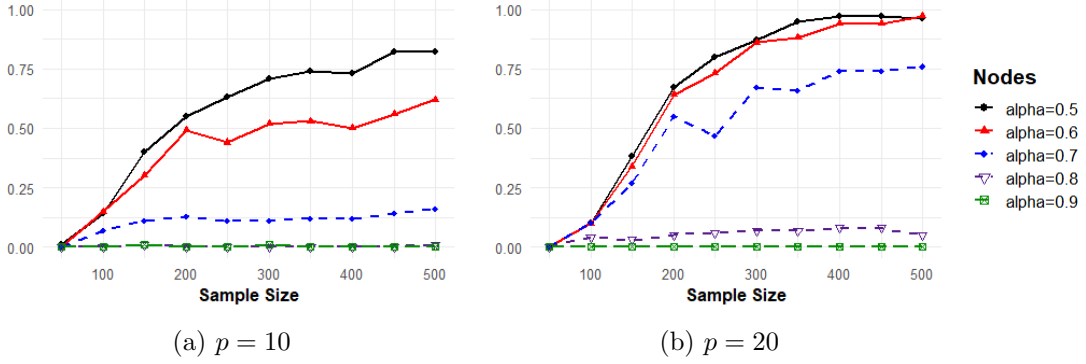


Figure 7: Performance of the proposed algorithm with various fractions of the sample size, $\alpha \in \{0.5, 0.6, \dots, 0.9\}$, for learning 10 and 20-node corrupted Gaussian linear SEMs with different maximum number of bad samples ($|B| \in \{1, 30, 60, 90\}$) when all observation are outliers. The empirical probability of successful graph recovery is shown versus sample size.

5.2 Random Graph with Outliers on All Node

This section verifies another main result that Algorithm 1 successfully learns corrupted Gaussian linear SEMs, even when there exist outliers on all nodes, that is, $\mathcal{I}_j \neq \emptyset$ for all $j \in V$. The considered setting is the worst case, where all observations are outliers on some nodes, that is, $\cup_{j \in V} \mathcal{I}_j = \{1, 2, \dots, n\}$. Hence, 100 sets of graphs and samples were generated under the procedure specified in Sections 5.1, except that i -th observation is an outlier on $X_{i \bmod p}$, where $a \bmod b$ is the remainder of a when divided by b . Then, the proposed algorithm and the comparison methods are evaluated in various settings, as done in Section 5.1.

Compared to the setting in Section 5.1, the major difference is that the number of outliers on each node is $|\mathcal{I}_j| = \frac{n}{p}$. In addition, the maximum number of bad samples is $|B| = \max_{j \in V} |\mathcal{I}_j \cup_{k \in \text{Ch}(j)} \mathcal{I}_k| \leq \frac{5n}{p}$ because the number of children can be $d = 4$. For instance, when the number of nodes are $p = 10$, the numbers of outliers on each node and maximum number of bad samples are $|\mathcal{I}_j| = 0.1n$ and (possibly) $|B| = 0.5n$, respectively.

Figure 7 compares the empirical probability of successful 10 and 20-node corrupted Gaussian linear SEMs estimation using Algorithm 1, with $\eta = 2$ and $\alpha \in \{0.5, 0.6, \dots, 0.9\}$, by varying sample size $n \in \{50, 100, \dots, 500\}$. Figure 7 supports the main contributions of this study that (i) the proposed algorithm can recover a graph when the number of trimmed samples are smaller than the maximum number of bad samples, $\alpha \times n < \max_{j \in V} |B_j|$ and (ii) the sample bound of the proposed algorithm relies on the maximum number of bad samples of variables, instead of total number of outliers. Figure 7 also confirms that (iii) the maximum number of bad samples depends on the outliers of a node and its child.

Specifically, when $p = 10$, the maximum number of bad samples are almost $\max_{j \in V} |B_j| = \frac{5n}{10}$ or $\frac{4n}{10}$ depending on the underlying structure. Hence, the proposed algorithm with $\alpha \in \{0.7, 0.8, 0.9\}$ fails to recover a graph, whereas the algorithm with $\alpha \in \{0.5, 0.6\}$ consistently recovers a graph. In addition, when $p = 20$, the maximum number of bad samples

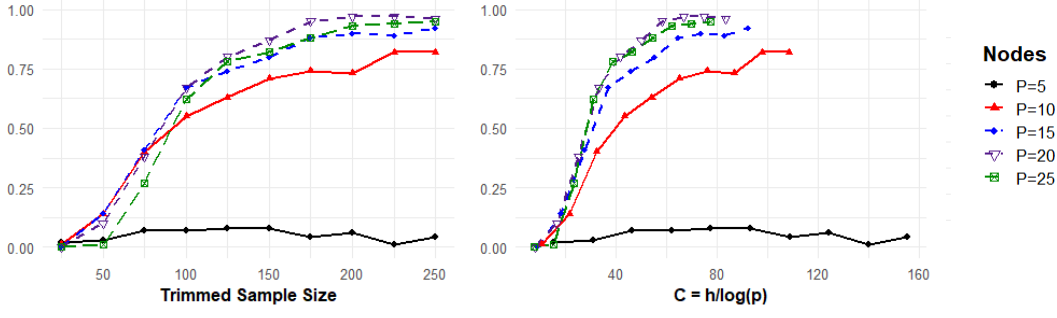


Figure 8: Performance of the proposed algorithm for learning 10-node corrupted Gaussian linear SEMs when all observation are outliers. The empirical probability of successful directed graph recovery is shown versus trimmed sample size h (left) and versus re-scaled sample size $C = h/(\log p)$ (right).

are $\max_{j \in V} |B_j| = \frac{5n}{20}$ or $\frac{4n}{20}$. Hence, due to the same reason, the proposed algorithm with $\alpha \in \{0.8, 0.9\}$ appears to fail in recovering a true graph, whereas the algorithm with $\alpha \in \{0.5, 0.6, 0.7\}$ can successfully recover a graph.

Figure 8 (a) illustrates the empirical probability of successful DAG recovery of Algorithm ?? with $\alpha = 0.5$ and $\eta = 2$, by varying sample size $n \in \{50, 100, \dots, 500\}$. As shown in Figure 8, the proposed algorithm consistently learns the model, even when all observations are outliers on some nodes, except when $p = 5$. This phenomenon can be explained by the breakdown point result because, when $p = 5$, there are possibly too many bad samples $|B| = \frac{5n}{p} \geq (1 - \alpha) \times n$. Hence, the proposed method is vulnerable to the outliers, according to Corollary 10. However, for $p \in \{10, 15, \dots, 25\}$, $|B| \leq \frac{5n}{p} \geq (1 - \alpha) \times n$. Hence, the proposed algorithm appears to consistently learn the model in Figure 8.

