Learning Linear Gaussian Polytree Models With Interventions

Daniele Tramontano, L. Waldmann, M. Drton, and Eliana Duarte

***Abstract*—We present a consistent and highly scalable local approach to learn the causal structure of a linear Gaussian polytree using data from interventional experiments with known intervention targets. Our methods first learn the skeleton of the polytree and then orient its edges. The output is a CPDAG representing the interventional equivalence class of the polytree of the true underlying distribution. The skeleton and orientation recovery procedures we use rely on second order statistics and low-dimensional marginal distributions. We assess the performance of our methods under different scenarios in synthetic data sets and apply our algorithm to learn a polytree in a gene expression interventional data set. Our simulation studies demonstrate that our approach is fast, has good accuracy in terms of structural Hamming distance, and handles problems with thousands of nodes.**

***Index Terms*—Causal discovery, interventions, linear structural equation model, polytrees.**

I. Introduction

THE DOMINANT approach in recent literature on causal discovery from interventional data is optimization of a model score. Although the scoring is straightforward in the sense that the optimization over DAGs refers to fully specified joint models for all observational and interventional data, the optimization landscape is very high-dimensional, making score-based algorithms infeasible for graphs with hundreds/thousands of nodes that are common in biological applications. This makes causal discovery difficult, in addition to many other challenges that remain such as departing from restrictive genericity assumptions on the underlying distributions and developing methodology for high-dimensional settings. In this article, to address these challenges, we depart from a score-based strategy and leverage special properties of polytrees to obtain a highly scalable "local" approach that learns from low-dimensional marginals.

Our methods yield fast and consistent algorithms to learn linear Gaussian polytrees from interventional data. The skeleton is learned by aggregating pairwise correlations from different experimental settings; edge orientations are found by testing pairwise regression coefficients on suitable subsets of the data. This removes the need to form a score that contemplates joint models for all data. Moreover, we allow the intervention targets to be arbitrary subsets (with no variable always intervened upon). We are not aware of any other work with these features. While it will be interesting to seek extensions to broader classes of graphs in future work, we stress that for very high-dimensional problems (e.g., [1]) it is of interest to target simpler computationally tractable objects that may be inferred reliably with moderate sample size [2]. For this reason, polytrees have received renewed interest.

II. Related Work

Directed acyclic graphs (DAGs) have been extensively used in causal modeling; the nodes of a graph represent the random variables of the model while the directed edges represent causal effects from source to sink. The effects of the parent nodes on the children are quantified by structural equations. Causal discovery is then the problem of inferring the graphical structure underlying the model.

Approaches for causal discovery using only observational data and under the assumption that all variables in the model are directly observed/measured, include constraint-based, score-based and hybrid methods (e.g., PC-algorithm [3], Greedy Equivalent Search (GES) [4], Greedy SP algorithm [5]). Without extra assumptions on the data generating process, these methods learn a completed partially directed graph (CPDAG), a mixed graph that encodes the causal information common to all the members of a Markov equivalence class (MEC). Classical constrained-based algorithms, such as the PC algorithm, can suffer from the elevated number of conditional independence tests that are needed to learn the CPDAG. A recent line of work [6], [7], [8], which has proven to be almost optimal in terms of the number of conditional independence tests performed and is also applicable in the presence of unobserved variables, exploits the idea of learning the graph recursively starting from Markov boundary information.

Manuscript received 13 May 2023; revised 21 September 2023; accepted 24 October 2023. Date of publication 30 October 2023; date of current version 15 November 2023. This work was supported by the European Research Council (ERC) through the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement 883818. The work of Daniele Tramontano was supported by the IGSSE/TUM-GS via a Technical University of Munich–Imperial College London Joint Academy of Doctoral Studies. The work of Eliana Duarte was supported in part by FCT under Grant 2020.01933.CEECIND, and in part by CMUP through FCT under Grant UIDB/00144/2020. *(Corresponding author: Daniele Tramontano.)*

Daniele Tramontano, L. Waldmann, and M. Drton are with the School of Computation, Information and Technology, Department of Mathematics and Munich Data Science Institute, Technical University of Munich, 80333 Munich, Germany (e-mail: d.tramontano@tum.de; l.waldmann@tum.de; m.drton@tum.de).

