Learning Linear Non-Gaussian Polytree Models

Abstract

In the context of graphical causal discovery, we adapt the versatile framework of linear non-Gaussian acyclic models (LiNGAMs) to propose new algorithms to efficiently learn graphs that are polytrees. Our approach combines the Chow–Liu algorithm, which first learns the undirected tree structure, with novel schemes to orient the edges by imposing an algebraic condition on moments of the data-generating distribution. This algebraic edge orientation gains significant efficiency over current algorithms that assess conditional independence. We establish high-dimensional consistency results and compare the algorithms in numerical experiments.

1 INTRODUCTION

Directed acyclic graphs (DAGs) have been extensively used in causal modeling; the nodes of the graph represent the random variables of the model while the edges represent directed causal effects from source to sink. The causal effects of the parent nodes on the children are quantified by structural equations. In this paper, we study the problem of inferring the graphical structure underlying the causal model, given only observational data [Drton and Maathuis, 2017]. Referred to as structural learning or causal discovery, it is an intractable problem due to the statistical curse of dimensionality and computational issues. Imposing restrictions on the random variables, graphical structure, or structural equations simplifies the problem and contributes to a better understanding of the overarching problem of causal discovery [Pearl et al., 2016, Peters et al., 2017]. Here, we consider a class of graphs that are trees, together with linear structural equations where the error terms are mutually independent and non-Gaussian.

Specifically, we work in the versatile causal discovery framework of *linear non-Gaussian acyclic models (LiNGAMs)* [Shimizu et al., 2006]. LiNGAMs postulate linear structural equations to describe the relation among observed variables featuring non-Gaussian noise terms. The non-Gaussianity assumption allows for consistent estimation of the graph encoding the model from observational data alone and for efficient structural learning algorithms [e.g., Shimizu et al., 2011, Hyvärinen and Smith, 2013, Wang and Drton, 2020]. Since the complexity of the structural learning problem depends directly the underlying graph, consistency results for causal discovery algorithms often require some restrictions on the graph. DAGs are a general framework that also control the complexity of the learning problem; in particular, the subset of DAGs whose underlying skeleton is a tree—a *polytree*—is the most scalable setting, offering low computational complexity whilst retaining model expressiveness [Pearl, 1988]. In this paper, we propose algorithms to learn a polytree underlying a LiNGAM model.

Learning a polytree may be decomposed into two tasks: extracting the skeleton and determining the orientation of the edges [Rebane and Pearl, 1987, Jakobsen et al., 2021]. Recovering the underlying skeleton may be achieved via the Chow–Liu algorithm [Chow and Liu, 1968]. Existing methods for edge orientation entail checking conditional independence, which is usually carried out by serial hypothesis testing and hinders efficiency. We bypass this limitation by imposing an algebraic condition and building on recent insights concerning algebraic relations among moments to determine edge orientation [Robeva and Seby, 2021, Améndola et al., 2021]. The result is an efficient approach that adapts a classical algorithm to recover the core causal tree structure and augments it with a novel algebraic strategy to determine edge orientation, thus fully learning the polytree from observational data alone, in a far more scalable manner than existing LiNGAM algorithms that learn more general graph structures.

The remainder of the paper is organized as follows. Section 2 sets the background and theory. Section 3 presents our contributions where a general population version and

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

three cases are studied in detail. Corresponding theoretical guarantees to our proposed algorithms are given in Section 4. Results to numerical experiments are presented in Section 5. We close with a discussion and suggestions for future research in Section 6. The proofs of all the results are provided in the appendix A.

2 LINEAR NON-GAUSSIAN STRUCTURAL CAUSAL MODELS

A directed graph (digraph) is a pair $G = (V, E)$, where V is the set of vertices and $E \subset V \times V$ is the set of directed edges. Let $V = [p] := \{1, \dots, p\}$; an element $(i, j) \in E$ may also be denoted by $i \to j$. A digraph G is acyclic (i.e., a DAG) if it does not contain any directed cycle: there is no sequence of vertices $i_0, \dots, i_k$ with $i_j \to i_{j+1} \in E$ for $j = 0, \dots, k-1$ and $i_0 = i_k$. A path G is a sequence of vertices $i_0, \dots, i_k$ such that $i_j \to i_{j+1} \in E$ or $i_{j+1} \to i_j \in E$ for all j. It is is directed if all the arrows point in the same direction. A *polytree* is a DAG in which there is a unique path between any two vertices. A polytree is a directed tree if all the connecting paths are directed paths.

If $i \to j \in E$, then i is a parent of j, and j is a child of i. If G contains a directed path from i to j, then i is an ancestor of j and j is a descendant of i. The sets of parents, children, ancestors, and descendants of $i \in V$ are denoted by $\mathrm{pa}(i), \mathrm{ch}(i), \mathrm{an}(i), \mathrm{de}(i)$, respectively.

Let $X = (X_i)_{i \in [p]}$ be a random vector indexed by the vertices of a DAG G. For $A \subset [p]$, let $X_A = (X_i)_{i \in A}$. When X_A is conditionally independent of X_B given X_C for disjoint subsets $A, B, C \subset [p]$, we write $A \perp\!\!\!\perp B | C$. The joint distribution of X satisfies the local Markov property with respect to G if $\{i\} \perp\!\!\!\perp [p] \setminus (\mathrm{pa}(i) \cup \mathrm{de}(i)) \mid \mathrm{pa}(i) \; \forall \, i \in [p]$. The Markov equivalence class of G is the set of all DAGs that encode the same conditional independence relations, i.e., for which the set of distributions satisfying the local Markov property is the same. See Maathuis et al. [2019, Chap. 1] for further details.

The skeleton of a DAG is the undirected graph obtained by replacing each directed edge by an undirected edge. Here, edges are denoted by $\{i, j\} \subseteq E$.

2.1 STRUCTURAL EQUATIONS

A structural equation model hypothesizes that every random variable in X is functionally related to its parent variables: $X_i = f_i(X_{\mathrm{pa}(i)}, \varepsilon_i), \; i \in V$, where the ε_i are independent noise terms and the f_i are measurable functions. If the f_i are linear, then we obtain a linear structural equation model (LSEM). An LSEM can be written in matrix form as

$$X = (I - \Lambda)^{-\top} \varepsilon, \tag{2.1}$$

where $\Lambda = (\lambda_{ij})$ with $\lambda_{ij} \neq 0$ only if $i \to j \in E$. An LSEM constrains the dependence structure on the coordinates of X, but not the mean. Hence, when working with the LSEM, we may assume without loss of generality that $\mathbb{E}[\varepsilon_i] = 0$, which implies $\mathbb{E}[X_i] = 0$ for all $i \in V$.

Let $\varepsilon^{(2)} = (\mathbb{E}[\varepsilon_i \varepsilon_j])_{ij}$ be the covariance matrix of ε, which is a diagonal matrix by independence, and write $\varepsilon_i^{(2)} := \mathbb{E}[\varepsilon_i^2] > 0$ for its ith diagonal entry. The covariance matrix of X is then the positive definite matrix

$$\Sigma = (I - \Lambda)^{-\top} \varepsilon^{(2)} (I - \Lambda)^{-1}. \tag{2.2}$$

2.2 CUMULANTS IN GAUSSIAN AND NON-GAUSSIAN MODELS

Cumulants can be seen as alternative representations of moments of a distribution. Here, we formalize the definition in higher order settings and discuss their implications under Gaussian and non-Gaussian errors.

Definition 2.1. *The kth cumulant tensor of a random vector $(X_1, \dots, X_p)$ is the k-way tensor in $\mathbb{R}^{p \times \cdots \times p} \equiv (\mathbb{R}^p)^k$ whose entry in position $(i_1, \dots, i_k)$ is the joint cumulant*

$$\mathrm{cum}(X_{i_1}, \dots, X_{i_k}) := \sum_{(A_1, \dots, A_L)} (-1)^{L-1} (L-1)! \mathbb{E}\Big[\prod_{j \in A_1} X_j\Big] \cdots \mathbb{E}\Big[\prod_{j \in A_L} X_j\Big],$$

where the sum is taken over all partitions $(A_1, \dots, A_L)$ of the multiset $\{i_1, \dots, i_k\}$.