In addition, Figure 8 (b) shows the empirical probability against re-scaled sample size $C = h/\log p$, as in Figure 4. However, unlike Figure 4, where the maximum number of bad samples is fixed, the empirical curves for different problem sizes do not closely align with this re-scaled sample size because the maximum number of bad samples are different, depending on the number of nodes. Specifically, it is clearly seen that as the number of nodes increases, the maximum number of bad samples are more likely to decrease, which leads to the increased empirical probability of successful graph recovery. Hence, these simulation results support the breakdown point and sample complexity results in Corollaries 10 and 14.

Figure 9 compares the empirical probability of successful 10-node corrupted Gaussian linear SEMs learning via Algorithm 1, with $\alpha = 0.5$ and $\eta \in \{1, 1.5, 2, 3, 100\}$, by varying sample size $n \in \{50, 100, \dots, 500\}$. As in Figure 5, these numerical experiments confirm that, under the required conditions, the proposed algorithm consistently learns the corrupted Gaussian linear SEMs, as long as η is sufficiently small, regardless of the number of corrupted variables and rowwise outliers.

Figure 10 evaluates the proposed algorithm (RGSM) and the state-of-the-art HLMS, HGSM, TD, US, and GDS algorithms in terms of the average precision and recall for learning 10-node corrupted Gaussian linear SEMs, when all observation are outliers, by varying sample size $n \in \{50, 100, \dots, 500\}$. Again, for simple presentation, the proposed algorithm

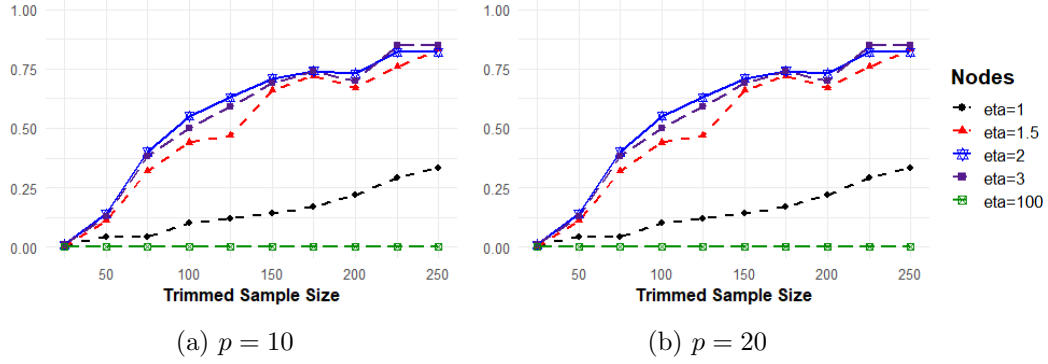


Figure 9: Performance of the proposed algorithm for learning 10- and 20-node corrupted Gaussian linear SEMs with various threshold for truncated conditional variances ($\eta \in \{1, 1.5, 2, 3, 100\}$) when all observation are outliers. The empirical probability of successful directed graph recovery is shown versus trimmed sample size h .

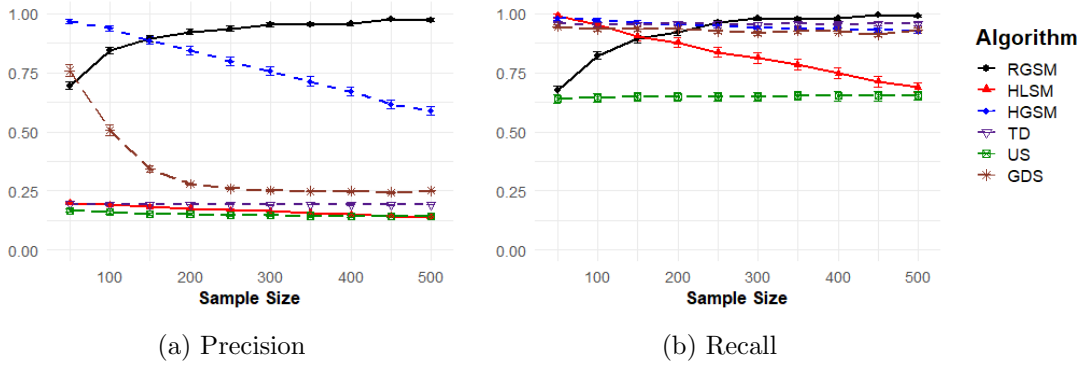


Figure 10: Comparison of the proposed algorithm (RGSM) against the HLSM, HGSM, TD, and US algorithms in terms of average Hamming distance for learning 10-node corrupted Gaussian linear SEMs when all observation are outliers.

with $\alpha = 0.5$ and $\eta = 2$ are only presented as in Figure 6. As expected, the simulation results are analogous to the results in Figure 6. Specifically, the proposed algorithm recovers the true directed edges better as the sample size increases, even when all observations are outliers, whereas the comparison methods are still inconsistent. Therefore, this reiterates the necessity and advantages of the proposed algorithm.

6. Discussion and Future Works

This study focuses on developing a statistically consistent outlier-robust algorithm for learning high-dimensional Gaussian linear SEMs using ℓ_1 -regularized LTS and a new robust identifiability condition, in terms of the truncated conditional variance. In addition, this study provides the theoretical guarantees on the proposed algorithm, and the theoretical results

of this study are empirically supported through various numerical experiments. However, there are several strong assumptions and difficulties on the appropriate choice of parameters. Furthermore, the proposed method has suffered seriously due to the heavy computational cost induced by ℓ_1 -regularized LTS. Nevertheless, to the best of my knowledge, it is the first theoretical result to learn Gaussian linear SEM from corrupted samples.