Eliana Duarte is with the Faculdade de Ciências, Universidade do Porto, 4169-007 Porto, Portugal (e-mail: eliana.gelvez@fc.up.pt).

This article has supplementary downloadable material available at https://doi.org/10.1109/JSAIT.2023.3328429, provided by the authors.

Digital Object Identifier 10.1109/JSAIT.2023.3328429

Learning only the CPDAG is not always satisfactory as DAGs in the same MEC can have opposite causal interpretations. However, using additional assumptions such as non- Gaussianity [9] and/or non-linearity [10] it is possible to identify the complete causal structure. Whenever these assumptions do not apply, such as in the linear Gaussian case, additional data from interventional experiments can help to improve the identifiability of directed edges by refining the MEC. This refinement is quantified in terms of interventional Markov equivalence classes ($\mathcal{I}$-MECs) [11]. An $\mathcal{I}$-MEC is a collection of DAGs that entails the same interventional distributions for a fixed choice of intervention targets $\mathcal{I}$.

Within the causal discovery methods that use interventional data, there are those in which intervention targets are known such as Greedy Interventional Equivalent Search (GIES) [12], Interventional GSP (IGSP) [13], and Joint Causal Inference [14]. Other methods accommodate for unknown interventions targets, for instance Differentiable Causal Discovery with Interventions (DCDI) [15], permutation based approaches [16], and Bayesian Causal Discovery with unknown Interventions (BaCaDI) [17]. For a recent review of causal discovery methods we refer the reader to [18], [19].

In this paper, we address the problem of causal discovery when both observational and interventional data are available and all variables are measured. Our focus is on linear Gaussian structural causal models in which the graph is a polytree and the intervention targets are known. As shown computationally in [20], the polytree assumption provides an effective compromise between computational complexity and model expressiveness. This property of polytrees has been effectively exploited in image segmentation [21], hardware optimization [22], and Ozone prediction [23]. The polytree assumption follows a recent paradigm in the causal discovery literature in which assumptions are made about the DAG underlying the causal model in order to reduce the complexity of learning algorithms. Other methods that follow a similar approach include the causal additive trees (CAT) method which assumes the underlying DAG is a tree [24] and the method from [25] which incorporates side information such as the assumption that the ground truth is a diamond free graph or that an upper bound on the clique number of the graph is known.

The polytree assumption has been studied since the early days of causal reasoning theory. Indeed, [26] use the Chow-Liu algorithm [27] to learn the skeleton of a polytree. Different variants of the Rebane and Pearl approach that work under different sets of assumptions have been developed in [28], [29], [30], a linear programming algorithm is developed in [31], while in [32] the graph is assumed to be locally a polytree around a targeted node allowing to infer the directed causes of the target node. Both [33], in the context of time series graphs, and [34] in the context of classical graphical models, introduce a notion of minimality for polytrees with hidden nodes and provide an algorithm for learning the graph under the assumption of minimality. Polytree learning has been proven to be an NP-hard problem in [35], the complexity of the problem is studied in full details in [36]. In [37], a complete characterization is given for the constraints that emerge between 2nd and 3rd order moments of a random vector that follows a polytree-based linear structural equation model with non-Gaussian error terms.

III. Preliminaries

A. Notation for Graphs

A *directed graph* is a pair $G = (V, E)$, where V is the set of vertices and $E \subset \{(u, v) : u, v \in V,\ u \neq v\}$ is the set of directed edges. We denote a pair $(u, v) \in E$ also by $u \to v$. A *walk* from node v to node w in G is an alternating sequence $(v_0, e_1, v_1, e_2, \ldots, v_{k-1}, e_k, v_k)$ consisting of nodes and edges of G such that $v_0 = v$, $v_k = w$, and $e_l \in \{(v_{l-1}, v_l), (v_l, v_{l-1})\}$ for $l = 1, \ldots, k$. A walk is a *directed path* if $e_i = (v_{i-1}, v_i)$ for all $i \in \{1, \ldots, k\}$ and a *directed cycle* if additionally $v_0 = v_k$. From now on we assume that the graph G is a *DAG* (directed acyclic graph), i.e., G does not contain any directed cycles.

