Generalized Criterion for Identifiability of Additive Noise Models Using Majorization

Aramayis Dallakyan
StataCorp LLC

Yang Ni
Department of Statistics
Texas A&M University

Abstract

The discovery of causal relationships from observational data is very challenging. Many recent approaches rely on complexity or uncertainty concepts to impose constraints on probability distributions, aiming to identify specific classes of directed acyclic graph (DAG) models. In this paper, we introduce a novel identifiability criterion for DAGs that places constraints on the conditional variances of additive noise models. We demonstrate that this criterion extends and generalizes existing identifiability criteria in the literature that employ (conditional) variances as measures of uncertainty in (conditional) distributions. For linear structural equation models, we present a new algorithm that leverages the concept of weak majorization applied to the diagonal elements of the Cholesky factor of the covariance matrix to learn a topological ordering of variables. Through extensive simulations and the analysis of bank connectivity data, we provide evidence of the effectiveness of our approach in successfully recovering DAGs. The code for reproducing the results in this paper is available in Supplementary Materials.

1 INTRODUCTION

One of the fundamental problems in science is learning causal relations from observational data. Directed acyclic graphical (DAG) models serve as valuable tools for representing conditional independence and causal relations among random variables. Nevertheless, learning DAGs solely from the observational data proves to be a difficult problem. This is primarily due to issues related to identifiability and the space of the potential DAGs growing super-exponentially as the number of nodes increases. A commonly used strategy for addressing the non-identifiability issue is to impose restrictions on the joint distribution. For instance, methods like PC (Spirtes et al., 2001), GES (Chickering, 2003) and other related methods (Zhang & Spirtes, 2015; Raskutti & Uhler, 2013) have demonstrated that, under certain assumptions such as faithfulness, it is possible to recover DAG up to the Markov equivalence class. However, the majority of Markov equivalence classes contain more than one graph, making it difficult to uniquely determine the true underlying graph.

Recent research has addressed the challenge of recovering causal structures by introducing various constraints aimed at capturing the asymmetry between cause and effect. Shimizu et al. (2006) propose a method for identifying linear non-Gaussian additive noise models (ANM). Hoyer et al. (2008); Zhang & Hyvärinen (2009); Peters et al. (2011) establish conditions for the identifiability of nonlinear ANMs. Park & Raskutti (2018); Park & Park (2019) employ higher-order moments of the conditional distribution to demonstrate the identifiability of DAGs. The ideas behind these approaches can be succinctly summarized using the concept of complexity or uncertainty (Mooij et al., 2016; Glymour et al., 2019). In particular, it is generally anticipated that the complexity of a physical process that generates effect from cause should be lower in some way than the complexity of the backward process. In other words, the relationship between two variables in a model is asymmetrical rather than symmetrical (Simon, 1977). These ideas have been explored using Kolmogorov complexity in Janzing & Schölkopf (2010). Also see Peters et al. (2017). The aforementioned methods harness various complexity measures for both marginal and conditional probability distributions to achieve identifiability.

In a similar vein, several studies (Peters & Bühlmann, 2013; Loh & Bühlmann, 2014; Ghoshal & Honorio, 2018;

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

Chen et al., 2019; Park, 2020) have employed (conditional) variance as a complexity measure to establish identifiability results. Notably, Peters & Bühlmann (2013) demonstrate the identifiability of linear structural equation models (SEMs) with equal variances. Ghoshal & Honorio (2017, 2018); Chen et al. (2019) note that the arrangement of specific conditional variances implies the identifiability of DAGs. Furthermore, Park & Kim (2020); Park (2020) introduce more comprehensive criteria concerning conditional variances for nonlinear ANMs with unknown heterogeneous error variances. Gao et al. (2020) relies on a similar criterion as Park (2020) to learn a nonlinear, nonparametric DAG. Gao & Aragam (2021) generalizes the latter results by imposing an identifiability criterion using entropy.

In this paper, we establish the identifiability of ANMs by leveraging the concept of majorization (see Section 2 for definitions). We prove that our approach constitutes a generalization of previous identifiability results that employ conditional variance as a complexity measure (Peters & Bühlmann, 2013; Rajaratnam & Salzman, 2013; Loh & Bühlmann, 2014; Ghoshal & Honorio, 2017, 2018; Chen et al., 2019; Park, 2020). Additionally, we introduce a novel DAG structure learning algorithm called **Ma**jorized **Cho**lesky diagonal (MaCho), which utilizes our established criterion to estimate a topological ordering of a DAG.

2 PRELIMINARIES

A DAG $G = (V, E)$ consists of a set of nodes $V = \{1, 2, \ldots, p\}$ and a set of directed edges $E \subset V \times V$ with no directed cycles. For a quick introduction to DAG models (also known as Bayesian networks) and related terminology, see Appendix I and Koller & Friedman (2009). Throughout the paper, we assume causal sufficiency, i.e. no hidden confounders, and causal minimality.

ANMs are a special case of DAG models in which, for the mean zero random vector $\boldsymbol{X} \in \mathbb{R}^p$, the joint distribution is defined by the following structural equation:

$$X_j = f_j(X_{\mathrm{pa}_j}) + \varepsilon_j, \tag{1}$$

where ε_j's are independent for $j \in V$ with possibly different distributions with mean zero and heterogeneous variances and pa_j denotes the parents of the variable j. A linear SEM is a special case of (1), in which f_j's are linear. In vector formulation, linear SEM can be written as

$$\boldsymbol{X} = B\boldsymbol{X} + \boldsymbol{\varepsilon}, \tag{2}$$

where $B \in \mathbb{R}^{p \times p}$ is a weighted adjacency matrix such that $\beta_{ij} \neq 0$ indicates $X_j \to X_i$. From (2), the covariance matrix $\Sigma = (I - B)^{-1}\Lambda[(I - B)^{-1}]^{'}$ and Λ is a diagonal matrix that contains error variances. We say a DAG admits a **topological ordering** $\rho(\cdot)$, if $\rho(j) < \rho(k)$, then k is not an ancestor of j, i.e. $k \notin \mathrm{An}(j)$, and a $p \times p$ permutation matrix P_ρ can be associated such that $P_\rho \boldsymbol{x} = (x_{\rho(1)}, \ldots, x_{\rho(p)})$, for $\boldsymbol{x} \in \mathbb{R}^p$. It is important to note that to learn a DAG G, it suffices to learn its topological ordering. In particular, given an ordering ρ, existing variable selection methods can be used to learn the parent set pa and hence the graph G. For more details, see Gao et al. (2020, Appendix A) and Section 5.2.

The existence of a topological ordering leads to the permutation-similarity of B to a strictly lower triangular matrix $B_\rho = P_\rho B P_\rho^{'}$ by permuting rows and columns of B simultaneously (Bollen, 1989) (see Figure 6 in Appendix I for an illustrative example). Therefore,

$$\Sigma_\rho = (I - B_\rho)^{-1}\Lambda_\rho[(I - B_\rho)^{-1}]^{'} = LL^{'}, \tag{3}$$

where $L = (I - B_\rho)^{-1}\Lambda_\rho^{1/2}$ is a Cholesky factor of Σ_ρ.

The following Lemma will be useful in Section 5. The proof is provided in Appendix A for completeness. Also see Rajaratnam & Salzman (2013), for a similar result for autoregressive processes.

Lemma 2.1. *Suppose a random vector $\boldsymbol{X}$ generated from the linear SEM (2) with an ordering ρ, then for $1 \leq j \leq p$*

$$L_{j,j}^2 = var(X_{\rho(j)} | X_{\rho(j-1)}, \ldots, X_{\rho(1)}) \tag{4}$$

2.1 Majorization

The notion of majorization arises in a variety of branches of mathematics and statistics. Informally, majorization describes the notion that the components of a vector $\boldsymbol{x} \in \mathbb{R}^p$ are less spread out or more nearly equal than the components of a vector $\boldsymbol{y} \in \mathbb{R}^p$.

Formally, let $\boldsymbol{x}_{[\cdot]}$ denote the vector x of a descending order, then we say $\boldsymbol{x}$ is **majorized** by $\boldsymbol{y}$, denoted $\boldsymbol{x} \prec \boldsymbol{y}$, if $\sum_{i=1}^k x_{[i]} \leq \sum_{i=1}^k y_{[i]}$, for $k = 1, \ldots, p-1$, and $\sum_{i=1}^p x_{[i]} = \sum_{i=1}^p y_{[i]}$. If $\sum_{i=1}^p x_{[i]} = \sum_{i=1}^p y_{[i]}$ does not hold, but $\sum_{i=1}^p x_{[i]} \leq \sum_{i=1}^p y_{[i]}$, then we say $\boldsymbol{y}$ **weakly majorizes** $\boldsymbol{x}$, i.e., $\boldsymbol{x} \prec_w \boldsymbol{y}$. (Weak) majorization defines a preordering on $\mathbb{R}^p$.

Definition 2.2. A linear transformation is called a T-transform, if

$$T = \lambda I + (1 - \lambda)Q,$$

where $0 \leq \lambda \leq 1$ and Q is a permutation matrix that interchanges two coordinates.

That is $T\boldsymbol{x} = (x_1, \ldots, x_{j-1}, \lambda x_j + (1-\lambda)x_k, x_{j+1}, \ldots, \lambda x_k + (1-\lambda)x_j, x_{k+1}, \ldots, x_p)'$.

Theorem 2.3. *(Hardy et al. (1952)) For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$, the following conditions are equivalent*

1. $\boldsymbol{x} \prec_w \boldsymbol{y}$
2. $\boldsymbol{x} = P\boldsymbol{y}$ *for some doubly substochastic matrix* P
3. $\boldsymbol{x}$ *can be derived from* $\boldsymbol{y}$ *by successive applications of a finite number of* T*-transformations,* $\boldsymbol{x} \leq T_1 \ldots T_k \boldsymbol{y}$

The next theorem will be useful to proving identification results in Section 3.

Theorem 2.4. *(Marshall et al. (2011, Theorems 3.A.8a and 3.C.2.D)) Let f be a real-valued function defined on the $\boldsymbol{a} \in \mathbb{R}^p$. Then*

$$\boldsymbol{x} \prec_w \boldsymbol{y} \implies f(\boldsymbol{x}) < f(\boldsymbol{y})$$

if and only if $f(\boldsymbol{a})$ is strictly increasing, symmetric, and strictly convex in $\boldsymbol{a}$.

3 IDENTIFIABILITY via MAJORIZATION

We denote a topological ordering of the graph G by ρ_0. Without loss of generality, we assume $\rho_0 = \{1, \ldots, p\}$ and use ρ to represent any other ordering. This notation simplification helps avoid the need for the longer notation $\rho_0(1), \ldots, \rho_0(p)$.

As we discussed in Section 1, our focus is on identifiability conditions whereas a complexity measure, the (conditional) variance, has been employed. Recall that the concept of complexity or uncertainty implies that it is expected that the physical process that is generated using a topological ordering of variables should be "simpler" than the process generated using other non-topological orderings. As a measure of the "simplicity" of the process and to capture the asymmetry between cause and effect, we impose a weak majorization assumption. This means if the vector $\boldsymbol{x}$ contains the conditional variances using a topological ordering

$$\boldsymbol{x} = \{\text{var}(X_1), \ldots, E[\text{var}(X_p | X_1, \ldots, X_{p-1})]\}' \quad (5)$$

and $\boldsymbol{y}_\rho$ contains conditional variances using any other ordering

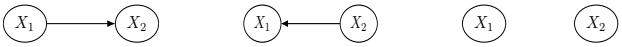
$$\boldsymbol{y}_\rho = \{\text{var}(X_{\rho(1)}), \ldots, E[\text{var}(X_{\rho(p)} | X_{\rho(1)}, \ldots, X_{\rho(p-1)})]\}', \quad (6)$$

then we anticipate $\boldsymbol{x}$ be less spread out then $\boldsymbol{y}_\rho$. Using language introduced in Section 2.1, $\boldsymbol{x} \prec_w \boldsymbol{y}_\rho$ for all permutations $\rho \in \mathfrak{S}_p$, where $\mathfrak{S}_p$ denotes the group of all permutations $\boldsymbol{X}$. To make it formal, we state the main assumption of this paper.