In our context, the variables have mean 0, so

$$\mathrm{cum}(X_i) = \mathbb{E}[X_i] = 0,$$
$$\mathrm{cum}(X_{i_1}, X_{i_2}) = \mathrm{Cov}[X_{i_1}, X_{i_2}] = \mathbb{E}[X_{i_1} X_{i_2}].$$

More generally, the sum can be restricted to the partitions in which all blocks A_i have at least two elements. In particular,

$$\begin{aligned}
&\mathrm{cum}(X_{i_1}, X_{i_2}, X_{i_3}) = \mathbb{E}[X_{i_1} X_{i_2} X_{i_3}], \\
&\mathrm{cum}(X_{i_1}, X_{i_2}, X_{i_3}, X_{i_4}) = \mathbb{E}[X_{i_1} X_{i_2} X_{i_3} X_{i_4}] \\
&- \mathbb{E}[X_{i_1} X_{i_2}] \mathbb{E}[X_{i_3} X_{i_4}] - \mathbb{E}[X_{i_1} X_{i_3}] \mathbb{E}[X_{i_2} X_{i_4}] \\
&- \mathbb{E}[X_{i_1} X_{i_4}] \mathbb{E}[X_{i_2} X_{i_3}].
\end{aligned}$$

The following powerful result dictates a simple condition for Gaussianity of X.

Theorem 2.2 (Marcinkiewicz)**.** *If there exists k such that $\mathrm{cum}(X_{i_1}, .., X_{i_j}) = 0$ for all $j \geq k$, then $k = 3$ and X has a multivariate Gaussian distribution.*

Furthermore, the following results dictate when the assumptions of Theorem 2.2 are satisfied, thus giving rise to Gaussianity, especially under LSEMs.

Lemma 2.3. *If the variables $X_1, \dots, X_n$ are independent, then* $\mathrm{cum}(X_{i_1}, \dots, X_{i_k}) = 0$ *unless* $i_1 = \cdots = i_k$.

Lemma 2.4. *Let the random vector X follow the LSEM from* (2.1) *with noise vector ε. Let $\mathcal{C}^{(k)}$ and $\varepsilon^{(k)}$ be the kth order cumulant tensors of X and ε, respectively. Then*

$$\begin{aligned}\mathcal{C}^{(k)} &= \varepsilon^{(k)} \bullet \left[(I-\Lambda)^{-1}\right]_{j=1}^{k} \\ &= \varepsilon^{(k)} \bullet (I-\Lambda)^{-1} \bullet \cdots \bullet (I-\Lambda)^{-1}\end{aligned}$$

is the Tucker product of $\varepsilon^{(k)}$ and k copies of $(I-\Lambda)^{-1}$.

Notice here that $\mathcal{C}^{(k)}$ reduces to (2.2) when $k = 2$.

See Comon and Jutten [2010] and references therein for proofs of Theorem 2.2 and Lemmas 2.3 and 2.4.

The next definition introduces the cumulant model obtained from the LSEM (2.1).

Definition 2.5. *Let $G = (V, E)$ be a DAG, and let $K \geq 2$ be an integer. The Kth cumulant model of G is the set of K-way tensors*

$$\begin{aligned}\mathcal{M}^{(K)}(G) = \{\varepsilon^{(K)} \bullet \left[(I-\Lambda)^{-1}\right]_{j=1}^{K} \ : \\ \Lambda \in \mathbb{R}^E,\ \varepsilon^{(K)} \in (\mathbb{R}^p)^K \text{ diagonal}\}.\end{aligned}$$

Here, $\mathbb{R}^E$ is the set of $p \times p$ matrices with support E. Further, the cumulants up to order K defined by G are modeled by

$$\mathcal{M}^{(\leq K)}(G) = \mathcal{M}^{(2)}(G) \times \cdots \times \mathcal{M}^{(K)}(G). \tag{2.3}$$

By Theorem 2.2, all multivariate Gaussian vectors X correspond to the zero element of $\mathcal{M}^{(K)}(G)$ for $k \geq 3$.

When the errors in an LSEM are Gaussian, all distributional information is captured by the covariance matrix and equivalence issues arise that hinder identifiability of the full graph. It then becomes necessary to expand to non-Gaussian settings. Relaxing the constraint of Gaussianity gives rise to the class of LiNGAMs where the underlying graph now becomes identifiable [Shimizu et al., 2006, 2011]. We will exploit this property algorithmically and use the the signal provided by higher cumulants; we do this by way of *treks*.

Definition 2.6 (Multi-Trek)**.** *A k-trek between vertices $i_1, \dots, i_k \in V$ of a DAG $G = (V, E)$ is a collection of directed paths $T = (P_1, \dots, P_k)$ in G that share the same source and have i_j as the sink of P_j for all j. The common source node is the top of the trek* $\mathrm{top}(T)$. *A trek is simple if the top node the unique node on all the paths. Denote the set of k-treks between $i_1, \dots, i_k$ by $\mathcal{T}(i_1, \dots, i_k)$ and the set of simple treks by $\mathcal{S}(i_1, \dots, i_k)$.*

If P is a directed path in the DAG $G = (V, E)$ and $\Lambda = (\lambda_{ij}) \in \mathbb{R}^E$, then λ^P is the path monomial given by the product of λ_{ij} with $i \to j$ an edge in P. For a k-trek $T = (P_1, \dots, P_k)$, set $\lambda^T := \lambda^{P_1} \cdots \lambda^{P_k}$.

Proposition 2.7 (Multi-Trek Rule)**.** *The kth order cumulant tensor $\mathcal{C}^{(k)}(G)$ of X can be expressed as*

$$\mathcal{C}^{(k)}_{i_1,\dots,i_k}(G) = \sum \varepsilon^{(k)}_{\mathrm{top}(T)} \lambda^T, \tag{2.4}$$

where the sum is over all the treks T in $\mathcal{T}(i_1, \dots, i_k)$ and $\varepsilon^{(k)}_{\mathrm{top}(T)}$ denotes the $\mathrm{top}(T)$ *diagonal entry of $\varepsilon^{(k)}$.*

Proposition 2.7 follows from Lemma 2.4 and expanding the entries of $(I-\Lambda)^{-1}$ into sums of path monomials as in the usual trek rule for covariances [Robeva and Seby, 2021].

Corollary 2.8 (Simple Multi-Trek Rule)**.** *The kth order cumulant tensor $\mathcal{C}^{(k)}(G)$ of X can be expressed as*

$$\mathcal{C}^{(k)}_{i_1,\dots,i_k}(G) = \sum \mathcal{C}^{(k)}_{\mathrm{top}(S)}(G) \lambda^S, \tag{2.5}$$

where the sum is extended to all the simple treks S in $\mathcal{S}(i_1, \dots, i_k)$.

Corollary 2.9. *The ith diagonal entry of $\mathcal{C}^{(k)}$ is*

$$\mathcal{C}^{(k)}_i(G) = \sum_{p_1,\dots,p_k \in \mathrm{pa}(i)} \lambda_{p_1,i} \cdots \lambda_{p_k,i} \mathcal{C}^{(k)}_{p_1,\dots,p_k}(G) + \varepsilon^{(k)}_i.$$

2.3 POLYTREE MODELS

For general graphs, the algebraic relations among the cumulants may be far more complicated than the bivariate case (see Example A.1) and have not yet been fully characterized. However, there exists a generalization of the rank-one constraint for polytrees, which we now discuss.

By consequence of there being at most one directed path between any two noes of a polytree G, there is at most one simple trek between any set of nodes $i_1, \dots, i_k$. The simple multi-trek rule then reduces to $C^{(k)}_{i_1,\dots,i_k}(G) = \lambda^S \mathcal{C}^{(k)}_{\mathrm{top}(S)}$ for a trek between nodes with S being the unique simple trek; denote the top of the simple trek between $i_1, \dots, i_k$, if it exists by $\mathrm{top}(i_1, \dots, i_k)$. Also, $C^{(k)}_{i_1,\dots,i_k}(G) = 0$ if there is no k-trek between the nodes.

For any two vertices $i \neq j$, let $c^{(i,j),k}_m$ denote the kth order cumulant $\mathcal{C}^{(k)}_{i\dots i,j\dots j}(G)$, where the first m indices are equal to i and the remaining $m-k$ equal j.

Proposition 2.10. *Let $e : i \to j$ be an edge of a polytree G. Then the following matrix is of rank one*

$$A^{e,K} = \begin{bmatrix} c^{e,k}_m \\ c^{e,k}_{m-1} \end{bmatrix} \mid 2 \leq m \leq k \leq K \Bigg]. \tag{2.6}$$

The first column of $A^{e,K}$ contains $\mathbb{E}[X_i^2] > 0$. Moreover, for every distribution induced by non-Gaussian errors, there exists k such that $\mathcal{C}^{(k)}_i \neq 0$. Hence, at least one minor of $A^{e,K}$ gives us an equation that is satisfied if $i \to j$ is in G, and is not satisfied in general for the graph with the edge reversed. This observation will provide the foundation for our learning algorithm, which we now present.

3 LEARNING NON-GAUSSIAN POLYTREES FROM MOMENTS

We now present our population algorithm for learning polytrees with three versions for learning the edge orientations. The first common phase is skeleton recovery.