A node v_l is a *collider* on a walk as above if $e_{l-1} = (v_{l-1}, v_l)$ and $e_l = (v_{l+1}, v_l)$. Moreover, v_l is an *unshielded collider* if neither (v_{l-1}, v_{l+1}) or (v_{l+1}, v_{l-1}) belongs to E.

A walk that does not contain a collider is called a *trek* from v to w. Every trek contains a unique node v_l that splits the trek into two directed paths from v_l to v and from v_l to w, respectively. This node is the *top* of the trek. Note that the top may be equal to v or w, in which case one of the two directed paths is trivial consisting of a single node and no edge. A trek is *simple* if it does not contain repeated nodes.

If $u \to v \in E$, then u is a *parent* of v, and v is a *child* of u. If G contains a directed path from u to v, then u is an *ancestor* of v and v is a *descendant* of u. The set of parents, children, ancestors, and descendants of u are denoted by $\mathrm{pa}(u)$, $\mathrm{ch}(u)$, $\mathrm{an}(u)$, $\mathrm{de}(u)$, respectively. The *skeleton* of a DAG is the undirected graph obtained by replacing each edge (u, v), by an undirected edge, denoted here by $\{u, v\}$.

A *mixed graph* is a triple $G = (V, E, U)$, where E is the set of directed edges defined as above, and $U \subset E \subset \{\{u, v\} : u, v \in V,\ u \neq v\}$ is the set of undirected edges. We assume all the graphs we consider to be *simple*, i.e., there is at most one edge, directed or undirected, between any two vertices.

B. Linear Structural Causal Models

Let $X = (X_u)_{u \in V}$ be a random vector indexed by the vertices of a DAG G. For $A \subset V$, let $X_A = (X_u)_{u \in A}$. When X_A is conditionally independent of X_B given X_C for disjoint subsets $A, B, C \subset V$, 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 [38, Ch. 1] for further details.

The Gaussian structural causal model given by G postulates that

$$X_v = \sum_{w \in \mathrm{pa}(v)} \lambda_{wv} X_w + \varepsilon_v, \qquad v \in V, \tag{1}$$

where the edge coefficients $\lambda_{wv} \in \mathbb{R}$ are unknown parameters and the errors $(\varepsilon_v)_{v \in V}$ are independent Gaussian random variables. Each error is assumed to have mean zero and unknown

variance $\omega_v > 0$; in symbols, $\varepsilon_v \sim \mathcal{N}(0, \omega_v)$. Let $\Lambda \in \mathbb{R}^{V\times V}$ be the matrix of edge coefficients, with zeros filled in at non-edges. Let $\Omega = \operatorname{diag}((\omega_v)_{v\in V}) \in \mathbb{R}^{V\times V}$ be the diagonal covariance matrix of $\varepsilon = (\varepsilon_v)_{v\in V}$. Solving (1), we obtain that $X = (\mathrm{I}-\Lambda^T)^{-1}\varepsilon$ is a Gaussian random vector with covariance matrix

$$\Sigma := \operatorname{Var}[X] \;=\; (\mathrm{I}-\Lambda^T)^{-1}\Omega(\mathrm{I}-\Lambda)^{-1}, \tag{2}$$

we stress that the matrix $(\mathrm{I}-\Lambda)$ is always invertible when the graph G is acylic, this is because the matrix Λ is strictly lower triangular, so the determinant of $(\mathrm{I}-\Lambda)$ is one.

As presented, X is modeled to have mean zero. This is without of loss of generality for the later results which solely pertain to the covariance structure. In the sequel, we denote the space of matrices supported on the edge set of G as

$$\mathbb{R}^E = \left\{\Lambda \in \mathbb{R}^{V\times V} : (v,w) \notin E \implies \Lambda_{vw} = 0\right\}.$$

We write D_+ for the set of diagonal matrices in $\mathbb{R}^{V\times V}$ with positive diagonal entries. The covariance model induced by the DAG G is the set of positive definite matrices

$$\mathcal{M}(G) \;:=\; \left\{(\mathrm{I}-\Lambda^T)^{-1}\Omega(\mathrm{I}-\Lambda)^{-1} : \Lambda \in \mathbb{R}^E,\ \Omega \in D_+\right\}.$$

Two graphs G_1 and G_2 are *Markov equivalent* if and only if $\mathcal{M}(G_1) = \mathcal{M}(G_2)$ (see, e.g., [39, Th. 8.13]). Combinatorially, the MEC is represented by its CPDAG [38, Ch. 1].