Assumption 1. Let $\boldsymbol{X}$ be generated from an ANM (1) with DAG G, then assume $\boldsymbol{x} \prec_w \boldsymbol{y}_\rho$ for all permutations $\rho \in \mathfrak{S}_p$, where $\boldsymbol{x}$ and $\boldsymbol{y}_\rho$ are defined in (5) and (6), respectively.

Remark 3.1. The notion of varsortability, introduced by Reisach et al. (2021), serves as a measure of agreement between the order of increasing marginal variance and the causal order. In Appendix H, we demonstrate that Assumption 1 holds even when the varsortability is low.

To provide an intuition how the Assumption 1 implies identification, we provide examples on bivariate ANMs generated from the following three DAGs:



(a) G_1 (b) G_2 (c) G_3

From Theorem 2.4, the following identification result holds:

$$\text{DAG} := \begin{cases} G_1, & \text{if } f(\boldsymbol{x}) < f(\boldsymbol{y}_\rho) \\ G_2, & \text{if } f(\boldsymbol{x}) > f(\boldsymbol{y}_\rho) \\ G_3, & \text{otherwise} \end{cases}, \quad (7)$$

where $\boldsymbol{x} = [\text{var}(X_1), E[\text{var}(X_2|X_1)]]'$, $\boldsymbol{y}_\rho = [\text{var}(X_2), E[\text{var}(X_2|X_1)]]'$ and

$$f(\boldsymbol{x}) = \|\boldsymbol{x}\|_2, \quad (8)$$

is the Euclidean norm of $\boldsymbol{x}$, which will be used throughout this paper. For other possible options for f, see Marshall et al. (2011). Empirical results, for bivariate case is provided in Section 6. The next theorem generalizes the discussed bivariate case. Proof is provided in Appendix B.

Theorem 3.2. *If the Assumption 1 holds then DAG G is uniquely identified.*

4 WEAK MAJORIZATION AS A GENERALIZED CONDITION

In this section, we show that the weak majorization, Assumption 1, is a generalization of the existing identification conditions for nonlinear ANM (Park, 2020), linear SEM (Peters & Bühlmann, 2013; Loh & Bühlmann, 2014; Ghoshal & Honorio, 2018; Chen et al., 2019; Park, 2020), and autoregressive processes (Rajaratnam & Salzman, 2013).

For the nonlinear ANMs, defined in (1), Hoyer et al. (2008); Mooij et al. (2009); Peters et al. (2011) prove the identifiable classes of ANM by imposing restrictions on the functions f_j or the joint distribution. Park (2020) used conditional independencies as an uncertainty measure to propose conditions listed in Assumption 2, for identification.

Assumption 2. Let $\boldsymbol{X}$ be generated from an ANM (1), then assume for $k \in \mathrm{De}(j)$ and $l \in \mathrm{An}(j)$

a. $\sigma_j^2 < \sigma_k^2 + E(\mathrm{var}(E(X_k|X_{\mathrm{pa}_k})|X_1, \ldots, X_{j-1}))$

b. $\sigma_j^2 > \sigma_l^2 - E(\mathrm{var}(E(X_l|X_1, \ldots X_{j-1} \backslash X_l)|X_{\mathrm{pa}_l}))$

Moreover, Park (2020, Theorem 4) shows that Assumption 2 is a generalization of conditions proposed in Peters & Bühlmann (2013); Loh & Bühlmann (2014); Ghoshal & Honorio (2018); Chen et al. (2019). Consequently, our goal is to show that Assumption 2 implies the weakly majorization of conditional variances, given in Assumption 1. Proof of the next theorem is provided in Appendix C.

Theorem 4.1. *If the conditions of Assumption 2 are satisfied, then Assumption 1 is also satisfied.*

4.1 Assumption 2 Is a Special Case Of Assumption 1

In Theorem 4.1 we showed that if Park (2020) identifiability conditions (Assumption 2) are satisfied then Assumption 1 is also satisfied. In this section, through a toy example we show that those assumptions are not equivalent, i.e. it is possible that the former is violated but the later is still satisfied.

Consider the following data generation process

$$X_1 := \varepsilon_1;\ X_2 := \beta_{12} X_1 + \varepsilon_2;\ X_3 := \beta_{13} X_1 + \varepsilon_3, \quad (9)$$

where $\sigma_1^2 = 4$, $\sigma_2^2 = 3$, $\sigma_3^2 = 1$, $\beta_{12} = -0.5$ and $\beta_{13} = 0.9$. That is $x = \{4, 3, 1\}$, where x is defined as in (5). One of the Park (2020) assumptions narrows down to $\sigma_1^2 < \sigma_2^2 + \beta_{12}^2 \sigma_1^2$. However, $4 \not< 3 + 0.25 \times 4$.

To show that the Assumption 1 holds, instead of computing the conditional variances for all possible permutations and checking the weak majorization, we rely on procedure introduced in Section D.1 to check that the weak majorization assumption indeed holds for all possible permutation. Table 1 reports vectors y_ρ, where for each permutation $\rho \in \mathfrak{S}_p$, y_ρ is defined as in (22). As can be seen, from Proposition D.1, for each ρ, $x \prec_w y_\rho$ is true.

ρ	y_ρ
{1,3,2}	{4,3,1}
{3,1,2}	{4.24,3,0.94}
{3,2,1}	{4.24,3.23,0.87}
{2,3,1}	{4,3.43,0.87}
{2,1,3}	{4,3,1}

Table 1: Estimated vectors y_ρ for the toy example.

5 ALGORITHM

The proof of Theorem 3.2 provides a recipe for learning topological ordering of a DAG. In particular, we can exploit Theorem 2.4 to compare the estimated value of (8) for all permutation $\rho \in \mathfrak{S}_p$ and choosing the one that obtains a minimum value

$$\min_\rho f(y_\rho).$$

Unfortunately, for large p this strategy is computationally infeasible, since $\mathfrak{S}_p$ is large and it requires comparing values of $p!$ permutations.

In this section, we propose an algorithm, named **Ma**jorized **Cho**lesky Diagonal (MaCho), that learns the topological ordering of the DAG generated from the linear SEM (2) with $O(p^3)$ computational cost. The intuition behind the proposed algorithm can be explained by first assuming that the index of the first variable is known, and then considering how the algorithm will learn the rest of the ordering on the Cholesky factor of the population covariance matrix. The MaCho algorithm proceeds recursively as follows. Given an already recovered ordering of the unknown permutation ρ, denoted $\rho_m = \{\rho(1), \rho(2), \ldots, \rho(m)\}$ of size $m < p$, the algorithm chooses the next index $i_{\rho(m+1)}$ to minimize the ℓ_2 norm of the diagonal of Cholesky factor

$$f(\mathrm{dichol}[\Sigma_{\rho_m \cup \rho(m+1), \rho_m \cup \rho(m+1)}]) = \sqrt{L^2_{\rho(1),\rho(1)} + L^2_{\rho(2),\rho(2)} + \cdots + L^2_{\rho(m+1),\rho(m+1)}}, \quad (10)$$

where $\mathrm{dichol}(A) = \mathrm{diag}(\mathrm{Cholesky}[A])$ denotes an operator that extracts the diagonal values of the Cholesky factor and Σ_{ρ_m,ρ_m} denote the submatrix of Σ with corresponding indices $\rho_m = \{\rho(1), \rho(2), \ldots, \rho(m)\}$. Lemma 2.1 and (8) provide a mathematical justification for (10).

When the index of the first variable is unknown, in each iteration we check if

$$f(\mathrm{dichol}[\Sigma_{\rho_m \cup \rho(m+1), \rho_m \cup \rho(m+1)}]) \leq f(\mathrm{dichol}[\Sigma_{\rho(m+1) \cup \rho_m, \rho(m+1) \cup \rho_m}]) \quad (11)$$

then we add $\rho(m+1)$ at the end of ρ_m, otherwise in front of ρ_m and $\rho(m+1)$ becomes the first index. Algorithm 1 summarizes the steps.

In practice, when the number of observations is greater than the dimension of variables, i.e., $n > p$, we suggest recovering the variable ordering using a modified version of the Cholesky factor of the sample covariance matrix $S = \hat{L}\hat{L}'$, which is asymptotically unbiased estimator of the population Cholesky factor. The modified version, is defined as $\tilde{L} = \hat{L}\Omega$, where

$$\Omega = \mathrm{diag}(1/n, \ldots, 1/n - j + 1, \ldots, 1/n - p + 1).$$

Algorithm 1 MaCho

Input: $p \times p$ sample or covariance matrix S
Select any index as $\rho(1)$.
Set $\rho_1 = \{\rho(1)\}$
for $i = 2$ **to** p **do**
  Choose the next index

$$\rho(i) = \min_{i,\ldots,p} f(\text{dichol}[S_{\rho_{i-1}\cup\rho(i),\rho_{i-1}\cup\rho(i)}])$$

  if (11) is TRUE **then**
    $\rho_i = \{\rho_{i-1}, \rho(i)\}$
  else
    $\rho_i = \{\rho(i), \rho_{i-1}\}$
  end if
end for
Output: ρ_p

The proof of the next lemma is relegated to Appendix F.

Lemma 5.1. *Let $\boldsymbol{X} \sim N_p(0, \Sigma)$. As $n \to \infty$, $E(\tilde{L}) = L$ with probability tending to 1.*

For high dimensional dataset, when $n < p$, the covariance matrix can be estimated using a well known sparse covariance matrix estimation methods such as Bien & Tibshirani (2011); Wang (2014). Throughout the paper, when $n < p$, we adopt covariance graphical lasso algorithm proposed in Wang (2014) and implemented in R package `covglasso` available via CRAN.

5.1 Computational Details

One of the main computational costs in Algorithm 1 is in the computation of $\min_{i,\ldots,p} f(\text{dichol}[S_{\rho_{i-1}\cup\rho(i),\rho_{i-1}\cup\rho(i)}])$ in each iteration i. Fortunately, there is no need to implement Cholesky factorization on the $i \times i$ matrix in each iteration i. The new Cholesky factor can be found utilizing the estimated $(i-1)\times(i-1)$ Cholesky factor $L^{(i-1)}$ from iteration $i-1$. In particular, the new added row of $L^{(i)}$ can be computed Golub & Van Loan (1996), for $j = 1$, $L^{(i)}_{i,j} = S_{i,j}/L^{(i-1)}_{j,j}$, for $j = 2, \ldots, i-1$, $L^{(i)}_{i,j} = 1/L^{(i-1)}(S_{i,j} - \sum_{k=1}^{j-1} L^{(i)}_{i,k} * L^{(i-1)}_{j,k})$ and for $j = i$, $L_{i,i} = \sqrt{S_{i,i} - \sum_{k=1}^{i-1}(L^{(i)})^2_{i,k}}$ This requires $O(i^2)$ flops in each iteration i. The next computational burden is in computing (11). Again this can be computed in $O(i^2)$ flops, using the Cholesky factorization update algorithm in Davis & Hager (2005) and Golub & Van Loan (1996, Chapter 12.5). From the above discussion, the following lemma easily follows.