Several topics remain for future work. As discussed, one of the critical challenges of the proposed algorithm is the exponentially growing computational cost in the number of observations. In addition, it is polynomially growing in the number of nodes. However, it is conjectured that the proposed algorithm can be improved in terms of the computational cost, as the algorithm may not require $O(p^2)$ ℓ_1 -regularized LTS. In addition, the result of LTS in the r -th iteration of the proposed algorithm can help the LTS in the next iteration, with reference to computational cost. Hence, it is possible to develop a more computationally efficient algorithm. Furthermore, the proposed algorithm needs restrictive Gaussian error and linearity assumptions. Hence, one may develop a new method with milder conditions and prove its consistency.

Acknowledgments

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF-2021R1C1C1004562)

Appendix A. Proof for Theorem 2

Proof Suppose that $r \in \{1, 2, \dots, p-1\}$, $k \in \{\pi_{r+1}, \dots, \pi_p\}$, and $T_k = \{\pi_1, \dots, \pi_{p+1-r}\} \setminus \{k\}$. Then, we have

$$X_k \mid X_{T_k} \sim N(\mathbb{E}(X_k \mid X_{T_k}), \text{Var}(X_k \mid X_{T_k})).$$

For ease of notation, let $\zeta_{k|T_k} = X_k - \mathbb{E}(X_k \mid X_{T_k})$ and $\eta_{k|T_k} = \eta / \sqrt{\text{Var}(X_k \mid X_{T_k})}$ for any given positive constant η . Then, the given truncated conditional variance $|X_k - \mathbb{E}(X_k \mid X_{T_k})| < \eta$ is as follows.

$$\text{Var}(X_k \mid X_{T_k}, |\zeta_{k|T_k}| < \eta) = \text{Var}(X_k \mid X_{T_k}) \left(1 - 2\eta \frac{\phi(\eta_{k|T_k})}{\sqrt{\text{Var}(X_k \mid X_{T_k})}(2\Phi(\eta_{k|T_k}) - 1)} \right)$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the density and cumulative distribution function of the standard normal distribution, respectively.

In addition, under the backward selection condition in Lemma 1, for $\pi_r = j$ and any $\ell \in \text{An}(j)$, we have

$$\text{Var}(X_\ell \mid X_{T_\ell}) < \text{Var}(X_j \mid X_{T_j}).$$

Since $\eta_{\ell|T_\ell} < \eta_{j|T_j}$, we obtain

$$\frac{\phi(\eta_{\ell|T_\ell})}{2\Phi(\eta_{\ell|T_\ell}) - 1} > \frac{\phi(\eta_{j|T_j})}{2\Phi(\eta_{j|T_j}) - 1}.$$

Therefore, simple algebra yields that

$$\begin{aligned} & \text{Var}(X_j | X_{T_j}, |\zeta_{j|T_j}| < \eta) - \text{Var}(X_\ell | X_{T_\ell}, |\zeta_{\ell|T_\ell}| < \eta) \\ & > (\text{Var}(X_j | X_{T_j}) - \text{Var}(X_\ell | X_{T_\ell})) - 2\eta(\sqrt{\text{Var}(X_j | X_{T_j})} - \sqrt{\text{Var}(X_\ell | X_{T_\ell})}) \frac{\phi(\eta_{j|T_j})}{2\Phi(\eta_{j|T_j}) - 1}. \end{aligned}$$

Hence, if the following holds,

$$\frac{2\eta\phi(\eta_{j|T_j})}{2\Phi(\eta_{j|T_j}) - 1} \leq \sqrt{\text{Var}(X_j | X_{T_j})} + \sqrt{\text{Var}(X_\ell | X_{T_\ell})},$$

then we obtain $\text{Var}(X_j | X_{T_j}, |\zeta_{j|T_j}| < \eta) > \text{Var}(X_\ell | X_{T_\ell}, |\zeta_{\ell|T_\ell}| < \eta)$.

Since the cumulative distribution function of standard normal distribution is approximately

$$\Phi(x) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2} \left[x + \frac{x^3}{3} + \frac{x^5}{3 \cdot 5} + \cdots + \frac{x^{2n+1}}{(2n+1)!!} + \cdots \right],$$

we have

$$\frac{2x\phi(x)}{2\Phi(x) - 1} = \frac{1}{\left[1 + \frac{x^2}{3} + \frac{x^4}{3 \cdot 5} + \cdots + \frac{x^{2n}}{(2n+1)!!} + \cdots \right]},$$

where !! is the double factorial.

Hence, we have

$$\frac{2\eta\phi(\eta_{j|T_j})}{2\Phi(\eta_{j|T_j}) - 1} \leq \sqrt{\text{Var}(X_j | X_{T_j})}.$$

Therefore, this completes the proof that for any $\eta > 0$,

$$\text{Var}(X_j | X_{T_j}, |\zeta_{j|T_j}| < \eta) > \text{Var}(X_\ell | X_{T_\ell}, |\zeta_{\ell|T_\ell}| < \eta).$$

■

Appendix B. Proof for Theorem 9

Proof This section proves the sign recovery of the weighted regularized problem in Equation (10) for given $r \in \{1, 2, \dots, p-1\}$ and $j \in \{\pi_1, \dots, \pi_{p+1-r}\}$. For a better presentation, the intercept θ_{j0} , subscript j , and parenthesis (r) on $(\theta_j(r), \omega_j(r), \lambda_j(r), S_j(r))$ are omitted. Furthermore, let $\omega_j(r)^{(i)}$ and $\epsilon_j(r)^{(i)}$ be ω_i and ϵ_i , respectively. Moreover, $X_j^{(i)}$ and $X_{S_j(r)}^{(i)}$ are denoted by $X_j^{(i)}$ and $X_S^{(i)}$, respectively.

Let (θ^*, ω^*) and $(\tilde{\theta}, \tilde{\omega})$ be the true and a solution to the weighted regularized problem in Equation (10). Here, ω_i^* is $\tilde{\omega}_i$, if i -th observation is non-corrupted; otherwise 0. This proof is mostly built upon the theoretical results in Yang et al. (2018).

Borrowing the notations in Yang et al. (2018), suppose that $\mathcal{L}(\theta, w)$ to denote

$$\frac{1}{h} \sum_{i=1}^n w_i \bar{L}(\theta; X^{(i)}) \text{ where } \bar{L}(\theta; X^{(i)}) = \frac{1}{2} \left(X_j^{(i)} - \langle X_S^{(i)}, \theta_j \rangle \right)^2$$

Then, in order to guarantee the bounded errors of $\tilde{\theta}$, the following assumptions are required:

Assumption 15 (Restricted strong convexity on θ) For any possible $\Delta = \theta - \theta^*$, the differentiable loss function \bar{L} satisfies

$$\langle \nabla_{\theta} \mathcal{L}(\theta^* + \Delta, \omega^*) - \nabla_{\theta} \mathcal{L}(\theta^*, \omega^*), \Delta \rangle \geq \kappa_{\ell} \|\Delta\|_2^2 - \tau_1(n, p) \|\Delta\|_1^2$$

where κ_{ℓ} is a curvature parameter, and $\tau_1(n, p)$ is a tolerance function on number of samples n and nodes p .