3.1 LEARNING THE SKELETON

In its original formulation, the Chow–Liu algorithm gives the maximum likelihood tree approximation of a given discrete distribution [Chow and Liu, 1968]. The tree obtained is the maximum weight spanning tree of the complete undirected graph with edge weights $w(i,j)$, given by the mutual information between X_i and X_j. [Rebane and Pearl, 1987, Theo. 1] show that under a non-degeneracy assumption, the same Chow–Liu algorithm can be used to recover skeletons in the polytree setting. Their proof is based on the following property of the mutual information.

Proposition 3.1. *If the polytree that defines the model contains the subgraph $i \to j \to l$ or $i \leftarrow j \to l$, then*

$$\min\{I(X_i, X_j), I(X_j, X_l)\} > I(X_i, X_l),$$

where $I(\cdot,\cdot)$ is the mutual information.

When working with an LSEM, a stronger result justifies the use of the absolute value of the correlation coefficient instead of the mutual information.

Lemma 3.2 (Wright's Formula, [Wright, 1960])**.** *In the LSEM defined by a polytree, the correlation $\rho_{i,j} = Corr[X_i, X_j]$ satisfies*

$$|\rho_{i,j}| = \begin{cases} \prod |\rho_e|, & \mathcal{T}(i,j) \neq \emptyset, \\ 0, & \textit{otherwise}, \end{cases} \tag{3.1}$$

where the product is taken over the edges of the unique trek from i to j, and ρ_e denotes the correlation between the random variables indexed by the endpoints of the edge e.

Definition 3.3. *Let $R = (\rho_{i,j})$ be the correlation matrix of a random vector X. The Chow–Liu tree $\mathcal{M}(R)$ is the (undirected) maximum weight spanning tree over $[p]$, with weights given by $|\rho_{i,j}|$.*

Kruskal's algorithm may be applied to compute the Chow–Liu tree Kruskal [1956].

Proposition 3.4. *Let $R = (\rho_{i,j})$ be the correlation matrix of a random vector $X = (X_1, \dots, X_p)$ that follows the LSEM given by a polytree G. If $0 < |\rho_{i,j}| < 1$ for every $e : i \to j \in E$, then $\mathcal{M}(R)$ equals the skeleton of G.*

The assumption $|\rho_{i,j}| < 1$ holds for all random vectors with positive definite covariance matrix. Moreover, in a polytree model, $|\rho_{i,j}| > 0$ for an edge (i,j) if $\lambda_{ij} \neq 0$.

3.2 LEARNING ORIENTATIONS

We now present three ways to orient the edges in the estimated skeleton. The three resulting orientation algorithms are all based on Proposition 2.10 and the following result.

Theorem 3.5. *Consider the LSEM given the polytree G, and let $e : i \to j$ be an edge of G. Then*

(i) $\operatorname{rank}(A^{i\to j,K}) = 1$,

(ii) $\operatorname{rank}(A^{j\to i,K}) = 2$, *for generic edge coefficients and error cumulants up to order K.*

Proposition 3.6. *Suppose the skeleton of the polytree G contains the subgraph $i - j - l$ with $\rho_{i,j}, \rho_{j,l} \neq 0$. Then the corresponding subgraph of G is $i \to j \leftarrow l$ iff $\rho_{i,l} = 0$.*

We now present *PairwiseOrientation_Pop*; Algorithm 1. This algorithm takes as input the list of unoriented edges and the parameter $K \geq 3$, which defines the highest order cumulant used in $A^{i\to j,K}$. It orients each edge separately by checking whether the rank of $A^{i\to j,K}$ is 1 or not.

Algorithm 1 PairwiseOrientation_Pop(E, K)

```
O ← ∅
for (i, j) ∈ E do
    if rank(A^{i→j,K}) = 1 then
        O ← O ∪ {i → j}
    else
        O ← O ∪ {j → i}
return O
```

Our second algorithm *PTO_Pop*, Algorithm 2, proceeds recursively. At each step, it takes as input the order K, a list of already oriented edges O, a list of still unoriented edges E, and, possibly, an oriented edge o. Here $t(o)$ is the target/sink of the edge and $E \cap t(o)$ is the (possibly empty) set of unoriented edges containing $t(o)$. The procedure checks if there are unoriented edges, and if so, it searches for triplets of the form $i \to j - k$, where the oriented edge $o = i \to j$ can come either from the previous call of the procedure or from checking the rank of $A^{i\to j,K}$. For such a triplet, the method checks if $\rho_{i,k} = 0$, orienting the other edge according to the result. The algorithm is initialized with $O = o = \emptyset$ and the full list of undirected edges, E.

Our third proposed algorithm *TPO_Pop*, Algorithm 3, can be seen as a direct extension of learning completed partially directed graphs (CPDAG)—a mixed graph that encodes the causal information common to all the members of a Markov equivalence class. Here, we first compute the CPDAG as in [Rebane and Pearl, 1987], then we orient all remaining undirected edges by checking the rank of $A^{i\to j,K}$. This ensures that no other unshielded colliders appear.

The following example compares our three algorithms.

Algorithm 2 PTO_Pop(E, K, O, o)

```
if E ≠ ∅ then
    if o = ∅ then
        (i, j) ← E[1]
        if rank(A^{i→j,K}) = 1 then
            o ← (i → j)
            O ← O ∪ {o}
        else
            o ← (j → i)
            O ← O ∪ {o}
    E_o ← E ∩ t(o)
    if E_o ≠ ∅ then
        E ← E \ E_o
        for t(o) − k ∈ E_o do
            if ρ_{s(o),k} = 0 then
                O ← O ∪ {k → t(o)}
            else
                O ← O ∪ {t(o) → w}
                o ← (t(o) → w)
                O, E ← PTO_Pop(E, K, O, o)
    O, E ← PTO_Pop(E, K, O, ∅)
return O,E
```

Algorithm 3 TPO_Pop(E, K)

```
O ← ∅
for i − j − k ∈ E do
    if ρ_{i,k} = 0 then
        E ← E \ {(i, j), (j, k)}
        O ← O ∪ {i → j, k → j}
for i → j ∈ O do
    for j − l ∈ E do
        E ← E \ {(j, l)}
        O ← O \ {j → l}
for (i, j) ∈ E do
    if rank(A^{i→j,K}) = 1 then
        O ← O ∪ {i → j}
        for j − l ∈ E do
            E ← E \ {(j, l)}
            O ← O \ {j → l}
    else
        O ← O ∪ {j → i}
        for i − l ∈ E do
            E ← E \ {(i, l)}
            O ← O \ {i → l}
return O
```

Example 3.7. *Consider the graph G with $1 \to 2 \to 3 \leftarrow 4$. With the skeleton $1-2-3-4$ inferred, the algorithm PairwiseOrientation_Pop sequentially computes the rank of $A^{i\to j,K}$ in the order of all edges and orients them according to the results. PTO_Pop orients $1-2$ using the rank condition and then checks if $\rho_{1,3}=0$. Since this is not the case, it orients $2-3$ using the rank condition and then $3-4$ checking that $\rho_{2,4}=0$. Finally, TPO_Pop first computes $\rho_{1,3}$ and $\rho_{2,4}$. Since $\rho_{2,4}=0$, it orients $2-3-4$, and then orients $1-2$ with the rank condition.*

Theorem 3.8. *The three versions of the algorithm are correct for generic edge coefficients and cumulants up to order K.*

4 LEARNING NON-GAUSSIAN POLYTREES FROM DATA

We now consider the empirical versions of our algorithms, which now learn a polytree from a dataset consisting of n i.i.d. random vectors. The algorithms then process the sample correlations $\hat{\rho}_{i,j}$ and sample cumulants $\hat{c}_m^{(i,j),k}$. Let $\hat{\Sigma}_{i,j}$ be the unbiased sample covariances. Then $\hat{\rho}_{i,j} = \Sigma_{i,j}/\sqrt{\Sigma_{i,i}\Sigma_{j,j}}$. Generalizing sample covariances, we take the sample cumulants $\hat{c}_m^{(i,j),k}$ to be the k-statistics that unbiasedly estimate $c_m^{(i,j),k}$ [McCullagh, 1987, §4.2].

We provide consistency results in a high dimensional setting where the size of the polytree grows at a faster rate than the sample size, subject to log-concavity of the variables. Specifically, we assume the errors ε_i and thus also the observation vector X are log-concave distributed. This setting allows for the following corollary that builds on the concentration inequality given in Lemma B.3 of Lin et al. [2016].