Lemma 5.2. *The Algorithm 1 has a computational cost $O(p^3)$.*

5.2 Estimation Algorithm

Algorithm 1 provides a procedure to estimate a topological ordering of the DAG. Given the ordering, there exist a rich literature on estimating the DAG structure, such as Shojaie & Michailidis (2010); Khare et al. (2019); Park (2020). Motivated by the encouraging results of thresholding estimators for the (inverse) covariance matrix Cai et al. (2016); Wang & Allen (2022), we propose to use the thresholded version of convex sparse Cholesky selection (CSCS) algorithm introduced in Khare et al. (2019). CSCS minimizes the following objective function

$$\text{tr}(LL^{'}S) - 2\log|L| + \lambda \sum_{1\le j<i\le p} |L_{ij}| \qquad (12)$$

Then our final estimator is

$$\hat{L}^{\tau} = \hat{L}1(|\hat{L}| > \tau), \qquad (13)$$

where $1(\cdot)$ is the indicator function, and $\hat{L}$ is a CSCS estimator. The next theorem shows the sign consistency of thresholded CSCS under the Assumption A1 - A4, listed in Appendix G:

Theorem 5.3. *Let the Assumptions 1-3, listed in Appendix G, be satisfied. Further, if Assumption 4 is satisfied with $c_1 > 2c_2$, where c_2 is defined in Lemma G.1, the thresholded CSCS estimate $\hat{L}^{\tau}$ with threshold level $\tau = c_2\sqrt{\frac{s\log p}{n}}$ satisfies:*

$$sign(\hat{L}^{\tau}) = sign(L_{ij}),\ \forall i \neq j \qquad (14)$$

with probability tending to 1.

The proof is provided in Appendix G. Theorem 5.3 provides an additional support to our usage of thresholded CSCS, since the estimated model with reversed signs can be misleading and hardly qualifies as a correctly selected model Zhao & Yu (2006).

6 NUMERICAL RESULTS

In this section, we assess the performance of the proposed criterion and the algorithm using various simulations.

6.1 Bivariate Non-linear ANM

We start our analysis by considering the bivariate case, as in (7), while incorporating heterogeneous error variances using both Gaussian and non-Gaussian distributions. Similar to Park (2020), as a non-linear ANM, we select a polynomial SEM where each variable is modeled as a 5th-degree polynomial in relation to its

parents. In the Gaussian scenario, error variances are randomly chosen from the range $\sigma_j^2 \in [0.7, 1.2]$, while in the non-Gaussian case, we employ Gaussian mixture models. For each of these cases, we generate 100 sets of samples, each with the number of observations set at $n = 500$.

We compare our results with the method proposed in Hoyer et al. (2008), referred to as NCD, and utilize the `Causal Discovery` package in Python (Kalainathan & Goudet, 2019). In our method, the non-linear ANM is estimated using random forest Breiman (2001). We generate causal effect relationships, such as $x \to y$, $x \leftarrow y$, and no effect, with probabilities of $\{0.4, 0.4, 0.2\}$, respectively. From (7), given that achieving the exact equality for the "no-effect" scenario is difficult, we introduce a tolerance measure denoted as μ, defined as: $\mu = |f(x) - f(y)| / \max(f(x), f(y))$, This measure quantifies the allowed relative difference in magnitude that still qualifies as the absence of a causal effect. Based on the simulation results, we recommend $\mu = 0.1$. A similar measure is also defined for the NCD method for the sake of comparison.

Table 2 reports the accuracy of both methods, where accuracy indicates the correct identification of the direction of the causal effect or the correct identification of a "no-effect". As evident from the table, our method outperforms NCD for both Gaussian and non-Gaussian cases.

Table 2: Estimated accuracy for the bivariate case

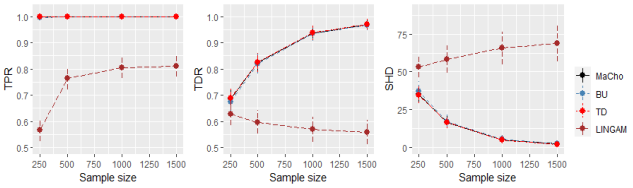
Method	Gaussian	Non-Gaussian
Proposed	**0.95**	**0.93**
NCD	0.91	0.90

6.2 Multivariate Linear SEM

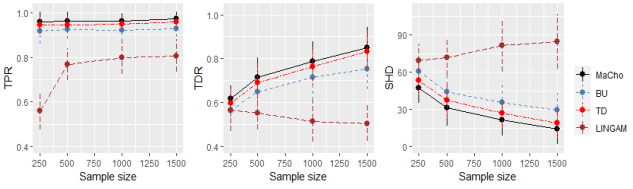
For the multivariate linear SEM, we explore three distinct scenarios: 1) Gaussian homogeneity, 2) Gaussian heterogeneity, and 3) non-Gaussian heterogeneity. In all three scenarios, we generate data based on the SEM described in Equation (2).

In the case of Gaussian homogeneity, we set all noise variances equal to $\sigma^2 = 0.7$. For the Gaussian heterogeneity scenario, error variances are randomly selected from the range $\sigma_j^2 \in [0.7, 1.7]$. Finally, in Case 3, data are generated from a Gaussian mixture.

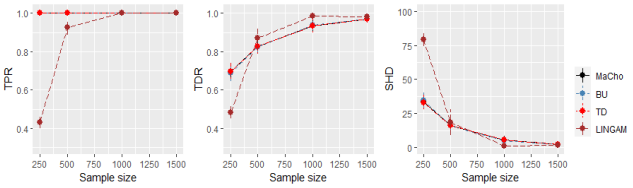
We compare the performance of our proposed MaCho algorithm with that of the bottom-up (BU) method Ghoshal & Honorio (2017), the top-down (TD) method Chen et al. (2019), and the LINGAM algorithm Shimizu et al. (2006). For BU and TD algorithms, we employ the `R` package `EqVarDAG`, and for LINGAM, we utilize `pcalg`. Both packages are available via CRAN. For BU and TD, we select the hyperparameter λ through a 10-fold cross-validation and for MaCho we use procedure described in Wang & Allen (2022, Section 3.4).



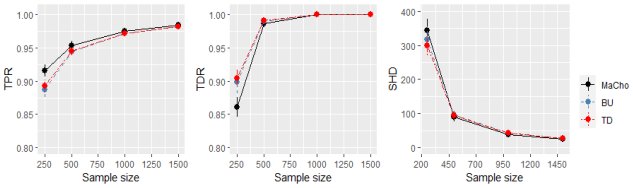
(a) Case 1, p = 20



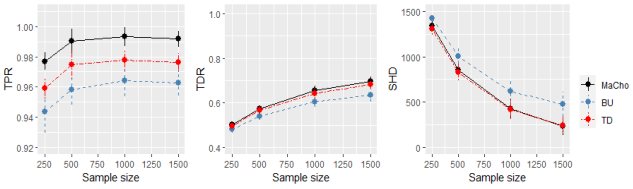
(b) Case 2, p = 20



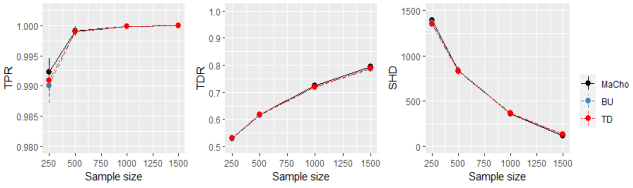
(c) Case 3, p = 20



(d) Case 1, p = 200



(e) Case 2, p = 200



(f) Case 3, p = 200

Figure 2: Comparison of the proposed algorithm MaCho, BU, TD, and LINGAM algorithms in terms of average TPR, TDR, and SHD for recovering linear SEM with different error variances and $p \in \{20, 200\}$.

We vary the sample size $n \in \{250, 500, 1000, 1500\}$ and the dimensionality $p \in \{20, 50, 200\}$. The performance is assessed by the true positive rate (TPR), true discovery rate (TDR), and structural Hamming distance (SHD). We repeat each simulation 100 times.

Figure 2(a)-(c) presents the results. It is evident that LINGAM performs the least effectively in both Case 1 and Case 2. In Case 3, LINGAM is similar to the other algorithms when the sample size n exceeds 1000. Overall, the performance of MaCho, BU, and TD are quite comparable. For a more detailed illustration of their performance (excluding LINGAM), please refer to Figure 10 in the Appendix. The figure reveals that in Case 2, the performance of MaCho slightly outperforms BU and TD across all three evaluation metrics.

In Figure 2(d)-(f), we present results for $p = 200$ without the LINGAM algorithm. It is evident that MaCho achieves the highest TPR in all three cases. In terms of the other metrics, the performance of MaCho and TD is very close for all cases. Appendix J.2 contains results for $p = 50$.

6.3 Case Study: Bank Network Connectedness

We employ the MaCho algorithm to detect the global bank network connectedness. The original dataset (Demirer et al., 2018) comprises 96 banks from 29 developed and emerging economies (countries), spanning the period from September 12, 2003, to February 7, 2014. To enhance clarity, we narrow our focus to economies with more than four banks, resulting in a selection of 54 banks (for further details, please refer to Demirer et al. (2018)). To analyze the data, we adopt a commonly used rolling window procedure Xue et al. (2012); Basu et al. (2023) estimating the model using a 3-year rolling window.

Our investigation into bank connectedness involves studying the evolution of bank networks through the average number of direct neighbors. Figure 3 depicts the average degree of the network over 3-year rolling windows. The plot clearly demonstrates that the connectedness, as measured by the count of neighbors, increases both before and during significant systematic events. Notably, we observe several prominent cycles in Figure 3, with three key events marked, the first two corresponding to the failures of Bear Stearns and Lehman Brothers, which marked the financial crisis of 2007-2009. The final event corresponds to the European Bailout. Our findings align with existing results in the literature, including those in Basu et al. (2023, Section 4) and Billio et al. (2012). For comparative analysis, please refer to Appendix J.1, where we present the results for LINGAM, which show no clear pattern.

In Figure 4, we present DAGs corresponding to bank connectedness both before and after the Lehman Brothers' failure. In this figure, the colors of the nodes represent the respective countries to which the banks belong. As anticipated, the DAG estimated after the Lehman Brothers' failure exhibits a notably denser structure.

Another intriguing question pertains to how bank connectedness evolved within individual countries. To address this, we treat each country as a distinct component and estimate both intra- and inter-connectedness. In Figure 9, located in Appendix J.1, the red dotted line represents the average degrees within banks in the USA and Italy. The black solid and blue dashed lines correspond to the estimated average degrees of in-edges and out-edges of the country as a component. In this context, in-edges signify changes within the country's causal relationships, while out-edges signify how banks in the country influence banks in other countries. A visible contrast emerges in the evolution of network connectedness between these two countries. In the USA, all three lines exhibit similar behavior, indicating that connectedness tends to grow before and during significant events. However, in Italy, the growth of within connectedness occurs much more gradually when compared to the average degree of out-edges.

7 CONCLUSION AND DISCUSSION

In this paper, we have introduced a novel identifiability condition for ANMs with heterogeneous error variances. Our work demonstrates that ANMs can achieve identifiability without the need to assume linearity and Gaussianity. For linear SEMs, we have presented the MaCho algorithm, which is designed for learning the topological ordering of DAGs. Our findings are supported by various numerical experiments, which empirically confirm that MaCho successfully learns the true topological ordering.

One prominent area for future research involves exploring the testing of the weakly majorization condition for ANMs when the dimensionality p is large. We have demonstrated that empirical testing of the weakly majorization assumption is feasible for smaller dimensions, but the computational costs become prohibitively high as p increases.

References

Basu, S., Das, S., Michailidis, G., and Purnanandam, A. A high-dimensional approach to measure connectivity in the financial sector. The Institute of Mathematical Statistics, 2023.

Bertsekas, D. P. *Convex Optimization Algorithms.* Athena Scientific, 2015.

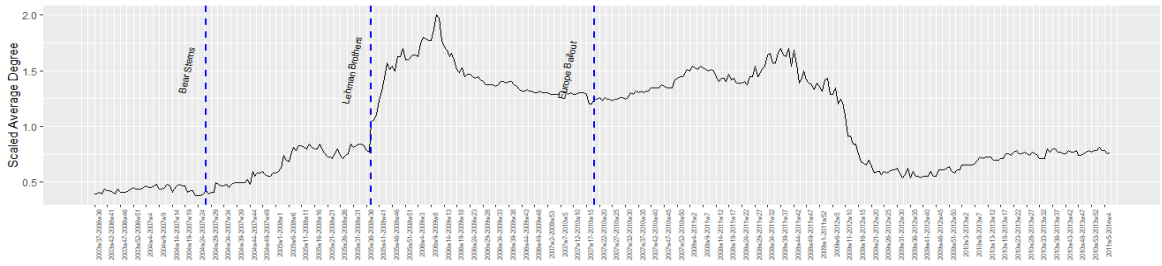


Figure 3: Evolution of average degree of bank connectedness scaled by their historic average (over 2003-2014).