Assumption 16 (Structural incoherence condition) Suppose that $\tilde{\Delta} = \tilde{\theta} - \theta^*$ and $\tilde{\Gamma} = \tilde{w} - w^* \in \{0, 1\}^n$. Then,

$$\langle \nabla_{\theta} \mathcal{L}(\theta^* + \tilde{\Delta}, \omega^* + \tilde{\Gamma}) - \nabla_{\theta} \mathcal{L}(\theta^* + \tilde{\Delta}, \omega^*), \tilde{\Delta} \rangle \geq \tau_2(n, p) \|\tilde{\Delta}\|_2^2 - \tau_3(n, p) \|\tilde{\Delta}\|_1^2,$$

where $\tau_2(n, p)$ and $\tau_3(n, p)$ are positive functions of number of samples n and nodes p .

Under these robust version of the standard restricted strong convexity and structural incoherence conditions discussed in Yang et al. (2018), we get the following result.

Theorem 17 (Theorem 1 of Yang et al., 2018) Let $(\tilde{\theta}, \tilde{w})$ be a solution to the weighted regularized problem in Equation (10). Suppose that Assumptions 15 and 16 are satisfied. In addition, suppose that the regularization parameter is set to

$$\lambda \geq 4 \max \{ \|\nabla_{\theta} \mathcal{L}(\theta^*, \omega^*)\|_{\infty}, 2\rho\tau_1(n, p) + \tau_3(n, p) \}. \quad (11)$$

Then, the following error bounds for $\tilde{\theta}$ are guaranteed for a given model space \mathcal{M} :

$$\|\tilde{\theta} - \theta^*\|_2 \leq \frac{1}{\kappa_{\ell}} \left(\frac{3\lambda\Psi}{2} + 2\tau_2(n, p) \right), \text{ where } \Psi := \sup_{u \in \mathcal{M} \setminus \{0\}} \|u\|_1 / \|u\|_2.$$

The remaining proof is focused such that Assumptions 15 and 16 are satisfied under Assumptions 6, 7, and 8 in Section 4.2. Throughout the proof, the following fact is applied that all elements in Γ corresponding to H_{opt} (set of non-corrupted observations) are all zeros: $\Gamma_{H_{opt}} = 0$ by construction.

Lemma 18 Assumption 15 is satisfied under Assumptions 6 and 8 if

$$\kappa_{\ell} = \kappa_1 \alpha_{\min}, \quad \text{and} \quad \tau_1(n, p) = \frac{\kappa_2 \log p}{h}.$$

Lemma 19 Assumption 16 is satisfied under Assumptions 7 if

$$\tau_2(n, p) = c_{\max} \max_{i \notin H_{opt}} |\delta_i| \frac{|B| \sqrt{\log p}}{h}, \quad \text{and} \quad \tau_3(n, p) = c_{\max} \max_{i \notin H_{opt}} |\delta_i| \frac{\sqrt{|B| \log p}}{h},$$

where c_{\max} is a positive constant.

Lemma 20 Suppose that Assumptions 6, 7, and 8 are satisfied. Then, for any positive constant $\epsilon > 0$, there exists a positive constant c_{ϵ} such that

$$\lambda = c_{\epsilon} \sqrt{\frac{\log p}{h}} \geq 4 \max \{ \|\nabla_{\theta} \mathcal{L}(\theta^*, \omega^*)\|_{\infty}, 2\rho\tau_1(n, p) + \tau_3(n, p) \}. \quad (12)$$

with probability at least $1 - \epsilon/p^2$.

The detailed proofs for Lemmas 18, 19, and 20 are provided in Sections C, D, and E, respectively. Plugging the results of the lemmas into Theorem 17, we get the following results.

$$\|\tilde{\theta} - \theta^*\|_\infty \leq \|\tilde{\theta} - \theta^*\|_2 \leq \frac{c_\epsilon}{\kappa_1 \alpha_{\min}} \left(\frac{3}{2} \sqrt{\frac{d \log p}{h}} + 2 \sqrt{\frac{|B| \log p}{h}} \right).$$

Applying Assumptions 6 and 8, if trimmed sample size $h \geq \left(\frac{c_\epsilon}{\kappa_1 \alpha_{\min} \theta_{\min}} \right)^2 \left(\frac{9}{4} d + 4|B| \right) \log p$.

$$\Pr \left(\text{sign} \left(\hat{\theta}_j(r) \right) = \text{sign} \left(\theta_j^*(r) \right) \right) \geq 1 - \frac{\epsilon}{p^2}.$$

■

Appendix C. Proof for Lemma 18

Proof Simple algebra yields that Assumption 15 can be re-written as

$$\frac{1}{h} \sum_{i=1}^n w_i^* \langle X_S^{(i)}, \tilde{\Delta} \rangle^2 \geq \kappa_l \|\tilde{\Delta}\|_2^2 - \tau_1(n, p) \|\tilde{\Delta}\|_1^2 \quad (13)$$

Recall that H_{opt} is an (optimal) index set of good observations following the true relationships with $|H_{opt}| = n - |B_j|$. Here, $|B_j|$ denotes the number of bad samples for $\theta_j^*(r)$. Suppose that \hat{H}_{opt} is an estimated index set used for parameter estimation in ℓ_1 -regularized LTS, such that $|H_{opt}| = h$ and it is a set of the indices corresponding h smallest values of $\bar{L}(\theta; X^{(i)})$. Lastly, for ease of notation, let $h^* = n - |B_j| \geq h$.

Again, recall that w^* is constructed as follows: w_i^* is simply set to \tilde{w}_i if $i \in H_{opt}$ and $w_i^* = 0$ for if $i \notin H_{opt}$. Hence, by construction, $\sum_{i \in H_{opt}} w_i^* \geq h - (n - h^*)$ (owing to $\sum_i \tilde{w}_i = h$). Thus, at least $h - (n - h^*)$ samples in H_{opt} have $\tilde{w}_i = w_i^* = 1$. Let \bar{H}_{opt} , which is the subset of H_{opt} , be the set of such samples, that is, $\bar{H}_{opt} = \{i \in H_{opt} \mid \tilde{w}_i = 1\}$.