Corollary 4.1. *Let $K \in \mathbb{N}$ and suppose that all moments up to order $2K$ of the random vector X are bounded in magnitude by a constant $M_K > 0$. For any $k \le K$, if $\hat{c}_m^{(i,j),k} = c_m^{(i,j),k} + \epsilon_m^{(i,j),k}$ is the k-statistic for a sample of size n, then for every $\delta > 0$ where $\frac{2}{LK^2\sqrt{M_K}}\left(\frac{\delta\sqrt{n}}{e}\right)^{\frac{1}{K}} > 2$, we have*

$$\mathbb{P}[|\epsilon_m^{(i,j),k}| > \delta] \le \exp\left\{-\frac{2}{LK^2\sqrt{M_K}}\left(\delta\sqrt{n}\right)^{\frac{1}{K}}\right\}.$$

4.1 LEARNING THE SKELETON CONSISTENTLY

Let $\rho_{\min}$ and $\rho_{\max}$ be the respective minimum and maximum of the absolute edge correlations in the set $\{|\rho_{i,j}| : i \to j \in E\}$ with $0 < \rho_{\min}, \rho_{\max} < 1$. We will use the following lemma on the correctness of the Chow–Liu tree $\mathcal{M}(R_n)$ computed from the sample correlation matrix $R_n = (\hat{\rho}_{i,j})$, together with Lemma 7 from Harris and Drton [2013], both restated below.

Lemma 4.2. *Let $\gamma = \rho_{\min}(1-\rho_{\max})/2$. Then the event $F := \bigcap\{|\hat{\rho}_{i,j} - \rho_{i,j}| \le \gamma\}$ satisfies $F \subset \{\mathcal{M}(\hat{R}_n) = \mathcal{S}(G)\}$.*

Lemma 4.3. *If A, B are 2×2 symmetric matrices, with A positive definite, $a_{1,1}, a_{2,2} \ge 1$, and $||A-B||_\infty < \delta$, then*

$$\left|\frac{a_{1,2}}{\sqrt{a_{1,1}a_{2,2}}} - \frac{b_{1,2}}{\sqrt{b_{1,1}b_{2,2}}}\right| < \frac{2\delta}{1-\delta}. \tag{4.1}$$

We now have the following consistency result for the Chow–Liu tree $\mathcal{M}(R_n)$.

Theorem 4.4. *Let $\lambda := \min\{\min_i \Sigma_{i,i}, 1\}$ and let γ and M_2 be defined as in Lemma 4.2 and Corollary 4.1 respectively, then*

$$\mathbb{P}(\mathcal{M}(R_n) = \mathcal{S}(G)) \ge 1 - \frac{3p(p-1)}{2}\exp\left\{-\frac{1}{2L\sqrt{M_2}}\left(\frac{\lambda\gamma\sqrt{n}}{2+\lambda}\right)^{\frac{1}{2}}\right\},$$

for all $n > \frac{e^2(2+\lambda)^2(4L^2\sqrt{M_2})^4}{\lambda^2\gamma^2}$.

4.2 LEARNING ORIENTATIONS CONSISTENTLY

For every edge $e = \{i,j\}$ in the skeleton $\mathcal{S}(G)$, let $v_r(e), v_w(e) \in \mathbb{R}^{B(K)}$ be the vectors containing the minors of $A^{(r(e)),K}$ and $A^{(w(e)),K}$ involving the first column. Here, $r(e)$ and $w(e)$ are the correct and incorrect orientations of e in G, respectively. Let $B(K) = K(K-1)/2-1$ be the size of the vectors. We assume $\exists\,\delta > 0$ such that $||v_w(e)|| > \delta$ for all $e \in \mathcal{S}(G)$, where $||\cdot||$ is the 2-norm. Let M_K, L, and $\epsilon_m^{(i,j),k}$ be defined as in Corollary 4.1. Moreover, let c be the vector containing all the cumulants $c_m^{(i,j),k}$ such that edge $\{i,j\} \in \mathcal{S}(G)$ and $0 \le m \le k \le K$. Write $\hat{c}_n$ for the vector containing the sample versions of these cumulants. Finally, let ϵ_n be the corresponding error vector tracking the differences between the true and sample cumulants.

Lemma 4.5. *If f is the difference of two monomials of degree 2 in the variables c, then*

$$|f(c+\epsilon_n) - f(c)| \le 4M_K||\epsilon_n||_\infty + 2||\epsilon_n||_\infty^2. \tag{4.2}$$

For use with data, the proposed algorithms in Section 3 must be modified to allow for sampling variability. In particular, instead of checking whether or not $\mathrm{rank}(A^{i\to j,K}) = 1$, we check $||\hat{v}_{i\to j}((i,j))|| < ||\hat{v}_{j\to i}((i,j))||$ instead. Similarly, for the independence test, we check whether or not the absolute correlation is below a threshold ρ_{thr}; Lemma 4.8 clarifies the possible choices of the threshold. The sample versions of the algorithms are given in Appendix B.

Let $\mathcal{A}_n^{PO}(E,K)$ be the output of Algorithm 4 applied to a sample of size n and let $E_{\mathcal{S}(G)}$ be the edge set of the true skeleton of G. Then we have the following consistency result.

Lemma 4.6. *Let $\delta' := \min\{\frac{\delta}{4M_K\sqrt{B(K)}}, \frac{\sqrt{\delta}}{\sqrt[4]{4B(K)}}\}$. Then*

$$\mathbb{P}(\mathcal{A}_n^{PO}(E_{\mathcal{S}(G)}, K) = G) \ge 1 - 4B(K)(p-1)\exp\left\{-\frac{2}{LK^2\sqrt{M_K}}\left(\delta'\sqrt{n}\right)^{\frac{1}{K}}\right\},$$

for all $n > \frac{e^2(LK^2\sqrt{M_K})^{2K}}{\delta'^2}$.

Theorem 4.7. *Suppose the data are an n-sample drawn from a distribution in the LSEM given by a polytree G. Let $\hat{G}$ be the polytree obtained by applying Algorithm 4 to the (undirected) edge set of the Chow–Liu tree $\mathcal{M}(R_n)$. Then $\hat{G} = G$ with probability greater than*

$$1 - 4B(K)(p-1)\exp\left\{-\frac{2}{LK^2\sqrt{M_K}}\left(\delta'\sqrt{n}\right)^{\frac{1}{K}}\right\} - \frac{3p(p-1)}{2}\exp\left\{-\frac{1}{2L\sqrt{M_2}}\left(\frac{\lambda\gamma\sqrt{n}}{2+\lambda}\right)^{\frac{1}{2}}\right\},$$

for all $n > \max\left\{\frac{e^2(2+\lambda)^2(4L^2\sqrt{M_2})^4}{\lambda^2\gamma^2}, \frac{e^2(LK^2\sqrt{M_K})^{2K}}{\delta'^2}\right\}$, with constants defined in Lemma 4.6 and Theorem 4.4.

Lemma 4.8. *Let $\tilde{\gamma} = \min\{\rho_{\min}/3, (1-\rho_{\max})/2\}\rho_{\min}$, and let λ and M_2 be as in Theorem 4.4. If $\tilde{\gamma} < \rho_{\text{thr}} < \rho_{\min}^2 - \tilde{\gamma}$, then the probability that all independence tests carried out by Algorithm 5 yield correct decisions is bounded from below by*

$$1 - \frac{3p(p-1)}{2}\exp\left\{-\frac{1}{2L\sqrt{M_2}}\left(\frac{\lambda\tilde{\gamma}\sqrt{n}}{2+\lambda}\right)^{\frac{1}{2}}\right\},$$

for all $n > \frac{e^2(2+\lambda)^2(4L^2\sqrt{M_2})^4}{\lambda^2\tilde{\gamma}^2}$. The same statement holds for Algorithm 6.

Theorem 4.9. *Suppose the data are an n-sample drawn from a distribution in the LSEM given by a polytree G. Let $\hat{G}$ be the polytree obtained by applying Algorithm 5 or 6 to the (undirected) edge set of the Chow–Liu tree $\mathcal{M}(R_n)$. If the threshold satisfies $\tilde{\gamma} < \rho_{\text{thr}} < \rho_{\min}^2 - \tilde{\gamma}$, then there exists $\alpha^* < p-1$ such that $\hat{G} = G$ with probability greater than*

$$1 - 4B(K)\alpha^*\exp\left\{-\frac{2}{LK^2\sqrt{M_K}}\left(\delta'\sqrt{n}\right)^{\frac{1}{K}}\right\} - \frac{3p(p-1)}{2}\exp\left\{-\frac{1}{2L\sqrt{M_2}}\left(\frac{\lambda\tilde{\gamma}\sqrt{n}}{2+\lambda}\right)^{\frac{1}{2}}\right\},$$

for all $n > \max\left\{\frac{e^2(2+\lambda)^2(4L^2\sqrt{M_2})^4}{\lambda^2\tilde{\gamma}^2}, \frac{e^2(LK^2\sqrt{M_K})^{2K}}{\delta'^2}\right\}$.