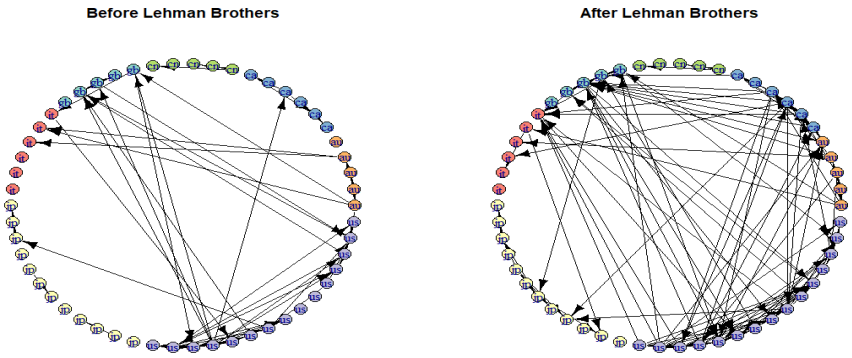


Figure 4: The DAG estimated from MaCho network for the pre and post Lehman Brothers failure.

Bien, J. and Tibshirani, R. J. Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820, 2011. ISSN 00063444, 14643510.

Billio, M., Getmansky, M., Lo, A. W., and Pelizzon, L. Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics*, 104(3):535–559, 2012. ISSN 0304-405X. doi: https://doi.org/10.1016/j.jfineco.2011.12.010. Market Institutions, Financial Market Risks and Financial Crisis.

Bollen, K. *Structural Equations with Latent Variables*. John Wiley and Sons, New York, 1989.

Breiman, L. Random forests. *Machine Learning*, 45 (1):5–32, 2001. doi: 10.1023/A:1010933404324.

Cai, T. T., Ren, Z., and Zhou, H. H. Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics*, 10:1–59, 2016.

Chen, W., Drton, M., and Wang, Y. S. On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980, 09 2019. ISSN 0006-3444.

Chickering, D. M. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3(null): 507–554, 2003. ISSN 1532-4435.

Davis, T. A. and Hager, W. W. Row modifications of a sparse cholesky factorization. *SIAM Journal on Matrix Analysis and Applications*, 26(3):621–639, 2005.

Demirer, M., Diebold, F. X., Liu, L., and Yilmaz, K. Estimating global bank network connectedness. *Journal of Applied Econometrics*, 33(1):1–15, 2018.

Gao, M. and Aragam, B. Efficient bayesian network structure learning via local markov boundary search. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21. Curran Associates Inc., 2021. ISBN 9781713845393.

Gao, M., Ding, Y., and Aragam, B. A polynomial-time algorithm for learning nonparametric causal graphs. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 11599–11611. Curran Associates, Inc., 2020.

Ghoshal, A. and Honorio, J. Learning identifiable gaussian bayesian networks in polynomial time and sample complexity. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Ghoshal, A. and Honorio, J. Learning linear structural equation models in polynomial time and sample complexity. In *Proceedings of the Twenty-First*

International Conference on Artificial Intelligence and Statistics, volume 84 of *Proceedings of Machine Learning Research*, pp. 1466–1475, 2018.

Glymour, C., Zhang, K., and Spirtes, P. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10, 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.00524. URL https://www.frontiersin.org/articles/10.3389/fgene.2019.00524.

Golub, G. H. and Van Loan, C. F. *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, Baltimore, MD, USA, 1996. ISBN 0-8018-5414-8.

Hardy, G., Littlewood, J., and Pólya, G. *Inequalities*. Cambridge Mathematical Library. Cambridge University Press, 1952. ISBN 9780521358804.

Horn, R. A. and Johnson, C. R. *Matrix Analysis*. Cambridge University Press, New York, NY, USA, 2nd edition, 2012.

Hoyer, P. O., Janzing, D., Mooij, J., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, NIPS'08, pp. 689–696, 2008. ISBN 9781605609492.

Janzing, D. and Schölkopf, B. Causal inference using the algorithmic markov condition. *IEEE Trans. Inf. Theor.*, 56(10):5168–5194, oct 2010. ISSN 0018-9448. doi: 10.1109/TIT.2010.2060095.

Kalainathan, D. and Goudet, O. Causal discovery toolbox: Uncover causal relationships in python, 2019.

Khare, K., Oh, S.-Y., Rahman, S., and Rajaratnam, B. A scalable sparse cholesky based approach for learning high-dimensional covariance matrices in ordered data. *Machine Learning*, 108(12):2061–2086, 2019. doi: 10.1007/s10994-019-05810-5.

Koller, D. and Friedman, N. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009. ISBN 0262013193.

Loh, P.-L. and Bühlmann, P. High-dimensional learning of linear causal networks via inverse covariance estimation. *Journal of Machine Learning Research*, 15(88):3065–3105, 2014. URL http://jmlr.org/papers/v15/loh14a.html.

Loh, P.-L. and Wainwright, M. J. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *J. Mach. Learn. Res.*, 16(1):559–616, 2015. ISSN 1532-4435.

Magnus, J. R. and Neudecker, H. Symmetry, 0-1 matrices and jacobians: A review. *Econometric Theory*, 2(2):157–190, 1986.

Marshall, A. W., Olkin, I., and Arnold, B. C. *Inequalities: Theory of Majorization and its Applications*, volume 143. Springer, second edition, 2011. doi: 10.1007/978-0-387-68276-1.

Mooij, J., Janzing, D., Peters, J., and Schölkopf, B. Regression by dependence minimization and its application to causal inference in additive noise models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 745–752, 2009. doi: 10.1145/1553374.1553470.

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. Distinguishing cause from effect using observational data: Methods and benchmarks. *J. Mach. Learn. Res.*, 17(1):1103–1204, jan 2016. ISSN 1532-4435.

Olkin, I. Estimating a cholesky decomposition. *Linear Algebra and its Applications*, 67:201 – 205, 1985. ISSN 0024-3795.

Park, G. Identifiability of additive noise models using conditional variances. *Journal of Machine Learning Research*, 21(75):1–34, 2020.

Park, G. and Kim, Y. Identifiability of gaussian linear structural equation models with homogeneous and heterogeneous error variances. *Journal of the Korean Statistical Society*, 49:276–292, 03 2020. doi: 10.1007/s42952-019-00019-7.

Park, G. and Park, H. Identifiability of generalized hypergeometric distribution (ghd) directed acyclic graphical models. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 158–166. PMLR, 16–18 Apr 2019.

Park, G. and Raskutti, G. Learning quadratic variance function (qvf) dag models via overdispersion scoring (ods). *Journal of Machine Learning Research*, 18 (224):1–44, 2018.

Peters, J. and Bühlmann, P. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 11 2013. ISSN 0006-3444.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. Identifiability of causal graphs using functional models. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, pp. 589–598, Arlington, Virginia, USA, 2011. AUAI Press. ISBN 9780974903972.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(58): 2009–2053, 2014. URL http://jmlr.org/papers/v15/peters14a.html.

Peters, J., Janzing, D., and Schlkopf, B. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017. ISBN 0262037319.

Rajaratnam, B. and Salzman, J. Best permutation analysis. *J. Multivar. Anal.*, 121:193–223, October 2013. doi: 10.1016/j.jmva.2013.03.001.

Raskutti, G. and Uhler, C. Learning directed acyclic graphs based on sparsest permutations. *Stat*, 7, 07 2013. doi: 10.1002/sta4.183.

Reisach, A., Seiler, C., and Weichwald, S. Beware of the simulated dag! causal discovery benchmarks may be easy to game. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 27772–27784. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/e987eff4a7c7b7e580d659feb6f60c1a-Paper.pdf`.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7 (72):2003–2030, 2006.

Shojaie, A. and Michailidis, G. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, 97(3):519–538, 07 2010. doi: 10.1093/biomet/asq038.

Simon, H. A. *Causal Ordering and Identifiability*, pp. 53–80. Springer Netherlands, 1977. ISBN 978-94-010-9521-1. doi: 10.1007/978-94-010-9521-1_5.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search, 2nd Edition*, volume 1 of *MIT Press Books*. The MIT Press, 2 edition, February 2001.

Tao, T. *Analysis II*. Singapore : Springer, third edition edition, 2016.

Wainwright, M. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. ISBN 9781108498029.

Wang, H. Coordinate descent algorithm for covariance graphical lasso. *Statistics and Computing*, 24:521–529, 2014.

Wang, M. and Allen, G. I. Thresholded graphical lasso adjusts for latent variables. *Biometrika*, 110(3):681–697, 11 2022. ISSN 1464-3510. doi: 10.1093/biomet/asac060.

Xue, L., Ma, S., and Zou, H. Positive-definite ℓ_1-penalized estimation of large covariance matrices. *Journal of the American Statistical Association*, 107 (500):1480–1491, 2012. doi: 10.1080/01621459.2012.725386.

Yu, G. and Bien, J. Learning local dependence in ordered data. *Journal of Machine Learning Research*, 18:1–60, 2017.

Zhang, J. and Spirtes, P. The three faces of faithfulness. *Synthese*, 193:1011 – 1027, 2015.

Zhang, K. and Hyvärinen, A. On the identifiability of the post-nonlinear causal model. UAI '09, pp. 647–655, Arlington, Virginia, USA, 2009. AUAI Press. ISBN 9780974903958.

Zhao, P. and Yu, B. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7:2541–2563, dec 2006. ISSN 1532-4435.

Supplementary Materials for Generalized Criterion for Identifiability of Additive Noise Models Using Majorization

A Proof of Lemma 2.1

In this section, we show that the diagonal elements of the Cholesky factor of the covariance matrix Σ_ρ have a probabilistic interpretation as the conditional variances.

From (2), $\boldsymbol{\varepsilon} = (I - B_\rho)^{-1}\boldsymbol{X}$. That is for $2 \leq j \leq p$,

$$\varepsilon_{\rho(j)} = X_{\rho(j)} + \sum_{i=1}^{j-1} U_{\rho(j),\rho_o(i)} X_{\rho(i)},$$

where $U_{\rho_0} = (I - B_\rho)^{-1}$. Since by definition of SEM, for $\rho(k) < \rho(j)$, $\varepsilon_{\rho(j)} \perp\!\!\!\perp X_{\rho(k)}$ then

$$\begin{aligned}\operatorname{var}(\varepsilon_{\rho(j)}) &= \operatorname{var}(\varepsilon_{\rho(j)}|X_{\rho(1)},\ldots,\rho(j-1)) = \operatorname{var}(X_{\rho(j)} + \sum_{i=1}^{j-1} U_{\rho(j),\rho(i)} X_{\rho(i)}|X_{\rho(1)},\ldots,X_{\rho(j-1)}) \\ &= \operatorname{var}(X_{\rho(j)}|X_{\rho(1)},\ldots,X_{\rho(j-1)})\end{aligned}$$

Since $\Sigma_\rho = LL^{'}$ is the unique Cholesky decomposition, from (3), the result follows.