Then, $\frac{1}{h} \sum_{i=1}^n w_i^* \langle X_S^{(i)}, \tilde{\Delta} \rangle^2$ can be lower bounded as follows:

$$\frac{1}{h} \sum_{i=1}^n w_i^* \langle X_S^{(i)}, \tilde{\Delta} \rangle^2 = \frac{1}{h} \sum_{i \in H_{opt}} w_i^* \langle X_S^{(i)}, \tilde{\Delta} \rangle^2 \geq \frac{1}{h} \sum_{i \in \bar{H}_{opt}} w_i^* \langle X_S^{(i)}, \tilde{\Delta} \rangle^2 \geq \frac{1}{h} \sum_{i \in \bar{H}_{opt}} \langle X_S^{(i)}, \tilde{\Delta} \rangle^2.$$

Noting that $(X_S^{(i)})_{i \in \bar{H}_{opt}}$ are i.i.d sampled from the distribution of the corrupted Gaussian linear SEM and referred as $N(0, \Sigma_S)$, where $\Sigma_S \in \mathbb{R}^{|S| \times |S|}$. Applying the result in Theorem 1 of Raskutti et al. (2010), there are some positive constants c_1 and c_2 such that

$$\frac{1}{|\bar{H}_{opt}|} \sum_{i \in \bar{H}_{opt}} \langle X_S^{(i)}, \tilde{\Delta} \rangle^2 \geq \left(\frac{1}{4} \|\Sigma_S^{1/2} \tilde{\Delta}\|_2 - \sqrt{\max_{j \in \{1, 2, \dots, |S|\}} [\Sigma_S]_{j,j} \frac{\log |S|}{|\bar{H}_{opt}|}} \|\tilde{\Delta}\|_1 \right)^2 \quad (14)$$

with probability at least $1 - c_1 \exp(-c_2 |\bar{H}_{opt}|)$.

Since $|S| < p$ and $\|\Sigma_S^{1/2}\tilde{\Delta}\|_2 \geq \sqrt{\Lambda_{\min}(\Sigma_S)}\|\tilde{\Delta}\|_2$ in which $\Lambda_{\min}(A)$ is the minimum eigenvalue of A ,

$$\frac{1}{|\bar{H}_{opt}|} \sum_{i \in \bar{H}_{opt}} \langle X_S^{(i)}, \tilde{\Delta} \rangle^2 \geq \kappa_1 \|\tilde{\Delta}\|_2^2 - \kappa_2 \frac{\log p}{|\bar{H}_{opt}|} \|\tilde{\Delta}\|_1^2 \quad (15)$$

where $\kappa_1 = \sqrt{\Lambda_{\min}(\Sigma_S)}$ and $\kappa_2 = \sqrt{\max_{j \in \{1, 2, \dots, |S|\}} [\Sigma_S]_{j,j}}$. Hence, κ_1 and κ_2 are strictly positive constants depending only on Σ_S .

Applying the fact that $|\bar{H}_{opt}| \geq h - (n - h^*)$,

$$\frac{1}{h} \sum_{i=1}^n w_i^* \langle X_S^{(i)}, \tilde{\Delta} \rangle^2 \geq \kappa_1 \frac{|\bar{H}_{opt}|}{h} \|\tilde{\Delta}\|_2^2 - \kappa_2 \frac{\log p}{h} \|\tilde{\Delta}\|_1^2 \geq \frac{\kappa_1(h - (n - h^*))}{h} \|\tilde{\Delta}\|_2^2 - \frac{\kappa_2 \log p}{h} \|\tilde{\Delta}\|_1^2$$

Applying Assumption 6 ($\frac{h - (n - h^*)}{h} \geq \alpha_{\min}$), Condition (13) holds with

$$\kappa_l = \kappa_1 \alpha_{\min}, \text{ and } \tau_1(n, p) = \frac{\kappa_2 \log p}{h}.$$

■

Appendix D. Proof for Lemma 19

Proof Assumption 16 can be restated as

$$\frac{1}{h} \tilde{\Gamma}_i \left(\langle X_S^{(i)}, \theta^* + \tilde{\Delta} \rangle - X_j^{(i)} \right) \langle X_S^{(i)}, \tilde{\Delta} \rangle \geq -\tau_2(n, p) \|\tilde{\Delta}\|_2 - \tau_3(n, p) \|\tilde{\Delta}\|_1. \quad (16)$$

Hence, this proof focuses on the lower bound for $\frac{1}{h} \tilde{\Gamma}_i \left(\langle X_S^{(i)}, \theta^* + \tilde{\Delta} \rangle - X_j^{(i)} \right) \langle X_S^{(i)}, \tilde{\Delta} \rangle$. Suppose that $X_j^{(i)} = \langle X_S^{(i)}, \theta^* \rangle + \epsilon_i I(i \in H_{opt}) + \delta_i I(i \notin H_{opt})$. Applying the fact that $\tilde{\Gamma}_i \geq 0$ because if $i \in H_{opt}$, $\tilde{\Gamma}_i = 0$. Otherwise, $\tilde{\Gamma}_i := \tilde{w}_i - w_i^* \geq 0$, and the following is obtained.

$$\begin{aligned} & \frac{1}{h} \tilde{\Gamma}_i \left(\langle X_S^{(i)}, \theta^* + \tilde{\Delta} \rangle - X_j^{(i)} \right) \langle X_S^{(i)}, \tilde{\Delta} \rangle \\ &= \frac{1}{h} \tilde{\Gamma}_i \langle X_S^{(i)}, \tilde{\Delta} \rangle^2 - \frac{1}{h} \sum_{i=1}^n \tilde{\Gamma}_i (\epsilon_i I(i \in H_{opt}) + \delta_i I(i \notin H_{opt})) \langle X_S^{(i)}, \tilde{\Delta} \rangle \\ &\geq -\frac{1}{h} \sum_{i=1}^n \tilde{\Gamma}_i (\epsilon_i I(i \in H_{opt}) + \delta_i I(i \notin H_{opt})) \langle X_S^{(i)}, \tilde{\Delta} \rangle, \end{aligned}$$

Applying the technique exploited in Candes et al. (2006), given $\tilde{\Gamma}$, we divide the index set of $\tilde{\Delta}$ into the disjoint exhaustive subsets Z_1, Z_2, \dots, Z_q of size $|H_{opt}^c|$, such that Z_1 contains the indices of the $|H_{opt}^c|$ largest absolute elements in $\tilde{\Delta}$, Z_2 contains the indices of the next $|H_{opt}^c|$ largest absolute elements, and so on. Hence, $q \leq \max\{1, \frac{p}{|H_{opt}^c|}\}$.