5 NUMERICAL EXPERIMENTS

We assess and compare the accuracy of our three proposed algorithms on synthetic data, simulated as follows: For any

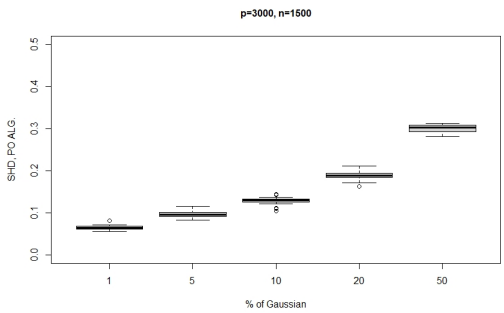


Figure 1: Performance for high dimensional experiments with varying % of Gaussian random variables over 25 runs.

fixed choice of n, p, and error distribution, we first generate a random undirected tree with p nodes using randomly generated Prüfer sequences [Prüfer, 1918] and then independently orient each edge. Next, we draw n samples for every node from the error distribution and uniformly draw the coefficients λ_{ij} from the interval $(-1, -0.3) \cup (0.3, 1)$. Finally, we multiply the matrix of sampled errors by the matrix $(I - \Lambda)^{-1}$ to obtain samples corresponding to the LSEM defined by the generated polytree.

The performance is measured by the structural Hamming distance, which is the number of incorrectly included edges plus the number of falsely omitted edges plus the number of incorrectly oriented edges divided by $2(p-1)$. Small distance indicates improved performance. We show the results in three settings: (i) low dimensional, with $25 \leq p \leq 200$ and $1 \leq n/p \leq 300$; (ii) high dimensional, with $500 \leq p \leq 3000$ and $0.5 \leq n/p \leq 1$; and (iii) a large scale setting with $10000 \leq p \leq 20000$ and $n/p = 0.1$. We set up experiments with errors drawn from the gamma and uniform distributions; the results are displayed in Figures 2, 3, and 4.

For choice of threshold required in Algorithms 5 and 6, we evaluate the algorithms on a grid of thresholds and report the value corresponding to the best result.

Gamma Distribution. Errors were drawn from $\Gamma(\alpha, \beta)$; the shape α and the scale β parameters are uniformly drawn from $(0.5, 5)$. Since $\Gamma(\alpha, \beta)$ is asymmetric, we tested the algorithms with $K = 3$.

The experimental results are coherent with our developed theory: for all three algorithms, the distance between the true and learned trees converges to 0 as the sample size and/or the dimension of the tree increases. We observe that Algorithm 4 performs better both in mean accuracy and variance, despite heavily relying on higher moments which is statistically disadvantageous. The improved performance is potentially due to the tuning-free nature of this algorithm compared to the other where tuning of the threshold parameter is required. Additionally, and opposed to the latter two algorithms, Algorithm 4 avoids any potential error propagation since the edges are oriented independently.

Uniform Distribution. Errors were drawn from $U(a, b)$, with the parameter a uniformly drawn from $(-10, -1)$ and b uniformly drawn from $(1, 10)$. Here, the uniform distribution is symmetric so the 3rd order cumulants are 0; we thus tested the algorithms with $K = 4$.

The experimental results here are also consistent with our developed theory. We also notice that overall, the experiments with uniform errors outperform those with gamma errors, which may be due to the greater higher order moments associated with the gamma distribution, which tend to increase the variance of the sample cumulants in Corollary 4.1.

6 CONCLUSION

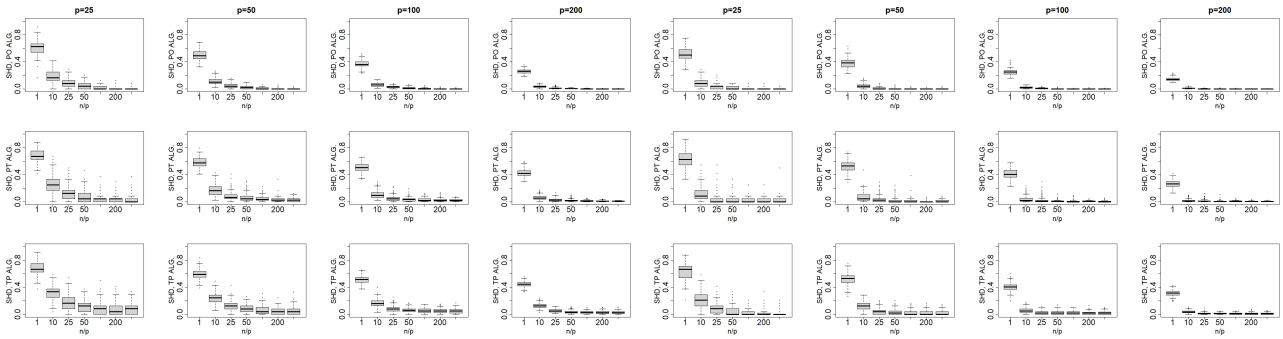
In this paper, we proposed three algorithmic schemes that learn linear non-Gaussian polytree models by using the Chow–Liu algorithm to infer the graph skeleton and then applying different approaches to orient edges leveraging non-Gaussianity and/or marginal uncorrelatedness. The algorithms differ from one another in how much information is taken from correlations versus from higher moments. The numerical experiments show that the algorithms are effective also in very high-dimensional problems.

Our work motivates the following questions for future research:

How to best handle Gaussian random variables when learning a polytree? In some settings we may encounter the situation that some but not all errors are non-Gaussian; see Hoyer et al. [2008] for a characterization of equivalence of graphs in this case. An interesting problem is then to determine how the respective performance of our algorithms is affected by such partial Gaussianity, and provide modifications that effectively learn a polytree equivalence class. As an illustration, Figure 1 shows that Algorithm 4 achieves 70% accuracy when a random choice of half of the random variables are allowed to be Gaussian.

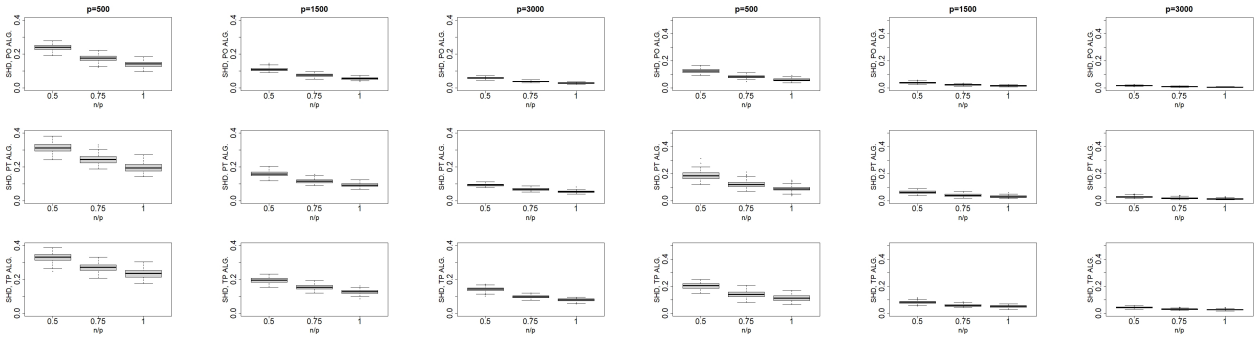
Which tree structures are the most difficult to learn? Tan et al. [2009] show that for undirected Gaussian tree models, the star and the chain represent the most difficult and the easiest trees to learn, respectively. The difficulty is due to the correlation decay. An interesting question to pursue is what the polytree analogues for the most difficult and easiest trees to learn would be.

What happens when the graph is not a tree? Boix-Adserà et al. [2021] prove a weakness result of the Chow–Liu algorithm under model misspecification for the Ising model and adapt it to achieve a form of optimality. It would be of interest to describe a similar optimality criterion in the LiNGAM setting and investigate how our algorithm performs under these terms.