B Proof of Theorem 3.2

The first part of the proof adapts a similar idea and technique as Theorem 1 in Peters & Bühlmann (2013). We assume that there are two models as in (1) with DAGs G and $G^{'}$ respectively, such that both induce the same joint distribution. We will show by contradiction that $G = G^{'}$. By definition, DAGs contain nodes that have no child, such that eliminating such nodes from the graph leads to a DAG again. Consequently, in both G and $G^{'}$, we remove all nodes that have no children but have the same parents in both graphs. If there are no remaining nodes, the two graphs are identical and the result follows. Otherwise, we obtain two smaller DAGs, which we again call G and $G^{'}$. It is important to note that, there exists at least one node L in G that has no children and is such that either the parents sets $\mathrm{Pa}_L^G \neq \mathrm{Pa}_L^{G^{'}}$ or $\mathrm{Ch}_L^{G^{'}} \neq \emptyset$. Therefore, from Markov property of the distribution with respect to G, all other nodes of the graph are independent from L given its parents

$$L \perp\!\!\!\perp \boldsymbol{X} \backslash (\mathrm{Pa}_L^G \cup L) | \mathrm{Pa}_L^G$$

For the rest of the proof, it is informative to partition the parents of L in G into $\boldsymbol{Y}, \boldsymbol{Z}$ and $\boldsymbol{W}$ (for illustration, see Figure 5), where $\boldsymbol{Z}$ are also parents of L in $G^{'}$, the $\boldsymbol{Y}$ are children of L in $G^{'}$, and the $\boldsymbol{W}$ are not adjacent to L in $G^{'}$. Let $\boldsymbol{D}$ be the $G^{'}$ parents of L that are not adjacent to L in G and $\boldsymbol{E}$ be the $G^{'}$ children of L that are not adjacent to L in G. Thus we have $\mathrm{Pa}_L^G = \boldsymbol{Y} \cup \boldsymbol{Z} \cup \boldsymbol{W}$, $\mathrm{Ch}_L^G = \emptyset$, $\mathrm{Pa}_L^{G^{'}} = \boldsymbol{Z} \cup \boldsymbol{D}$ and $\mathrm{Ch}_L^{G^{'}} = \boldsymbol{Y} \cup \boldsymbol{E}$.

(a) Part of G (b) Part of $G^{'}$

Figure 5: Partition of the nodes in G and $G^{'}$.

From Peters et al. (2014, Proposition 29), there exist a node $Y \in \boldsymbol{Y}$, such that for the sets $\boldsymbol{Q} := \mathrm{Pa}_L^G \backslash \{Y\}$, $\boldsymbol{R} := \mathrm{Pa}_Y^{G^{'}} \backslash \{L\}$ and $\boldsymbol{S} := \boldsymbol{Q} \cup \boldsymbol{R}$ we have $\boldsymbol{S} \subset \mathrm{Nd}_L^G \backslash \{Y\}$ and $\boldsymbol{S} \subset \mathrm{Nd}_Y^{G^{'}} \backslash \{L\}$. Take any $\boldsymbol{s} = (\boldsymbol{q}, \boldsymbol{r})$ and denote $L^* := L|_{\boldsymbol{S}=\boldsymbol{s}}$ and $Y^* := Y|_{\boldsymbol{S}=\boldsymbol{s}}$. From G, we can write

$$L^* = f_L(\boldsymbol{q}, Y^*) + \varepsilon_L$$

such that $\varepsilon_L \perp\!\!\!\perp Y^*$, since $\boldsymbol{S} \subset \mathrm{Nd}_L^G \backslash \{Y\}$. Applying the majorization Assumption 1 and denoting $\boldsymbol{x} = [\mathrm{var}(Y^*), E[\mathrm{var}(L^*|Y^*)]]^{'}$, from Theorem 2.4, $f(\boldsymbol{x}) < f(\tilde{\boldsymbol{x}})$, where $\tilde{\boldsymbol{x}} = [\mathrm{var}(L^*), E[\mathrm{var}(Y^*|L^*)]]^{'}$.

Similarly from $G^{'}$ we have

$$Y^* = g_Y(\boldsymbol{r}, L^*) + \varepsilon_Y$$

such that $\varepsilon_Y \perp\!\!\!\perp L^*$, since $\boldsymbol{S} \subset \mathrm{Nd}_Y^{G^{'}} \backslash \{L\}$. Now, the majorization Assumption 1 for $G^{'}$ results in $f(\tilde{\boldsymbol{x}}) < f(\boldsymbol{x})$, which is contradiction.

C Proof of Theorem 4.1

The following results will be used in the proof of the theorem. Recall that, we denoted a topological ordering of the graph G by $\rho_0 = \{1, \ldots, p\}$.

Lemma C.1. *Let $j \in De(i)$, i.e, $i < j$ and Assumption 2 is satisfied, then*

a. $E[var(X_i|X_1, \ldots, X_{i-1})] < E[var(X_j|X_1, \ldots, X_{i-1})]$

b. $var(X_j|X_1, \ldots, X_{j-1}) > var(X_i|X_1, \ldots, X_j \backslash X_i)$

Proof of a.

From the Law of Total Variance

$$\text{var}(X_j|X_1,\ldots,X_{i-1}) = E[\text{var}(X_j|\text{Pa}j)] + \text{var}(E[X_j|\text{Pa}j]|X_1,\ldots,X_{i-1})$$

After taking expectation from both sides and using Assumption 2(a) the result follows.

Proof of b. From the Law of Total Variance

$$E[\text{var}(X_i|\text{pa}_i)] = \text{var}(X_i|X_1,\ldots,X_j\backslash X_i) + E[\text{var}(E\{X_i|X_1,\ldots,X_j\backslash X_i\}|\text{pa}i)]$$

Then, the result directly follows from Assumption 2(b).

Lemma C.2. *Let $\boldsymbol{X}$ be generated from an ANM (1) with DAG G and $j \in De(i)$, then*

a. $var(X_j|X_1,\ldots,X_{i-1}) \geq var(X_j|X_1,\ldots,X_{j-1})$

b. $var(X_i|X_1,\ldots,X_{i-1}) \geq E[var(X_i|X_1,\ldots,X_j\backslash X_i)]$

Proof of a.

From the definition of ANM (1) for $k < j$, $\varepsilon_j \perp\!\!\!\perp X_k$ and $\text{var}(\varepsilon_j) = \text{var}(X_j|X_1,\ldots,X_{j-1})$. Define $\text{pa}j = U \cup V$, such that $U \cap V = \emptyset$, $U \subset \{X_1,\ldots,X_{j-1}\}$ and for all $v \in V$, $v \notin \{X_1,\ldots,X_{i-1}\}$. Now, from (1) and $i < j$

$$\text{var}(X_j|X_1,\ldots,X_{i-1}) = \text{var}(f_j(U,V)|X_1,\ldots,X_{i-1}) + \text{var}(\varepsilon_j)$$

Since $\text{var}(f_j(U,V)|X_1,\ldots,X_{i-1}) \geq 0$, the result follows.

Proof of b. This is a general result from a well known method of variance reduction by conditioning and follows from the Law of Total Variance.

Proof of Theorem 4.1

We show that if the Assumption 2 is satisfied then there exist finite number of T-transformations such that $\boldsymbol{x} \leq T_1 \ldots T_k \boldsymbol{y}$ (see Theorem 2.3). As a consequence of Marshall et al. (2011, Lemma B.1 and 3.A.5), without loss of generality, to proof results of majorization we can consider only cases when $\boldsymbol{x}$ and $\boldsymbol{y}$ differ in only two components.

Consider T-transformation $T = \lambda I + (1-\lambda)Q$, where Q is a permutation matrix that interchanges two coordinates i and j, such that $j \in \text{De}(i)$ in a topological ordering. That is $[X_1,\ldots,X_i,\ldots,X_j,\ldots,X_p]'$ is permuted to $[X_1,\ldots,X_j,\ldots,X_i,\ldots,X_p]'$. Let $\tilde{\boldsymbol{X}} = Q\boldsymbol{X}$. Using the above notation we define

$$\boldsymbol{x} = [\text{var}(X_1),\ldots,\text{var}(X_i|\{1,\ldots,i-1\}),\ldots,\text{var}(X_j|\{1,\ldots,j-1\}),\ldots,\text{var}(X_p|\{1,\ldots,p-1\}]', \tag{15}$$

and

$$\tilde{\boldsymbol{x}} = [\text{var}(X_1),\ldots,\text{var}(X_j|\{1,\ldots,i-1\}),\ldots,\text{var}(X_i|\{1,\ldots,j\}\backslash i),\ldots,\text{var}(X_p|\{1,\ldots,p-1\}]', \tag{16}$$

where we denote $\text{var}(X_i|X_m,X_k)$ as $\text{var}(X_i|\{m,k\})$ and $\tilde{\boldsymbol{x}}$ contains conditional variances using the new ordering with i and j interchanged. From Theorem 2.3 and the definition of T-transformation, our goal is to show that

$$\boldsymbol{x}_i \leq \lambda\tilde{\boldsymbol{x}}_i + (1-\lambda)\tilde{\boldsymbol{x}}_j \tag{17}$$

and

$$\boldsymbol{x}_j \leq \lambda\tilde{\boldsymbol{x}}_j + (1-\lambda)\tilde{\boldsymbol{x}}_i, \tag{18}$$

where $\boldsymbol{x}_i = \text{var}(X_i|\{1,\ldots,i-1\}$, $\tilde{\boldsymbol{x}}_i = \text{var}(X_j|\{1,\ldots,i-1\}$, $\boldsymbol{x}_j = \text{var}(X_j|\{1,\ldots,j-1\})$ and $\tilde{\boldsymbol{x}}_j = \text{var}(X_i|\{1,\ldots,j\backslash i\}$.

From Lemma C.1

$$\boldsymbol{x}_i < \tilde{\boldsymbol{x}_i} \text{ and } \tilde{\boldsymbol{x}_j} < \boldsymbol{x}_j. \tag{19}$$

Similarly, from Lemma C.2

$$\boldsymbol{x}_j \leq \tilde{\boldsymbol{x}_i} \text{ and } \tilde{\boldsymbol{x}_j} \leq \boldsymbol{x}_i \tag{20}$$

We examine the following cases

Case 1: $\boldsymbol{x}_i \geq \boldsymbol{x}_j$

From (19) and (20)

$$\tilde{\boldsymbol{x}_j} < \boldsymbol{x}_j \leq \boldsymbol{x}_i < \tilde{\boldsymbol{x}_i}$$

Let $\delta = \min(\tilde{\boldsymbol{x}_i} - \boldsymbol{x}_i, \boldsymbol{x}_j - \tilde{\boldsymbol{x}_j})$ and $1 - \lambda = \delta/(\tilde{\boldsymbol{x}_i} - \tilde{\boldsymbol{x}_j})$. Then (17) and (18) follow.

Case 2: $\boldsymbol{x}_i < \boldsymbol{x}_j$

Then $\tilde{\boldsymbol{x}_j} \leq \boldsymbol{x}_i < \boldsymbol{x}_j \leq \tilde{\boldsymbol{x}_i}$ and by defining δ and λ accordingly the result follow.

D Majorization and Park (2020) identifiability conditions

D.1 Check of majorization assumption for linear ANMs

The following proposition will be useful to check the weak majorization assumption. Let $\mathcal{D} = \{z : z_1 \geq \cdots \geq z_n\}$ and

Proposition D.1. *(Marshall et al. (2011, Proposition B.1.A)) Let $x \in \mathcal{D}_{++}$ and $\sum x_i \leq \sum y_i$, if for some k, $1 \leq k \leq n$, $x_i \leq y_i$ for $i = 1, \ldots, k$ and $x_i \geq y_i$ for $i = k+1, \ldots, n$ then $x \prec_w y$.*

In this section, we rely on Lemma 2.1 to propose a quick way to check whether a majorization is satisfied for the simulated data generated from the linear ANM:

$$\boldsymbol{X} = B\boldsymbol{X} + \boldsymbol{\varepsilon}$$

Let

$$x = \text{diag}(\text{var}(\boldsymbol{\varepsilon})) \tag{21}$$

be the diagonal elements of the covariance matrix, and Σ_ρ is a covariance matrix of $\boldsymbol{X}$ after row and column permutation for some $\rho \in \mathfrak{S}_p$. Define

$$y = \text{diag}(\text{chol}(\Sigma_\rho)^2), \tag{22}$$

then, from Lemma 2.1 the majorization assumption holds if $x \prec_w y$ for all $\rho \in \mathfrak{S}_p$. We summarize the Assumption 1 check in Algorithm 2

Algorithm 2 IsMajorized?