Then, by the definition of $\tilde{\Gamma}$ and the Cauchy-Schwarz inequalities,

$$\begin{aligned}
 & \left| \sum_{i=1}^n \tilde{\Gamma}_i (\epsilon_i I(i \in H_{opt}) + \delta_i I(i \notin H_{opt})) \langle X_S^{(i)}, \tilde{\Delta} \rangle \right| \\
 &= \left| \sum_{i \notin H_{opt}} \tilde{\Gamma}_i \delta_i \langle X_S^{(i)}, \tilde{\Delta} \rangle \right| = \left| \sum_{i \notin H_{opt}} \tilde{\Gamma}_i \delta_i \sum_{j=1}^q \langle X_{Z_j}^{(i)}, \tilde{\Delta}_{Z_j} \rangle \right| \\
 &\leq \sum_j \left| \sum_{i \notin H_{opt}} \tilde{\Gamma}_i \delta_i \langle X_{Z_j}^{(i)}, \tilde{\Delta}_{Z_j} \rangle \right| \leq \sum_j \sqrt{\sum_{i \notin H_{opt}} \tilde{\Gamma}_i^2 \delta_i^2} \sqrt{\sum_{i \notin H_{opt}} \langle X_{Z_j}^{(i)}, \tilde{\Delta}_{Z_j} \rangle^2}.
 \end{aligned}$$

Applying the property of the norm, we have

$$\begin{aligned}
 \sum_j \sqrt{\sum_{i \notin H_{opt}} \tilde{\Gamma}_i^2 \delta_i^2} \sqrt{\sum_{i \notin H_{opt}} \langle X_{Z_j}^{(i)}, \tilde{\Delta}_{Z_j} \rangle^2} &\leq \sqrt{\sum_{i \notin H_{opt}} \tilde{\Gamma}_i^2 \delta_i^2} \left(\max_j \|X_{Z_j}^{H_{opt}^c}\| \right) \sum_j \|\tilde{\Delta}_{Z_j}\|_2 \\
 &\leq \sqrt{|H_{opt}^c|} \max_{i \notin H_{opt}} |\tilde{\Gamma}_i \delta_i| \left(\max_j \|X_{Z_j}^{H_{opt}^c}\| \right) \sum_j \|\tilde{\Delta}_{Z_j}\|_2,
 \end{aligned}$$

where $X_{Z_j}^{H_{opt}^c}$ denotes $|H_{opt}^c| \times |Z_j|$ sub-matrix of $X_{Z_j}^{H_{opt}^c} \in \mathbb{R}^{|H_{opt}^c| \times |Z_j|}$ corresponding only to indices Z_j .

Applying Assumption 7 and the property of a spectral norm, we have $\max_j \|X_{Z_j}^{H_{opt}^c}\| \leq c_{\max} \sqrt{|B| \log p}$, where $|B|$ is the maximum number of bad samples, and hence,

$$\sqrt{|H_{opt}^c|} \max_{i \notin H_{opt}} |\tilde{\Gamma}_i \delta_i| \left(\max_j \|X_{Z_j}^{H_{opt}^c}\| \right) \sum_j \|\tilde{\Delta}_{Z_j}\|_2 \leq \sqrt{|H_{opt}^c|} \max_{i \notin H_{opt}} |\delta_i| c_{\max} \sqrt{|B| \log p} \sum_j \|\tilde{\Delta}_{Z_j}\|_2$$

Applying the bound discussed in Candes et al. (2006), we have

$$\begin{aligned}
 \sum_{j=1}^q \|\tilde{\Delta}_{Z_j}\|_2 &= \|[\tilde{\Delta}]_{Z_1}\|_2 + \sum_{j=2}^q \|\tilde{\Delta}_{Z_j}\|_2 \leq \|[\tilde{\Delta}]_{Z_1}\|_2 + \frac{1}{\sqrt{|H_{opt}^c|}} \sum_{j=1}^q \|\tilde{\Delta}_{Z_j}\|_1 \\
 &\leq \|\tilde{\Delta}\|_2 + \frac{1}{\sqrt{|H_{opt}^c|}} \|\tilde{\Delta}\|_1
 \end{aligned}$$

Putting all together and applying $|H_{opt}^c| \leq |B|$,

$$\begin{aligned}
 \frac{1}{h} \tilde{\Gamma}_i \left(\langle X_S^{(i)}, \theta^* + \tilde{\Delta} \rangle - X_j^{(i)} \right) \langle X_S^{(i)}, \tilde{\Delta} \rangle &\geq -\frac{1}{h} \sum_{i=1}^n \tilde{\Gamma}_i (\epsilon_i + \delta_i) \langle X_S^{(i)}, \tilde{\Delta} \rangle \\
 &\geq -|B| \max_{i \notin H_{opt}} |\delta_i| c_{\max} \sqrt{\log p} \left(\|\tilde{\Delta}\|_2 + \frac{1}{\sqrt{|B|}} \|\tilde{\Delta}\|_1 \right).
 \end{aligned}$$

Hence, Condition (16) holds with functions,

$$\begin{aligned}\tau_2(n, p) &= c_{\max} \max_{i \notin H_{opt}} |\delta_i| \frac{|B| \sqrt{\log p}}{h} \\ \tau_3(n, p) &= c_{\max} \max_{i \notin H_{opt}} |\delta_i| \frac{\sqrt{|B| \log p}}{h}.\end{aligned}$$

■

Appendix E. Proof for Lemma 20

Proof In the Gaussian linear SEM setting, simple algebra yields that

$$\|\nabla_{\theta} \mathcal{L}(\theta^*, \omega^*)\|_{\infty} = \left\| \frac{1}{h} \sum_{i=1}^n w_i^* (\langle X_S^{(i)}, \theta^* \rangle - X_j^{(i)}) X_S^{(i)} \right\|_{\infty} = \left\| \frac{1}{h} \sum_{i \in H_{opt}} w_i^* \epsilon_i X_S^{(i)} \right\|_{\infty}.$$