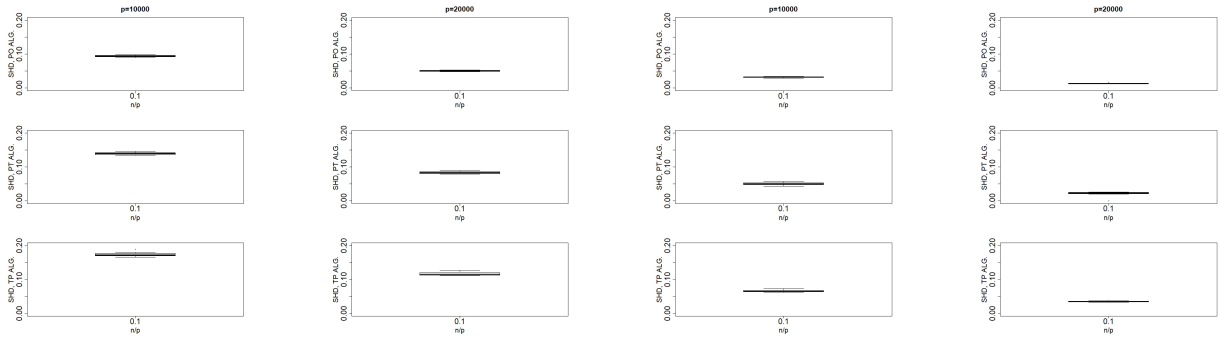
(a) Gamma Distribution (b) Uniform Distribution

Figure 2: Performance for low dimensional experiments with $25 \leq p \leq 200$ and $1 \leq n/p \leq 300$ over 200 runs.



(a) Gamma Distribution (b) Uniform Distribution

Figure 3: Performance for high dimensional experiments with $500 \leq p \leq 3000$ and $0.5 \leq n/p \leq 1$ over 100 runs.



(a) Gamma Distribution (b) Uniform Distribution

Figure 4: Performance for large scale experiments with $10000 \leq p \leq 20000$ and $n/p = 0.1$ over 10 runs.

References

Carlos Améndola, Mathias Drton, Alexandros Grosdos, Roser Homs, and Elina Robeva. Third-order moment varieties of linear non-Gaussian graphical models, 2021.

Enric Boix-Adserà, Guy Bresler, and Frederic Koehler. Chow-liu++: Optimal prediction-centric learning of tree ising models. *ArXiv*, abs/2106.03969, 2021.

C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968.

Pierre Comon and Christian Jutten. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press, Inc., USA, 1st edition, 2010.

Mathias Drton and Marloes H. Maathuis. Structure learning in graphical modeling. *Annu. Rev. Stat. Appl.*, 4:365–393, 2017.

Naftali Harris and Mathias Drton. PC algorithm for nonparanormal graphical models. *J. Mach. Learn. Res.*, 14: 3365–3383, 2013.

Patrik O. Hoyer, Aapo Hyvärinen, Richard Scheines, Peter L. Spirtes, Joseph Ramsey, Gustavo Lacerda, and Shohei Shimizu. Causal discovery of linear acyclic models with arbitrary distributions. In *UAI*, 2008.

Aapo Hyvärinen and Stephen M. Smith. Pairwise likelihood ratios for estimation of non-Gaussian structural equation models. *J. Mach. Learn. Res.*, 14, 2013.

Martin Emil Jakobsen, Rajen D. Shah, Peter Bühlmann, and Jonas Peters. Structure learning for directed trees, 2021.

Joseph B. Kruskal, Jr. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. Amer. Math. Soc.*, 7:48–50, 1956.

Lina Lin, Mathias Drton, and Ali Shojaie. Estimation of high-dimensional graphical models using regularized score matching. *Electron. J. Stat.*, 10(1):806–854, 2016.

Xingmei Lou, Yu Hu, and Xiaodong Li. Linear polytree structural equation models: Structural learning and inverse correlation estimation, 2021.

Marloes Maathuis, Mathias Drton, Steffen Lauritzen, and Martin Wainwright, editors. *Handbook of graphical models*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL, 2019.

Peter McCullagh. *Tensor methods in statistics*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1987. ISBN 0-412-27480-9.

Masashi Okamoto. Distinctness of the Eigenvalues of a Quadratic form in a Multivariate Sample. *The Annals of Statistics*, 1(4):763 – 765, 1973.

Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. The Morgan Kaufmann Series in Representation and Reasoning. Morgan Kaufmann, San Mateo, CA, 1988.

Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. *Causal inference in statistics*. John Wiley & Sons, Ltd., Chichester, 2016. A primer.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2017. Foundations and learning algorithms.

Heinz Prüfer. Neuer Beweis eines Satzes über Permutationen. *Archiv der Mathematischen Physik*, 27:742–744, 1918.

George Rebane and Judea Pearl. The recovery of causal poly-trees from statistical data. In *Int. J. Approx. Reason.*, 1987.

Elina Robeva and Jean-Baptiste Seby. Multi-trek separation in linear structural equation models. *SIAM J. Appl. Algebra Geom.*, 5(2):278–303, 2021.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-Gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.*, 7:2003–2030, 2006.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O. Hoyer, and Kenneth Bollen. DirectLiNGAM: a direct method for learning a linear non-Gaussian structural equation model. *J. Mach. Learn. Res.*, 12:1225–1248, 2011.

Vincent Y. F. Tan, Animashree Anandkumar, and Alan S. Willsky. How do the structure and the parameters of gaussian tree models affect structure learning? In *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 684–691, 2009. doi: 10.1109/ALLERTON.2009.5394929.

Y. Samuel Wang and Mathias Drton. High-dimensional causal discovery under non-Gaussianity. *Biometrika*, 107 (1):41–59, 2020.

Sewall Wright. Path coefficients and path regressions: Alternative or complementary concepts? *Biometrics*, 16(2): 189–202, 1960.

A LEMMAS AND PROOFS

A.1 SECTION 2

Example A.1. *Assume we have two random variables X_1 and X_2, coming from an unknown xSEM, and we wish to find the DAG G that generates the model. Assume further that X_1 and X_2 are correlated, leaving us with the two options $G_1 = ([2], \{1 \to 2\})$ and $G_2 = ([2], \{2 \to 1\})$. Under the model given by G_1, we have $X_1 = \varepsilon_1$ and $X_2 = \lambda_{12}X_1 + \varepsilon_2$, and the covariance matrix from* (2.2) *becomes*

$$\Sigma = \begin{bmatrix} \varepsilon_1^{(2)} & \lambda_{12}\varepsilon_1^{(2)} \\ \lambda_{12}\omega_1 & \lambda_{12}^2\varepsilon_1^{(2)} + \varepsilon_2^{(2)} \end{bmatrix}. \tag{A.1}$$

Observe that any positive definite 2×2 matrix Σ can be written in this way; set $\varepsilon_1^{(2)} = \Sigma_{11} > 0$, $\lambda_{12} = \Sigma_{21}/\Sigma_{11}$ and $\varepsilon_2^{(2)} = \Sigma_{22} - \Sigma_{12}^2/\Sigma_{11} > 0$. By symmetry, the model given by G_2 also allows its covariance matrices to be any positive definite 2×2 matrix. Hence, the sets of covariance matrices resulting from G_1 versus G_2 are the same, and nothing can be said about the graph on the basis of covariances alone.

Proof of Corollary 2.8. We say that a k-trek $T = (P_1, \ldots, P_k)$ factorizes through k-trek $S = (Q_1, \ldots, Q_k)$ if $Q_j \subset P_j$ for all j. Indeed, it is easy to see that $\lambda^T = \lambda^S \lambda^{T-S}$, where $T - S = (P_1 - Q_1, \ldots, P_k - Q_k) \in \mathcal{T}^S = \mathcal{T}(\mathrm{top}(S), \ldots, \mathrm{top}(S))$ with $P_j - Q_j$ being the directed path from $\mathrm{top}(T)$ to $\mathrm{top}(S)$ that remains after removing the edges in Q_j from P_j.

For the sake of readability, when the considered set of vertices is clear from the context, we denote the set of treks (and simple treks) by only $\mathcal{T}$ (and $\mathcal{S}$). Now note that every trek factorizes along one and only one simple trek. Hence, we can partition $\mathcal{T} = \dot{\bigcup}_{S \in \mathcal{S}} \mathcal{T}^S$, according to the simple trek through which the factorization occurs. The expression in (2.4) may thus be rewritten as

$$\mathcal{C}^{(k)}_{i_1,\ldots,i_k}(G) = \sum_{S \in \mathcal{S}} \lambda^S \left(\sum_{T \in \mathcal{T}^S} \varepsilon^{(k)}_{\mathrm{top}(T)} \lambda^{T-S} \right).$$

By the multi-trek rule, the term in parentheses is $\mathcal{C}^{(k)}_{\mathrm{top}(S)}(G)$. □

Proof of Corollary 2.9. Since the graph is acyclic, the only trek in $\mathcal{T}(i, .., i)$ that has i as *top* is the trivial trek, from which come the $\varepsilon_i^{(k)}$ term. All the other treks in $\mathcal{T}(i, .., i)$ factorize through a set of distinct parents of i, so we can write the sum in 2.4 as

$$\sum_{p_1,\ldots,p_k \in \mathrm{pa}(i)} \lambda_{p_1,i} \cdots \lambda_{p_k,i} \Big(\sum_{T \in \mathcal{T}(p_1,..,p_k)} \lambda^T \varepsilon^{(k)}_{top(T)} \Big) + \varepsilon_i^{(k)}$$

and the term inside the internal parenthesis is $\mathcal{C}^{(k)}_{p_1,\ldots,p_k}(G)$. □

Proof of Proposition 2.10. By the simple multi-trek rule for polytrees $c_m^{e,k} = \lambda_{ij}^{k-m}\mathcal{C}_i^{(k)}$. Therefore, $c_m^{e,k} = \lambda_{ij} c_{m+1}^{e,k}$ so that the second row of $A^{e,K}$ equals λ_{ij} times the first row. □

A.2 SECTION 3

Proof of Proposition 3.4. As noted, $\mathcal{M}(R)$ may be computed using Kruskal's algorithm, which considers all edges in decreasing order of their weights and adds them to the spanning tree if their presence does not create a (undirected) cycle.