Input: x, y as in (21) and (22)
Sort x in a descending order
if criterion in Proposition D.1 holds **then**
  Return TRUE
else
  Return FALSE
end if

E Proof of Lemma 5.2

As we discussed in Section 5.1, in each iteration, the worst computational cost is $O(p^2)$. Consequently, the total cost is $O(p^3)$.

F Proof of Lemma 5.1

Let the Cholesky factors of the sample and the population covariance matrices be $S = \hat{L}\hat{L}^{'}$ and $\Sigma = LL^{'}$.

The from from Olkin (1985) the joint distribution of the elements of $\hat{L}$ is

$$p(\hat{L}) = 2^p c\Big(\prod_{i=1}^{p} L_{ii}^{-n}\Big)\Big(\prod_{i=1}^{p} \hat{L}_{ii}^{n-p-1}\Big)\Big(\prod_{i=1}^{p} \hat{L}_{ii}^{p-i+1}\Big)\exp\Big(-\frac{1}{2}\text{tr}L^{-1}\hat{L}\hat{L}^{'}L^{-'}\Big), \tag{23}$$

where $c^{-1} = 2^{np/2}\pi^{p(p-1)/4}\prod_{i=1}^{p}\Gamma\Big(\frac{n-i+1}{2}\Big)$.

Let $U = L^{-1}\hat{L}$, then for $i > j$, $U_{ij} \sim N(0,1)$ and $U_{ii} \sim \chi^1_{n-i+1}$. As a result

$$\hat{L}_{ij} = L_{ij}U_{jj} + \sum_{l=j+1}^{i} L_{il}U_{lj} \tag{24}$$

$$E[\hat{L}_{ij}] = L_{ij}E[U_{jj}] \tag{25}$$

and

$$E[\hat{L}_{ii}] = L_{ii}E[U_{ii}] \tag{26}$$

where $E[U_{ii}] = \frac{\sqrt{2}\Gamma[(n-i+2)/2]}{\Gamma[(n-i+1)/2]}$

Consider the random variable $V_{ii} = U_{ii}/\sqrt{n-i+1}$. Since $U_{ii} = \sqrt{U_{ii}^2} = \|Z_1,\ldots,Z_{n-i+1}\|_2/\sqrt{n-i+1}$, where $Z_i \sim N(0,1)$ and the Euclidean norm is a 1-Lipschitz function, from Wainwright (2019, Theorem 2.26)

$$P(V_{ii} - E[V_{ii}] \geq t) \leq \exp\Big(-\frac{(n-i+1)t^2}{2}\Big)$$

From Jensen's inequality

$$E[V_{ii}] \leq \sqrt{E[V_{ii}^2]} = \Big(\frac{1}{n-i+1}\sum_{k=1}^{n-i+1} E[Z_k^2]\Big)^{1/2} = 1$$

Thus,

$$P(V_{ii} - 1 \geq t) \leq P(V_{ii} - E[V_{ii}] \geq t) \leq \exp\Big(-\frac{(n-i+1)t^2}{2}\Big) \tag{27}$$

and as $n \to \infty$, V_ii goes to 1 with high probability. Now,

$$E[\tilde{L}_{ij}] = L_{ij}E[V_{ij}]$$

and the result follows from (27).

G Proof of Theorem 5.3

The proof of Theorem 5.3 relies on the following assumptions.

- A1 *Marginal sub-Gaussian assumption:* The sample matrix $X \in \mathcal{R}^{n\times p}$ has n independent rows with each row drawn from the distribution of a zero-mean random vector $X = (X_1,\ldots,X_p)^t$ with covariance Σ and sub-Gaussian marginals; i.e.,
$$E[\exp(tX_j/\sqrt{\Sigma_{jj}})] \leq \exp(Ct^2)$$
for all $j = 1,\ldots,p$, $t \leq 0$ and for some constant $C > 0$.

- A2 *Sparsity Assumption:* The true Cholesky factor $L \in \mathcal{R}^{p\times p}$ is the lower triangular matrix with positive diagonal elements and support $\mathcal{S}(L) = \{(i,j), i \neq j | L_{ij} \neq 0\}$. We denote by $s = |S|$ cardinality of the set S.
- A3 *Bounded eigenvalues:* There exist a constant κ such that

$$0 < \kappa^{-1} \leq \lambda_{min}(L) \leq \lambda_{max}(L) \leq \kappa$$

- A4 The minimum edge strength:

$$\ell_{\min} := \min_{1\leq j<i\leq p} |L_{ij}| > c_1\sqrt{\frac{s\log p}{n}} \tag{28}$$

The following lemma, which proof is for a general penalty functions $\rho(\cdot,\lambda)$ that include ℓ_1 penalty as in (12) and satisfy the conditions below, will be useful in the proof of Theorem 5.3.

- The function $\rho(\cdot,\lambda)$ satisfies $\rho(0,\lambda) = 0$ and is symmetric around zero.
- On the non negative real line, the function $\rho(\cdot,\lambda)$ is nondecreasing.
- For $t > 0$, the function $t \to \rho(\cdot,\lambda)/t$ is nonincreasing in t.
- The function $\rho(\cdot,\lambda)$ is differentiable for all $t \neq 0$ and subdifferentiable at $t = 0$, with $\lim_{t\to 0^+}\rho'(t,\lambda) = \lambda C$.
- There exists $\mu > 0$ such that $\rho_\mu(t,\lambda) = \rho(t,\lambda) + \frac{\mu}{2}t^2$ is convex.

Recall that a matrix $\hat{L} \in \mathcal{L}_p$ is a stationary point for (12) if it satisfies (Bertsekas, 2015)

$$\langle \nabla\mathcal{L}_n(\hat{L}) + \nabla\rho(\hat{L},\lambda), L - \hat{L}\rangle \geq 0, \text{ for } L \in \mathcal{L}_p, \tag{29}$$

where $\mathcal{L}_n(L) = \mathrm{tr}(SLL^t) - 2\log|L|$ and $\nabla\rho(\cdot,\cdot)$ is the subgradient.

Lemma G.1. *Under Assumptions A1-A3, with tuning parameter λ of scale $\sqrt{\frac{\log p}{n}}$, and $\frac{3}{4\gamma} < (\kappa+1)^{-2}$, the scaling $(s+p)\log p = o(n)$ is sufficient for any stationary point $\hat{L}$ of the (12) to satisfy the following estimation bounds:*

$$\|\hat{L} - L\|_F = \mathcal{O}_p\Big(\sqrt{\frac{(s+p)\log p}{n}}\Big)$$

$$\|\hat{L}_{\mathit{off}} - L_{\mathit{off}}\|_F = \mathcal{O}_p\Big(\sqrt{\frac{s\log p}{n}}\Big),$$

where L_{off} refers to all the off-diagonal entries of a matrix L.

The proof is provided in Section G.1.

Now, we are ready to provide the proof of Theorem 5.3. From Assumption 2, for $(i,j), i \neq j \notin \mathcal{S}(L)$, we have $L_{ij} = 0$ and from Lemma G.1 there exist $c_2 > 0$ such that

$$|\hat{L}_{ij}| < c_2\frac{s\log p}{n}$$

Then by the definition of thresholded CSCS (13), it follows that $\hat{L}^\tau = 0$.

For $(i,j) \in \mathcal{S}(L)$, from the Assumption 4

$$L_{ij} > c_1\sqrt{\frac{s\log p}{n}}$$

and from Lemma G.1 for $i \neq j$

$$|\hat{L}_{ij} - L_{ij}| < c_2\sqrt{\frac{s\log p}{n}}.$$

Since, we assumed $c_1 > 2c_2$, it follow that $|\hat{L}_{ij}| > c_2\sqrt{\frac{s\log p}{n}}$. Then by the definition of thresholded CSCS (13), it follows that $\hat{L}^\tau \neq 0$ and the result follows.

G.1 Proof of Lemma G.1

We start by showing that $\mathcal{L}_n$ satisfies RSC conditions. Recall that the differentiable function $\mathcal{L}_n : \mathcal{R}^{p\times p} \to \mathcal{R}$ satisfies RSC condition if:

$$\langle \nabla\mathcal{L}_n(L+\Delta) - \nabla\mathcal{L}_n(L), \Delta\rangle \geq$$

$$\geq \begin{cases} \alpha_1\|\Delta\|_F^2 - \tau_1\dfrac{\log p}{n}\|\Delta\|_1^2, & \forall\|\Delta\|_F \leq 1 \quad (30) \\ \alpha_2\|\Delta\|_F - \tau_2\sqrt{\dfrac{\log p}{n}}\|\Delta\|_2, & \forall\|\Delta\|_F \geq 1 \quad (31) \end{cases}$$

where the α_j's are strictly positive constants and the τ_j's are nonnegative constants. From Loh & Wainwright (2015, Lemma 4), under conditions of Lemma G.1, if (30) holds then (31) holds. Thus, we concentrate only on showing that (30) holds for $\|\Delta\|_F \leq 1$. Recall that

$$\mathcal{L}_n(L) = \mathrm{tr}(SL^tL) - 2\log|L| \tag{32}$$

Lemma G.2. *The cost function (32) satisfies RSC condition with $\alpha_1 = (\kappa+1)^{-2}$ and $\tau_1 = 0$; i.e.,*

$$\langle \nabla\mathcal{L}_n(L+\Delta) - \nabla\mathcal{L}_n(L), \Delta\rangle \geq (\kappa+1)^{-2}\|\Delta\|_F^2,\ \forall\|\Delta\|_F \leq 1 \tag{33}$$

The proof is provided in Section G.2.

From the penalty conditions listed above, $\rho_\mu(L,\lambda) = \rho(L,\lambda) + \frac{\mu}{2}\|L\|_F^2$ is convex. Thus,

$$\begin{aligned}\rho_\mu(L,\lambda) - \rho_\mu(\hat{L},\lambda) &\geq \langle\nabla\rho_\mu(\hat{L},\lambda), L-\hat{L}\rangle \\ &= \langle\nabla\rho(\hat{L},\lambda) + \mu\hat{L}, L-\hat{L}\rangle,\end{aligned}$$

which implies that

$$\langle\nabla\rho(\hat{L},\lambda), L-\hat{L}\rangle \leq \rho(L,\lambda) - \rho(\hat{L},\lambda) + \frac{\mu}{2}\|\hat{L}-L\|_F^2 \tag{34}$$

From stationarity condition (29)

$$\langle\nabla\mathcal{L}_n(\hat{L}), L-\hat{L}\rangle \geq -\langle\nabla\rho(\hat{L},\lambda), L-\hat{L}\rangle$$

and combining above result with (33)

$$\begin{aligned}(1+\kappa)^{-2}\|\Delta\|_F^2 &\leq \langle\mathcal{L}_n(\hat{L}),\Delta\rangle - \langle\nabla\mathcal{L}_n(L),\Delta\rangle \\ &\leq \langle\nabla\rho(\hat{L},\lambda), L-\hat{L}\rangle - \langle\nabla\mathcal{L}_n(L),\Delta\rangle \\ &\leq \rho(L,\lambda) - \rho(\hat{L},\lambda) + \frac{\mu}{2}\|\hat{L}-L\|_F^2 \\ &\quad - \langle\nabla\mathcal{L}_n(L),\Delta\rangle\end{aligned}$$

After rearrangement and H$\acute{}$older inequality

$$\begin{aligned}\Big((1+\kappa)^{-2}) - \frac{\mu}{2}\Big)\|\Delta\|_F^2 &\leq \rho(L,\lambda) - \rho(\hat{L},\lambda) \\ &\quad + \|\nabla\mathcal{L}_n(L)\|_\infty\|\Delta\|_1\end{aligned}$$