By the Gaussian property: for any fixed vector v such that $\|v\|_2 = 1$, and for all $t > 0$,

$$P(|\langle v, \epsilon \rangle|) \leq 2 \exp\left(-\frac{t^2}{2\sigma_j^2}\right).$$

Recall that $\omega_i^* = \tilde{\omega}_i$, if $i \in H_{opt}$; otherwise 0. In addition, recall that $\bar{H}_{opt} = \{i \in H_{opt} \mid \tilde{\omega}_i = 1\}$. Then, for any $k \in S$

$$P\left(\left|\frac{1}{h} \sum_{i \in \bar{H}_{opt}} X_k^{(i)} \epsilon_j^{(i)}\right| > t\right) \leq 2 \exp\left(-\frac{ht^2}{2\sigma_j^2 \|X_k^{1:n}\|_2^2}\right).$$

Applying the union bound and $|S| < p$,

$$P\left(\max_{k \in S} \left(\frac{1}{h} \sum_{i \in \bar{H}_{opt}} X_{S_j(r)}^{(i)} \epsilon_j^{(i)}\right) > t\right) \leq 2p \exp\left(-\frac{ht^2}{2\sigma_j^2 \max_{k \in S} \|X_k^{1:n}\|_2^2}\right).$$

Applying $|S_j(r)| < p$, and setting $t^2 = 2a_{\epsilon} \sigma_j^2 \max_k \|X_k^{1:n}\|_2^2 (\frac{\log p}{h})$ for some small $\epsilon > 0$ and $a_{\epsilon} \in (1, \infty)$, we obtain

$$\max_k \left|\frac{1}{h} \sum_{i \in \bar{H}_{opt}} X_k^{(i)} \epsilon_j^{(i)}\right| \leq \sqrt{2} \sqrt{a_{\epsilon}} \sigma_j \max_k \|X_k^{1:n}\|_2 \sqrt{\frac{\log p}{h}}$$

with probability at least $1 - 2p^{1-a_{\epsilon}}$.

Therefore, for $p > 2$ and some sufficient large constant $c_{\epsilon} \geq 4\sqrt{(3 - \log \epsilon/2)} \max_{j \in V} \sigma_j \|X_j^{1:n}\|_2$, we have

$$4\|\nabla_{\theta} \mathcal{L}(\theta^*, \omega^*)\|_{\infty} \leq c_{\epsilon} \lambda$$

with probability at least $1 - \frac{\epsilon}{p^2}$.

For the upper bound of $2\rho\tau_1(n, p) + \tau_3(n, p)$, applying the results in Sections C and D, and Assumption 8,

$$\begin{aligned} 2\rho\tau_1(n, p) + \tau_3(n, p) &= 2\rho \frac{\kappa_2 \log p}{h} + c_{\max} \max_{i \notin H_{opt}} |\delta_i| \frac{\sqrt{|B| \log p}}{h} \\ &\leq 2\kappa_2 \sqrt{\frac{\log p}{h}} + c_{\max} \max_{i \notin H_{opt}} |\delta_i| \frac{\sqrt{|B| \log p}}{h} = \left(2\kappa_2 + c_{\max} \max_{i \notin H_{opt}} |\delta_i| \right) \sqrt{\frac{\log p}{h}}. \end{aligned}$$

Hence, the following result is reached.

$$4(2\rho\tau_1(n, p) + \tau_3(n, p)) \leq \{2\kappa_2 + c_{\max} \max_{i \notin H_{opt}} |\delta_i|\} \sqrt{\frac{\log p}{h}}.$$

Therefore, when $c_\epsilon \geq \max\{4\sqrt{(3 - \log \epsilon/2)} \max_{j \in V} \sigma_j \|X_j^{1:n}\|_2, 2\kappa_2 + c_{\max} \max_{i \notin H_{opt}} |\delta_i|\}$, the proof is completed. \blacksquare

Appendix F. Proof for Theorem 13

Proof This section proves that Algorithm 1 accurately estimates the ordering with high probability, given that $\theta_j^*(r)$ are well estimated from ℓ_1 -regularized LTS. The following proof is analogous with the one developed in Park (2020); Park et al. (2021). Here, the proof is restated in the framework of corrupted Gaussian linear SEMs.

Suppose that $\pi = (\pi_1, \pi_2, \dots, \pi_p) = (1, 2, \dots, p)$ satisfies the backward selection condition in Lemma 1. In addition, for ease of notation, let $\pi_{1:j} = (\pi_1, \pi_2, \dots, \pi_j)$, and omit subscripted j and r in parentheses, as done in Appendix B. Let $\widehat{\text{Var}}(X_j | X_S)$ denote the estimation of truncated conditional variance $\text{Var}(X_j | X_S, |X_j - \mathbb{E}(X_j | X_S)| < \eta)$. Lastly, suppose that $\eta \in (0, \eta_{\min})$, where η_{\min} is the outlier thresholding constant in Assumption 11. Then, applying the robust identifiability condition in Theorem 2, the probability that the ordering is correctly estimated from Algorithm 1 is

$$\Pr(\widehat{\pi} = \pi) = \Pr\left(\min_{\substack{j=2, \dots, p \\ k=1, \dots, j-1}} \widehat{\text{Var}}(X_j | X_{\pi_{1:j} \setminus j}) - \widehat{\text{Var}}(X_k | X_{\pi_{1:j} \setminus k}) > 0\right).$$

Since it can be decomposed into the following two terms, we have

$$\begin{aligned} \Pr(\widehat{\pi} = \pi) &\geq \Pr\left(\min_{\substack{j=2, \dots, p \\ k=1, \dots, j-1}} \left\{ \text{Var}(X_j | X_{\pi_{1:j} \setminus j}) - \text{Var}(X_k | X_{\pi_{1:j} \setminus k}) \right\} > \tau_{\min}, \text{ and} \right. \\ &\quad \left. \max_{\substack{j=2, \dots, p \\ k=1, \dots, j}} \left| \text{Var}(X_k | X_{\pi_{1:j} \setminus k}) - \widehat{\text{Var}}(X_k | X_{\pi_{1:j} \setminus k}) \right| < \frac{\tau_{\min}}{2} \right). \end{aligned}$$