Let $\mathcal{S}(G)$ be the skeleton of G. For a contradiction, assume that $\mathcal{M}(R) \neq \mathcal{S}(G)$. Since both graphs are trees, we have $L := \mathcal{M}(R) \setminus \mathcal{S}(G) \neq \emptyset$. Take $\tilde{e} = \{i, j\} = \arg\max_{e \in L} |\rho_e|$. Then $|\rho_{\tilde{e}}| \neq 0$ and, thus, $\mathcal{T}(i, j) \neq \emptyset$. The unique trek in G that connects i and j must contain an edge e that is not in $\mathcal{M}(R)$; otherwise we would have two paths between i and j in $\mathcal{M}(R)$ which cannot occur as $\mathcal{M}(R)$ is a tree. Moreover, Wright's formula from Lemma 3.2 and the assumption made on the correlation coefficients imply that $|\rho_{\tilde{e}}| < |\rho_e|$. But then e would appear before $\tilde{e}$ in Kruskal's algorithm and e would be added to $\mathcal{M}(R)$ since all the edges in $\mathcal{M}(R)$ with a weight higher than the weight of $\tilde{e}$ are correctly classified and so the presence of e could not create a loop. We have arrived at a contradiction. □

Proof of Theorem 3.5. The first claim is a restatement of Proposition 2.10. To prove the second claim, we need to show that for generic error cumulants, at least one of the 2×2 subdeterminants of $A^{j \to i,K}$ is nonzero. Since these minors are polynomials in the cumulants, it is enough to prove that they are not identically zero; see, e.g., the lemma in Okamoto [1973].

By the simple multi-trek rule (Corollary 2.8),

$$\begin{aligned} c_i^{j \to i,2} &= \Sigma_{i,j} = \lambda_{i,j}\Sigma_{i,i}, \\ c_m^{j \to i,k} &= \lambda_{i,j}^m \mathcal{C}_i^{(k)} \quad \forall m < k \leq K. \end{aligned}$$

By Corollary 2.9,

$$\begin{aligned} c_2^{j \to i,2} = \Sigma_{j,j} &= \sum_{p,q \in \mathrm{pa}(i)} \lambda_{p,j}\lambda_{q,j}\Sigma_{p,q} + \varepsilon_j^{(2)} \\ &= \sum_{p \in \mathrm{pa}(j)} \lambda_{p,j}^2 \Sigma_{p,p} + \varepsilon_j^{(2)}, \end{aligned}$$

because in a polytree any two distinct parents p and q have $\mathcal{T}(p, q) = \emptyset$ and, thus, $\Sigma_{p,q} = 0$ by Lemma 3.2.

Consider now a minor involving the first column together

with any other column with $m < k$. This minor is

$$\begin{aligned}
&\Sigma_{j,j} c_{m-1}^{j\to i,k} - \Sigma_{i,j} c_m^{j\to i,k} \\
&= \Big[\sum_{p\in \mathrm{pa}(j)} \lambda_{p,j}^2 \Sigma_{p,p} + \varepsilon_j^{(2)} \Big] \lambda_{i,j}^{m-1} \mathcal{C}_i^{(k)} - \lambda_{i,j}\Sigma_{i,i}\lambda_{i,j}^m \mathcal{C}_i^{(k)} \\
&= \lambda_{i,j}^{m-1}\mathcal{C}_i^{(k)} \Big[\sum_{p\in \mathrm{pa}(j)} \lambda_{p,j}^2 \Sigma_{p,p} + \varepsilon_j^{(2)} - \lambda_{i,j}^2 \Sigma_{i,i} \Big] \\
&= \lambda_{i,j}^{m-1}\mathcal{C}_i^{(k)} \Big[\sum_{p\in \mathrm{pa}(j)\setminus i} \lambda_{p,j}^2 \Sigma_{p,p} + \varepsilon_j^{(2)} \Big].
\end{aligned}$$

The term in parentheses is always positive, while the front factor is nonzero provided $\lambda_{i,j} \neq 0$ and the error cumulants are chosen such that $\mathcal{C}_i^{(k)} \neq 0$ (e.g., take any distribution with $\varepsilon_k^{(k)} = 0$, for every $k \neq j$, and $\varepsilon_j^{(k)} \neq 0$). □

Proof of Proposition 3.6. The claim follows from Lemma 3.2 as $i \to j \leftarrow l$ is the only case in which there are no treks between i, l. □

Proof. From Theorem 3.5 we derive the correctness of the rank condition and thus of the entire algorithm *Pairwise-Orientation_Pop* as well as the relevant parts of the other two algorithms. Proposition 3.6 yields the correctness of the remaining parts of the other two algorithms. □

A.3 SECTION 4

Lemma A.2. *Consider a degree k polynomial $f(X) = f(X_1, \dots, X_m)$, where $X_1, \dots, X_m$ are possibly dependent random variables with log-concave joint distribution on $\mathbb{R}^m$. Then exists a constant $L > 0$ such that for all δ with*

$$\frac{2}{L}\left(\frac{\delta}{e\sqrt{\mathrm{var}[f(X)]}}\right)^{\frac{1}{K}} > 2,$$

we have

$$\begin{aligned}
&\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| > \delta] \\
&\leq \exp\left\{ -\frac{2}{L}\left(\frac{\delta}{\sqrt{\mathrm{var}[f(X)]}}\right)^{\frac{1}{K}} \right\}.
\end{aligned}$$

Proof of Corollary 4.1. The results follows by bounding the variance of $\hat{c}_m^{(i,j),k}$. To this end, we may express the variance as a linear combination of products of moments based on the definition of cumulants. We may then bound each product of moments by a power of M_K and note that no weight may exceed $(K-1)!$. Applying Stirling's approximation for the factorial and bounding the Bell numbers that count the number of summands gives the result. □

Proof of Theorem 4.4. Let F be the event defined in Lemma 4.2. Then

$$\begin{aligned}
&\mathbb{P}(\mathcal{M}_n(G) = \mathcal{S}(G)) \geq \mathbb{P}(E) \\
&\geq 1 - \sum_{i\leq j} \mathbb{P}(|\hat{\rho}_{i,j} - \rho_{i,j}| > \gamma),
\end{aligned}$$

where the last inequality comes from the union bound.

Correlations are scale-invariant, thus the value of $\rho_{i,j}$ and the distribution of $\hat{\rho}_{i,j}$ do not change under rescaling of the observed variables to $\tilde{X}_i = X_i/\lambda$. Now, we apply Lemma 4.3 with $A = \tilde{\Sigma}$ and $B = \tilde{\Sigma}^*$ as the new population and sample covariance matrices, respectively. Let $||\tilde{\epsilon}||_\infty = ||\tilde{\Sigma} - \tilde{\Sigma}^*||_\infty$. Then $||\tilde{\epsilon}||_\infty \leq \frac{||\epsilon^{(i,j),2}||_\infty}{\lambda}$, where we indicate with $\epsilon^{(i,j),2}$ the vector containing the errors $\epsilon_m^{(i,j),2}$, for $m = 0, 1, 2$. This allows us to write

$$\begin{aligned}
\mathbb{P}(|\rho_{i,j}^* - \rho_{i,j}| > \gamma) &\leq \mathbb{P}(||\tilde{\epsilon}||_\infty > \frac{\gamma}{2+\gamma}) \\
&\leq \mathbb{P}(||\epsilon^{(i,j),2}||_\infty > \frac{\lambda\gamma}{2+\gamma}) \\
&\leq 3\exp\left\{ -\frac{1}{2L\sqrt{M_2}}\left(\frac{\lambda\gamma\sqrt{n}}{2+\lambda}\right)^{\frac{1}{2}} \right\},
\end{aligned}$$

where the last inequality comes from a union bound and Corollary 4.1 with $K = 2$. □