From Loh & Wainwright (2015, Lemma 4)

$$\lambda\|\Delta\|_1 \leq \rho(\Delta,\lambda) + \frac{\mu}{2}\|\Delta\|_F^2$$

and from Yu & Bien (2017, Lemma 15) under the assumed scaling of λ

$$\|\nabla\mathcal{L}_n(L)\|_\infty \leq \frac{\lambda}{2}$$

with probability going to 1. Combining above two results and using subadditive property; i.e., $\rho(\Delta, \lambda) \leq \rho(L, \lambda) + \rho(\hat{L}, \lambda)$:

$$\begin{aligned}\Big((1+\kappa)^{-2} - \frac{\mu}{2}\Big)\|\Delta\|_F^2 &\leq \rho(L,\lambda) - \rho(\hat{L},\lambda) + \frac{\lambda}{2}\|\hat{\Delta}\|_1 \\ &\leq \rho(L,\lambda) - \rho(\hat{L},\lambda) \\ &\quad + \frac{\rho(\Delta,\lambda)}{2} + \frac{\mu}{4}\|\Delta\|_F^2 \\ &\leq \rho(L,\lambda) - \rho(\hat{L},\lambda) \\ &\quad + \frac{\rho(L,\lambda) + \rho(\hat{L},\lambda)}{2} + \frac{\mu}{4}\|\Delta\|_F^2\end{aligned}$$

After rearranging and using $3/4\mu \leq (1+\kappa)^{-2}$

$$0 \leq \Big((1+\kappa)^{-2} - \frac{3}{4}\mu\Big)\|\Delta\|_F^2 \leq 3\rho(L,\lambda) - \rho(\hat{L},\lambda) \tag{35}$$

From (35) and Loh & Wainwright (2015, Lemma 5) follows

$$\rho(L,\lambda) - \rho(L,\lambda) \leq 2\lambda\|\Delta_S\| - \lambda\|\Delta_{S^c}\| \Rightarrow \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1$$

Thus,

$$\begin{aligned}\Big(2(1+\kappa)^{-2} - \frac{3}{2}\mu\Big)\|\Delta\|_F^2 &\leq \lambda\|\Delta_S\|_2 - \lambda\|\Delta_{S^c}\|_1 \\ &\leq \lambda\|\Delta_S\|_1 \leq \lambda\sqrt{p+s}\|\Delta\|_F,\end{aligned}$$

from which we conclude that

$$\|\Delta\|_F \leq \frac{6\lambda\sqrt{p+s}}{4(1+\kappa)^{-2} - 3\mu}, \tag{36}$$

and the result follows from the chosen scaling of λ.

For the precision matrix bound, from page 45 of Yu & Bien (2017) we note that

$$\hat{L}^t\hat{L} - L^tL = (\hat{L}-L)^t(\hat{L}-L) + (\hat{L}-L)^TL + L^t(\hat{L}-L)$$

and

$$\|L^t(\hat{L}-L)\|_F \leq |||L|||_2\|\hat{L}-L\|_F$$

From submultiplicativity property of matrix norm

$$\|(\hat{L}-L)^t(\hat{L}-L)\|_F \leq \|(\hat{L}-L)\|_F^2$$

Therefore,

$$\|\hat{L}^t\hat{L} - L^tL\|_F \leq (\|\hat{L}-L\|_F + 2|||L|||_2)\|\hat{L}-L\|_F \tag{37}$$

The upper bound for the off-diagonal elements $\hat{L}_{\text{off}}$ easily follows from the above proof. For example, see Wang & Allen (2022, Proposition 1).

G.2 Proof of Lemma G.2

The following facts will be useful in the proof.

Fact 1

1. $(K_{pp})^{-1} = K_{pp}$
2. $\lambda_{max}(K_{pp}) = 1$
3. $\text{tr}(ABCD) = \text{vec}(D^t)(C^t \otimes A)\text{vec}(B)$
4. $\lambda_{max}(A \otimes B) = \lambda_{max}(A)\lambda_{max}(B)$

where K_{pp} is the commutation matrix such that $\text{vec}(L) = K_{pp}\text{vec}(L^t)$. The proof of the facts can be found in Magnus & Neudecker (1986, Section 4).

To show the RSC condition, we rely on the directional derivatives (for example see Tao (2016, Section 6.3)). In particular, if we denote by $D_\Delta \mathcal{L}_n(L)$ the directional derivative with respect to the direction Δ, then from Tao (2016, Lemma 6.3.5) :

$$\langle \nabla \mathcal{L}_n(L), \Delta \rangle = D_\Delta \mathcal{L}_n(L) = 2\text{tr}[(SL^t - L^{-1})\Delta] \tag{38}$$

Similarly

$$\begin{aligned} \langle \nabla \mathcal{L}_n(L+\Delta), \Delta \rangle &= D_\Delta \mathcal{L}_n(L+\Delta) \\ &= 2\text{tr}[(S(L+\Delta)^t - (L+\Delta)^{-1})\Delta] \end{aligned} \tag{39}$$

From Woodbury identity Horn & Johnson (2012)

$$(L+\Delta)^{-1} = L^{-1} - L^{-1}\Delta(L+\Delta)^{-1}$$

Plugging back into (39) and after some algebra

$$\begin{aligned} \langle \nabla \mathcal{L}_n(L+\Delta), \Delta \rangle &= 2\text{tr}[(S(L+\Delta)^t - (L+\Delta)^{-1})\Delta] \\ &\quad + 2\text{tr}[S\Delta^t\Delta + L^{-1}\Delta(L+\Delta)^{-1}\Delta] \end{aligned} \tag{40}$$

Thus, from (38) and (40)

$$\begin{aligned} &\langle \nabla \mathcal{L}_n(L+\Delta) - \nabla \mathcal{L}_n(L), \Delta \rangle = \\ &\qquad = 2\text{tr}[\Delta^t S\Delta + L^{-1}\Delta(L+\Delta)^{-1}\Delta] \\ &\qquad \geq \text{vec}(\Delta)^t K_{pp}((L+\Delta)^{-t} \otimes L^{-1})\text{vec}(\Delta) \\ &\qquad = \text{vec}(\Delta)^t[((L+\Delta)^t \otimes L)K_{pp}^{-1}]^{-1}\text{vec}(\Delta) \\ &\qquad \geq \lambda_{min}([((L+\Delta)^t \otimes L)K_{pp}^{-1}]^{-1})\|\Delta\|_F^2, \end{aligned} \tag{41}$$

where for the first inequality we used the fact that S is positive semi-definite and the second equality follows from the Fact 1. Now, since

$$\begin{aligned} &\lambda_{min}([((L+\Delta)^t \otimes L)K_{pp}^{-1}]^{-1}) = \\ &\qquad = \lambda_{max}^{-1}[((L+\Delta)^t \otimes L)K_{pp}^{-1}] \\ &\qquad \geq \lambda_{max}^{-1}(K_{pp}^{-1})\lambda_{max}^{-1}(L)\lambda_{max}^{-1}(L+\Delta) \\ &\qquad \geq (\kappa+1)^{-2}, \end{aligned} \tag{42}$$

where the first inequality follows from the submultiplicativity property of the norm and lower-triangularity of the L and Δ. The second inequality follows from the triangular property, the fact that $\|\Delta\|_2 \leq \|\Delta\|_F \leq 1$ and, properties of the K_{pp} stated in Fact 1. After plugging (42) into (41), the result follows.

H Assumption 1 and varsortability

In this section, we show that Assumption 1 holds even when varsortability is low. Consider the following data generation process

$$\begin{aligned} X_1 &:= \varepsilon_1 \\ X_2 &:= X_1 + \varepsilon_2 \\ X_3 &:= 0.85X_2 + \varepsilon_3 \\ X_4 &:= 0.79X_2 + \varepsilon_4, \end{aligned} \tag{43}$$

where $\sigma_1^2 = 4$, $\sigma_2^2 = 2$, $\sigma_3^2 = 1$, and $\sigma_4^2 = 1.5$. That is $x = \{4, 2, 1, 1.5\}$, where x is defined as in (5).

After generating data from (43) with $n = 1000$ and using the function `varsortability()` available from `https://github.com/Scriddie/Varsortability/`, it can be shown that, for this case, varsortability is equal to 0.6. Recall that when varsitability is one then ordering by marginal variance is a valid causal ordering.

We use procedure described in Appendix D.1 to check that Assumption 1 holds. Table 3 reports the vectors y_ρ, where for each permutation $\rho \in \mathfrak{S}_p$, y_ρ is defined as in (6). As can be seen, from Proposition D.1, $x \prec_w y_\rho$ is true for all permutations.

ρ	y_ρ
$\{1,2,4,3\}$	$\{4.00, 2.00, 1.50, 1.00\}$
$\{1,4,2,3\}$	$\{4.00, 2.75, 1.09, 1.00\}$
$\{4,1,2,3\}$	$\{5.24, 2.10, 1.09, 1.00\}$
$\{4,1,3,2\}$	$\{5.24, 2.10, 1.79, 0.61\}$
$\{1,4,3,2\}$	$\{4.00, 2.75, 1.79, 0.61\}$
$\{1,3,4,2\}$	$\{4.00, 2.44, 2.01, 0.61\}$
$\{1,3,2,4\}$	$\{4.00, 2.44, 1.50, 0.82\}$
$\{3,1,2,4\}$	$\{5.33, 1.83, 1.50, 0.82\}$
$\{3,1,4,2\}$	$\{5.33, 2.01, 1.83, 0.61\}$
$\{3,4,2,1\}$	$\{5.33, 2.20, 1.67, 0.61\}$
$\{4,3,1,2\}$	$\{5.24, 2.24, 1.67, 0.61\}$
$\{4,3,2,1\}$	$\{5.24, 2.24, 1.33, 0.77\}$
$\{3,4,2,1\}$	$\{5.33, 2.20, 1.33, 0.77\}$
$\{3,2,4,1\}$	$\{5.33, 1.50, 1.33, 1.12\}$
$\{3,2,1,4\}$	$\{5.33, 1.50, 1.33, 1.12\}$
$\{2,3,1,4\}$	$\{6.00, 1.50, 1.33, 1.00\}$
$\{2,3,4,1\}$	$\{6.00, 1.50, 1.33, 1.00\}$
$\{2,4,3,1\}$	$\{6.00, 1.50, 1.33, 1.00\}$
$\{4,2,3,1\}$	$\{5.24, 1.72, 1.33, 1.00\}$
$\{4,2,1,3\}$	$\{5.24, 1.72, 1.33, 1.00\}$
$\{2,4,1,3\}$	$\{6.00, 1.50, 1.33, 1.00\}$
$\{2,1,4,3\}$	$\{6.00, 1.50, 1.33, 1.00\}$
$\{2,1,3,4\}$	$\{6.00, 1.50, 1.33, 1.00\}$

Table 3: Estimated vectors y_ρ for the data generation process (43).

I Bayesian Networks

We start by introducing the following graphical concepts. If the graph $\mathcal{G}$ contains a directed edge from the node $k \to j$, then k is a parent of its child j. We write $\Pi_j^{\mathcal{G}}$ for the set of all parents of a node j. If there exists a directed path $k \to \cdots \to j$, then k is an ancestor of its descendant j. A *Bayesian Network* is a directed acyclic graph $\mathcal{G}$ whose nodes represent random variables $X_1, \ldots, X_p$. Then $\mathcal{G}$ encodes a set of conditional independencies and conditional probability distributions for each variable. The DAG $\mathcal{G} = (V, E)$ is characterized by the node set $V = \{1, \ldots, p\}$ and the edge set $E = \{(i,j) : i \in \Pi_j^{\mathcal{G}}\} \subset V \times V$. It is well-known that for a BN, the joint distribution factorizes as:

$$P(X_1, \ldots, X_p) = \prod_{j=1}^{p} P(X_j | \Pi_j^{\mathcal{G}}) \tag{44}$$

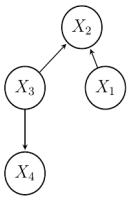
J Additional Results

J.1 Bank Connedtedness analysis for LINGAM

Figure 7 illustrates average degrees for the network estimated using LINGAM. As can be seen, the plot is noisy and no significant pattern can be discerned.