The first term in the above probability is always satisfied because $\min\{\text{Var}(X_j | X_{\pi_{1:j} \setminus j}) - \text{Var}(X_k | X_{\pi_{1:j} \setminus k})\} > \tau_{\min}$ from Assumption 12. Hence, the probability that the ordering is

correctly estimated from Algorithm 1 is reduced to

$$\Pr(\hat{\pi} = \pi) \geq \Pr\left(\max_{\substack{j=2,\dots,p \\ k=1,\dots,j}} \left| \text{Var}(X_k | X_{\pi_{1:j} \setminus k}) - \widehat{\text{Var}}(X_k | X_{\pi_{1:j} \setminus k}) \right| < \frac{\tau_{\min}}{2}\right).$$

Applying the union bound, we have

$$\Pr(\hat{\pi} = \pi) \geq 1 - p^2 \max_{j,k} \Pr\left(\left| \text{Var}(X_k | X_{\pi_{1:j} \setminus k}) - \widehat{\text{Var}}(X_k | X_{\pi_{1:j} \setminus k}) \right| < \frac{\tau_{\min}}{2}\right).$$

Now, the focus is on the consistency rate of the two-stage ℓ_1 -regularization based conditional variance estimator in Equation (4.2.2):

$$\widehat{\text{Var}}(X_j | X_S) = [\widehat{\Sigma}_{S'}^{\widehat{G}}]_{j,j}^{-1}$$

where $\widehat{\Sigma}_{S'}^{\widehat{G}}$ is a sample covariance matrix for $X_{S'}^{\widehat{G}}$, in which $S' := \{j\} \cup \text{Supp}(\theta^*)$ and \widehat{G} is the estimated index set of good observations estimated by the magnitude of residuals using Equation (7).

Since the truncated distribution of X_j given X_S has the (sub-)Gaussian tail property, Lemma 13 of Park et al. (2021) implies that for any positive $\epsilon > 0$, there are some positive constants C_1 and C_2 such that

$$\Pr\left(\left| \widehat{\text{Var}}(X | X_S) - \text{Var}(X | X_S) \right| < \epsilon\right) \geq 1 - C_1 \cdot \exp\left(C_2 \frac{-n}{(d+1)^2}\right).$$

Finally, applying the consistency result of the truncated conditional variance estimator, we complete the proof. For a corrupted Gaussian linear SEM, there are positive constants $C_1, C_2 > 0$ such that

$$\Pr(\hat{\pi} = \pi) > 1 - C_1 p^2 \cdot \exp\left(C_2 \frac{-n}{(d+1)^2}\right).$$

■

References

- Andreas Alfons, Christophe Croux, and Sarah Gelper. Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics*, pages 226–248, 2013.
- Fatemah Alqallaf, Stefan Van Aelst, Victor J Yohai, and Ruben H Zamar. Propagation of outliers in multivariate data. *The Annals of Statistics*, pages 311–331, 2009.
- Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8): 1207–1223, 2006.

- Mengjie Chen, Chao Gao, and Zhao Ren. Robust covariance and scatter matrix estimation under huber’s contamination model. *The Annals of Statistics*, 46(5):1932–1960, 2018.
- Wenyu Chen, Mathias Drton, and Y Samuel Wang. On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980, 2019.
- David L Donoho and Peter J Huber. The notion of breakdown point. *A festschrift for Erich L. Lehmann*, 157184, 1983.
- Marco F Eigenmann, Preetam Nandy, and Marloes H Maathuis. Structure learning of linear gaussian structural equation models with weak edges. *arXiv preprint arXiv:1707.07560*, 2017.
- Asish Ghoshal and Jean Honorio. Learning linear structural equation models in polynomial time and sample complexity. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1466–1475, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR.
- Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992.
- Steffen L Lauritzen. *Graphical models*. Oxford University Press, 1996.
- Po-Ling Loh and Peter Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *The Journal of Machine Learning Research*, 15(1):3065–3105, 2014.
- Po-Ling Loh and Xin Lu Tan. High-dimensional robust precision matrix estimation: Cell-wise corruption under ϵ -contamination. *Electronic Journal of Statistics*, 12(1):1429–1467, 2018.
- Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m -estimators with decomposable regularizers. *Statistical science*, 27(4):538–557, 2012.
- Gunwoong Park. Identifiability of additive noise models using conditional variances. *Journal of Machine Learning Research*, 21(75):1–34, 2020.
- Gunwoong Park and Yesool Kim. Learning high-dimensional gaussian linear structural equation models with heterogeneous error variances. *Computational Statistics & Data Analysis*, 154:107084, 2021.
- Gunwoong Park and Youngwhan Kim. Identifiability of gaussian linear structural equation models with homogeneous and heterogeneous error variances. *Journal of the Korean Statistical Society*, 49(1):276–292, 2020.
- Gunwoong Park, Sang Jun Moon, Sion Park, and Jong-June Jeon. Learning a high-dimensional linear structural equation model via l1-regularized regression. *Journal of Machine Learning Research*, 22(102):1–41, 2021.

- Jonas Peters and Peter Bühlmann. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.
- Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.
- Basil Saeed, Anastasiya Belyaeva, Yuhao Wang, and Caroline Uhler. Anchored causal inference in the presence of measurement error. In *Conference on Uncertainty in Artificial Intelligence*, pages 619–628. PMLR, 2020.
- Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.
- Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.
- Yuhao Wang, Santiago Segarra, and Caroline Uhler. High-dimensional joint estimation of multiple directed gaussian graphical models. *Electronic Journal of Statistics*, 14(1):2439–2483, 2020.
- Marten Wegkamp and Yue Zhao. Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. *Bernoulli*, 22(2):1184–1226, 2016.
- Lingzhou Xue and Hui Zou. Regularized rank-based estimation of high-dimensional non-paranormal graphical models. *The Annals of Statistics*, 40(5):2541–2571, 2012.
- Eunho Yang, Aurélie C Lozano, and Aleksandr Aravkin. A general family of trimmed estimators for robust high-dimensional data analysis. *Electronic Journal of Statistics*, 12(2):3519–3553, 2018.
- Kun Zhang, Mingming Gong, Joseph Ramsey, Kayhan Batmanghelich, Peter Spirtes, and Clark Glymour. Causal discovery in the presence of measurement error: Identifiability conditions. *arXiv preprint arXiv:1706.03768*, 2017.