Proof of Lemma 4.5. By Taylor expansion, we have

$$f(c+\epsilon) - f(c) = \nabla(f)_{|c} \cdot \epsilon + \frac{\epsilon^t \cdot H(f)_{|c} \cdot \epsilon}{2}, \quad \text{(A.2)}$$

where $H(f)_{|c}$ is the Hessian matrix of f computed in c. For the special quadratic polynomial f, the Hessian is constant with entries ± 1, and the gradient contains entries of c, possibly negated. The result follows by triangle inequality. □

Proof of Theorem 4.4. The event of correctly reconstructing the polytree G is the intersection of the two events of correctly recovering of the skeleton and correctly orienting all the edges. The probability of this intersection is bounded from below by the sum of the probabilities of the two events minus one. Theorem 4.4 and Lemma 4.6 imply the result. □

Proof of Lemma 4.6. Given that we start with the correct skeleton, a mistake in the orientation appears if and only if there is an edge, e in the skeleton for which $||\hat{v}_r(e)|| \geq ||\hat{v}_w(e)|$, so the union bound leads to the following lower bound

$$\begin{aligned}
&\mathbb{P}(\mathcal{A}_n^{PO}(E_{\mathcal{S}(G)}, K) = G) \\
&= \mathbb{P}(||\hat{v}_r(e)|| < ||\hat{v}_w(e)||, \forall e \in \mathcal{S}(G)) \\
&\geq 1 - (p-1)\mathbb{P}(||\hat{v}_r(e)|| \geq ||\hat{v}_w(e)||).
\end{aligned}$$

Now we need an upper bound on $\mathbb{P}(||\hat{v}_r(e)|| \geq ||\hat{v}_w(e)||)$, adding and subtracting $||v_w(e)||$ on the right hand side of the inequality, and using $||v_w(e)|| > \delta$, led to

$$\begin{aligned}
&\mathbb{P}(||\hat{v}_r(e)|| \geq ||\hat{v}_w(e)||) \\
&\leq \mathbb{P}(||\hat{v}_r(e)|| + (||v_w(e)|| - ||\hat{v}_w(e)||) \geq \delta) \\
&\leq \mathbb{P}(||\hat{v}_r(e)|| + (||v_w(e) - \hat{v}_w(e)||) \geq \delta) \\
&\leq \mathbb{P}(||\hat{v}_r(e)|| \geq \tfrac{\delta}{2}) + \mathbb{P}(||v_w(e) - \hat{v}_w(e)|| \geq \tfrac{\delta}{2}) \\
&= \mathbb{P}(||\hat{v}_r(e)||^2 \geq \tfrac{\delta^2}{4}) + \mathbb{P}(||v_w(e) - \hat{v}_w(e)||^2 \geq \tfrac{\delta^2}{4}).
\end{aligned}$$

Now, using that $v_r(e) = 0$ and applying Lemma 4.5, we obtain that

$$\begin{aligned}
&\mathbb{P}(||\hat{v}_r(e)||^2 \geq \tfrac{\delta^2}{4}) + \mathbb{P}(||v_w(e) - \hat{v}_w(e)||^2 \geq \tfrac{\delta^2}{4}) \\
&\leq \mathbb{P}(||\hat{v}_r^i(e)||^2 \geq \tfrac{\delta^2}{4B(K)}) + \\
&\qquad + \mathbb{P}(||v_w^i(e) - \hat{v}_w^i(e)||^2 \geq \tfrac{\delta^2}{4B(K)}) \\
&\leq 2\mathbb{P}(4M_K||\epsilon||_\infty + 2||\epsilon||_\infty^2 \geq \tfrac{\delta}{2\sqrt{B(K)}}),
\end{aligned}$$

where v^i is one of the components of the vector. Finally, we obtain that

$$\begin{aligned}
&\mathbb{P}(4M_K||\epsilon||_\infty + 2||\epsilon||_\infty^2 \geq \tfrac{\delta}{2\sqrt{B(K)}}) \\
&\leq \mathbb{P}(||\epsilon||_\infty \geq \tfrac{\delta}{4M_K\sqrt{B(K)}}) + \mathbb{P}(||\epsilon||_\infty \geq \tfrac{\sqrt{\delta}}{\sqrt[4]{4B(K)}}).
\end{aligned}$$

With $\delta' := \min\{\frac{\delta}{4M_K\sqrt{B(K)}}, \frac{\sqrt{\delta}}{\sqrt[4]{4B(K)}}\}$ we can bound the right hand side of the last inequality by

$$2\mathbb{P}(||\epsilon||_\infty \geq \delta') \leq 2B(K)\exp\left\{-\tfrac{2}{LK^2\sqrt{M_K}}\left(\delta'\sqrt{n}\right)^{\frac{1}{K}}\right\}$$

whenever $\frac{2}{LK^2\sqrt{M_K}}\left(\frac{\delta'\sqrt{n}}{e}\right)^{\frac{1}{K}} > 2$ holds, and this concludes the proof. □

Proof of Lemma 4.8. It suffices to show that under the event $\tilde{F} = \bigcap\{|\hat{\rho}_{i,j} - \rho_{i,j}| < \tilde{\gamma}\}$ all independence tests give the correct decision. This claim was proven in Theorem 3.5 of [Lou et al., 2021]. The rest of the proof is the same as for Theorem 4.4. □

Proof of Theorem 4.9. Having $\tilde{\gamma} \leq \gamma$ from Lemma 4.2 and Lemma 4.8, the probability that the skeleton is recovered correctly and that all the independence tests give the right answer is bounded from below by the probability of the event considered in Lemma 4.2. Hence, it only remains to bound the probability that all cumulant tests correctly orient their respective edge. This can be done as in Lemma 4.6. Finally, carrying over the arguments from the proof of Theorem 4.7 gives the result. Notice that in the case of Algorithm 5, the constant $\alpha^* < p - 1 - CP$, where CP is the number of directed edges in the CPAG associated to G. □

B SAMPLE VERSION OF THE ALGORITHMS

Algorithm 4 PairwiseOrientation(E, K)

$O \leftarrow \emptyset$
for $e = (i, j) \in E$ **do**
  if $||\hat{v}_{i\to j}((i,j))|| < ||\hat{v}_{j\to i}((i,j))||$ **then**
    $O \leftarrow O \cup \{i \to j\}$
  else
    $O \leftarrow O \cup \{j \to i\}$
return O

Algorithm 5 TPO$(E, K, \rho_{\text{thr}})$

$O \leftarrow \emptyset$
for $i - j - k \in E$ **do**
  if $|\rho_{i,k}| < \rho_{\text{thr}}$ **then**
    $E \leftarrow E \setminus \{(i,j),(j,k)\}$
    $O \leftarrow O \cup \{i \to j, k \to j\}$
for $i \to j \in O$ **do**
  for $j - l \in E$ **do**
    $E \leftarrow E \setminus \{j - l\}$
    $O \leftarrow O \setminus \{j \to l\}$
for $(i, j) \in E$ **do**
  if $\text{rank}(A^{i\to j,K}) = 1$ **then**
    $O \leftarrow O \cup \{i \to j\}$
    for $j - l \in E$ **do**
      $E \leftarrow E \setminus \{j - l\}$
      $O \leftarrow O \setminus \{j \to l\}$
  else
    $O \leftarrow O \cup \{j \to i\}$
    for $i - l \in E$ **do**
      $E \leftarrow E \setminus \{i - l\}$
      $O \leftarrow O \setminus \{i \to l\}$
return O

Algorithm 6 PTO$(E, K, O, o, \rho_{\text{thr}})$

if $E \neq \emptyset$ **then**
  if $o = \emptyset$ **then**
    $(i, j) \leftarrow E[1]$
    if $\text{rank}(A^{i\to j,K}) = 1$ **then**
      $o \leftarrow (i \to j)$
      $O \leftarrow O \cup \{o\}$
    else
      $o \leftarrow (j \to i)$
      $O \leftarrow O \cup \{o\}$
  $E_o \leftarrow E \cap t(o)$
  if $E_o \neq \emptyset$ **then**
    $E \leftarrow E \setminus E_o$
    for $t(o) - k \in E_o$ **do**
      if $|\rho_{s(o),k|} < \rho_{\text{thr}}$ **then**
        $O \leftarrow O \cup \{k \to t(o)\}$
      else
        $O \leftarrow O \cup \{t(o) \to w\}$
        $o \leftarrow (t(o) \to w)$
        $O, E \leftarrow PTO_Pop(E, K, O, o, \rho_{\text{thr}})$
  $O, E \leftarrow PTO_Pop(E, K, O, \emptyset, \rho_{\text{thr}})$
return O,E