J.2 Multivariate Linear Simulation Results

Figures 10 and 11 report the results for $p = 20$ and $p = 50$ without LINGAM algorithm.



$$B = \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \\ X_4 \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & X_3 & X_4 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 \\ \beta_{21} & 0 & \beta_{23} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \beta_{43} & 0 \end{pmatrix} \end{array} \quad P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$B_\pi = \begin{array}{c} \\ X_1 \\ X_3 \\ X_2 \\ X_4 \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_3 & X_2 & X_4 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \beta_{21} & \beta_{23} & 0 & 0 \\ 0 & \beta_{43} & 0 & 0 \end{pmatrix} \end{array}$$

Figure 6: Illustration of DAG $\mathcal{G}$, corresponding coefficient matrix B, permutation matrix P, and permuted strictly lower triangular matrix B_π.

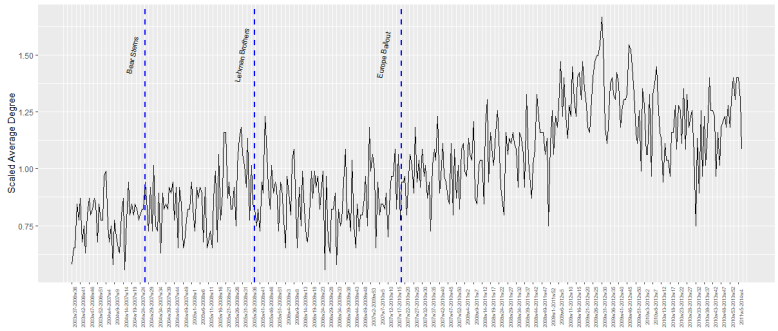


Figure 7: Evolution of average degree of bank connectedness scaled by their historic average (over 2003- 2014) using LINGAM.

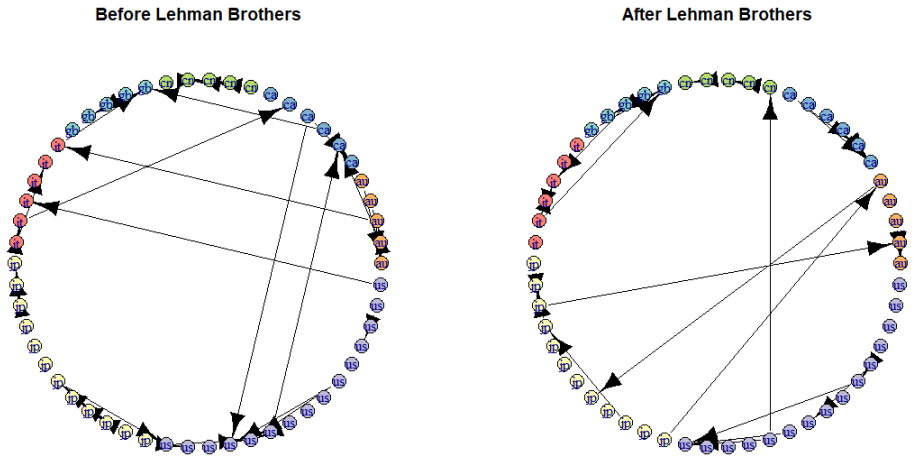


Figure 8: The DAG estimated from LINGAM does not show significant changes between before and after Lehman Brothers failure.

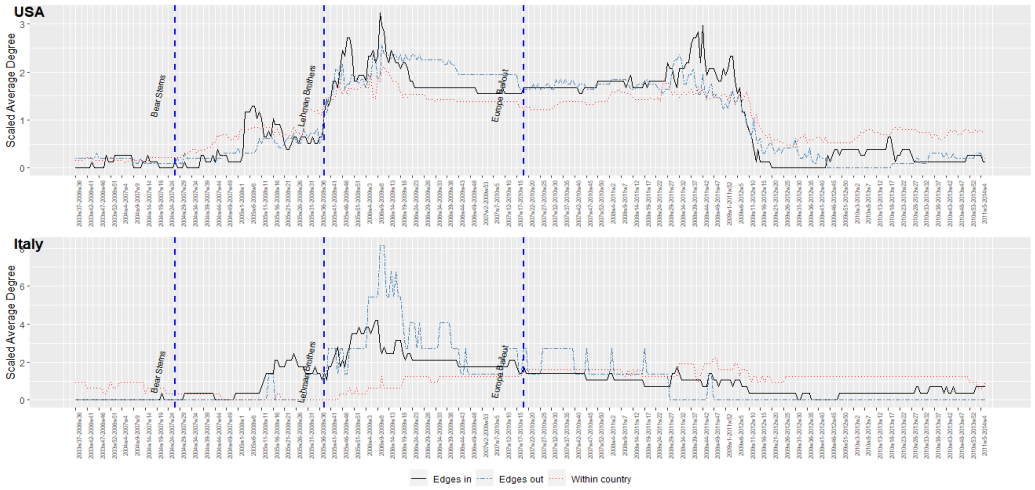
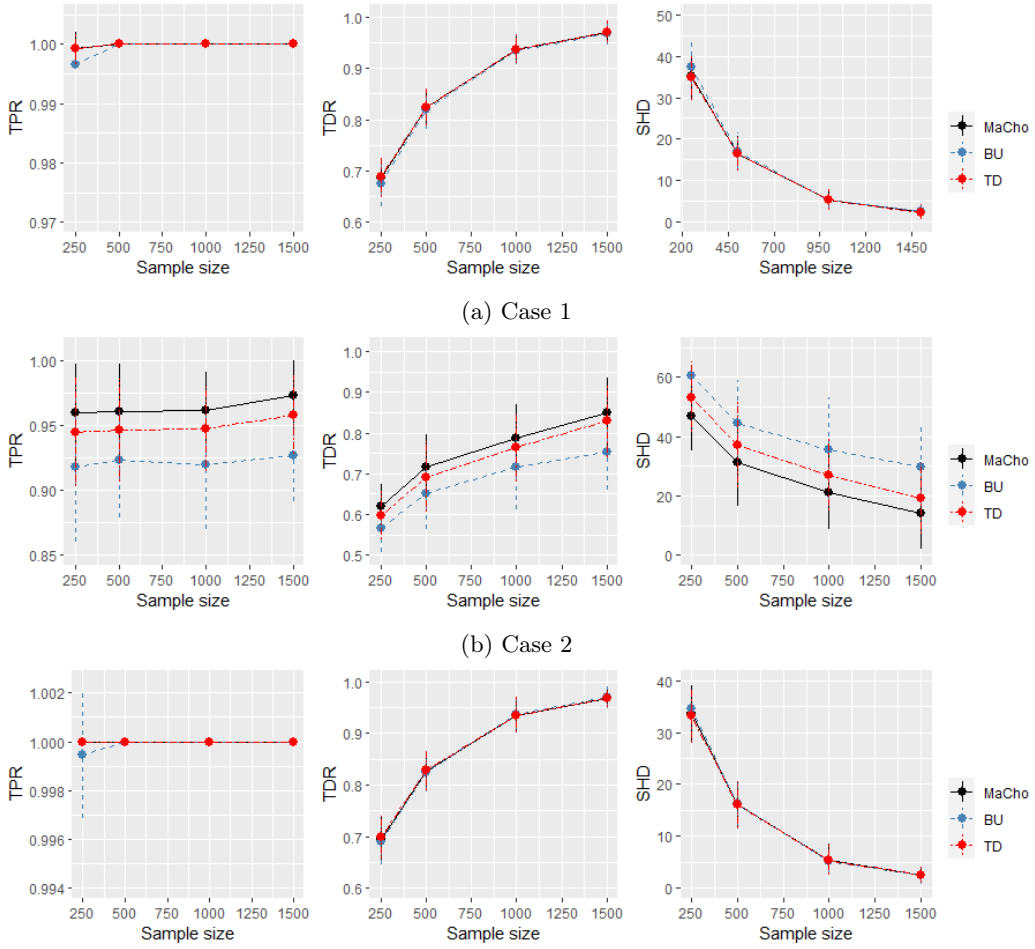


Figure 9: Evolution of average intra and inter-degree of bank connectedness scaled by their historic average (over 2003- 2014). (Top) USA and (Bottom) Italy.

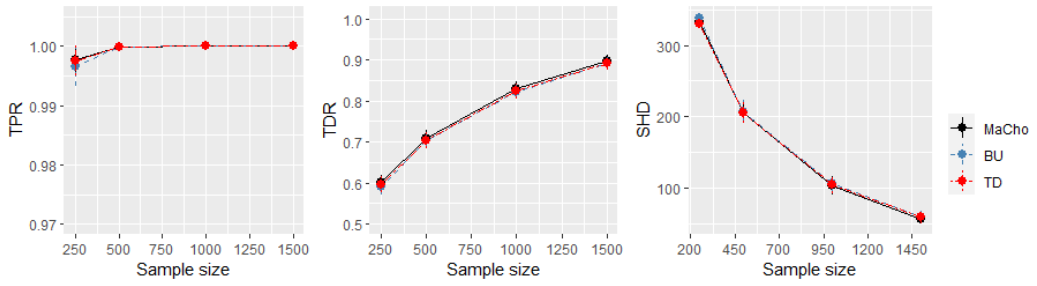


(a) Case 1

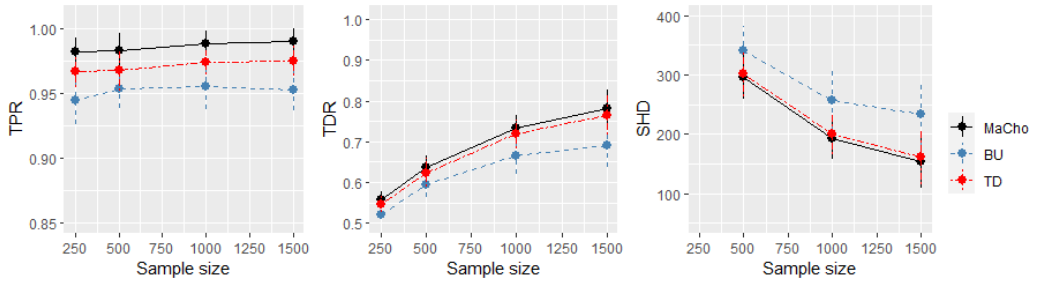
(b) Case 2

(c) Case 3

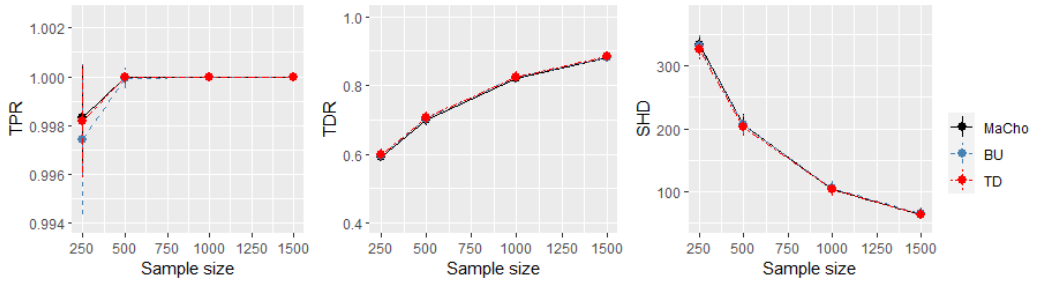
Figure 10: Comparison of the proposed algorithm MaCho, BU, and TD algorithms in terms of average TPR, TDR, and SHD for recovering linear SEM with different error variances and $p = 20$.



(a) Case 1



(b) Case 2



(c) Case 3

Figure 11: Comparison of the proposed algorithm MaCho, BU, and TD algorithms in terms of average TPR, TDR, and SHD for recovering linear SEM with different error variances and $p = 50$.