Writing explicitly the entries of the a covariance matrix $\Sigma \in \mathcal{M}(G)$ using Eq. (2), one can get two useful parametrizations for the set $\mathcal{M}(G)$, we report the two parametrizations here, and refer to [40] and the references therein for further details.

Proposition 1 (Trek-Rule): Let $\mathcal{T}(v,w)$ be the set of all treks from v to w. The matrix Σ from Eq. (2) has its entries

$$\Sigma_{vw} = \sum_{\tau\in\mathcal{T}(v,w)} \omega_{\operatorname{top}(\tau)} \prod_{e\in\tau} \lambda_e, \quad v,w \in V.$$

Moreover, the entries of Σ satisfy the recursive relation

$$\Sigma_{vw} = \sum_{\tau\in\mathcal{S}(v,w)} \Sigma_{\operatorname{top}(\tau),\operatorname{top}(\tau)} \prod_{e\in\tau} \lambda_e, \quad v,w \in V,$$

where $\mathcal{S}(v,w)$ is the set of simple treks from v to w.

Let $\Sigma = (\sigma_{vw})$ be the covariance matrix of a random vector X, with diagonal entries $\sigma_{vv} > 0$. The correlation matrix $R(\Sigma) = (\rho_{vw})$ of Σ is the matrix with entries $\rho_{vw} = \sigma_{vw}/\sqrt{\sigma_{vv}\sigma_{ww}}$. It is also the covariance matrix of the standardized random vector $(X_v/\sqrt{\sigma_{vv}})_{v\in V}$.

Proposition 2 (Correlation Matrices): If Σ is a covariance matrix in $\mathcal{M}(G)$, then its correlation matrix $R(\Sigma)$ is also in $\mathcal{M}(G)$. Hence, there exists $\Lambda = (\lambda_{vw}) \in \mathbb{R}^E$ such that

$$R(\Sigma)_{vw} = \sum_{\tau\in\mathcal{S}(v,w)} \prod_{e\in\tau} \lambda_e, \quad v,w \in V.$$

Proof: Write $\Sigma = (\mathrm{I}-\Lambda')^{-T}\Omega'(\mathrm{I}-\Lambda')^{-1}$ for $\Lambda' \in \mathbb{R}^E$ and $\Omega' \in D_+$. Then $R(\Sigma) = (\mathrm{I}-\Lambda^T)^{-1}\Omega(\mathrm{I}-\Lambda)^{-1}$, where the entries of $\Lambda \in \mathbb{R}^E$ are $\Lambda_{vw} = \Lambda'_{vw}\sqrt{\Sigma_{vv}}/\sqrt{\Sigma_{ww}}$ and those of $\Omega \in D_+$ are $\Omega_{vv} = \Omega'_{vv}/\Sigma_{vv}$.

The second assertion follows from the simple trek-rule in Proposition 1, observing that $R(\Sigma)_{vv} = 1$ for all v. ∎

C. Interventions

It is useful to formally consider the collection of linear Gaussian structural causal models that arise from G and a set of interventions. In a *soft intervention*, a subset $I \subset V$ of target nodes is selected, and for each $v \in I$ the conditional distribution of X_v given $X_{\operatorname{pa}(v)}$ is modified. When X_v is made independent of its parents, the intervention is *perfect*. It is common that different interventions are performed and hence we have a collection of intervention targets denoted by $\mathcal{I}$, so $\mathcal{I} \subseteq 2^V$. Without loss of generality we assume that $\emptyset \in \mathcal{I}$ and refer to this as the observational experiment. [11, Th. 3.14] justifies this assumption, which subsumes the case of conservative intervention targets treated by [12].

We assume that each interventional experiment obeys a linear Gaussian structural causal model defined by G (recall Section III-B). Namely, for each $I \in \mathcal{I}$, we have a random vector $X^I := (X_v^{(I)})_{v\in V}$ with structural equations

$$X_v^{(I)} = \sum_{w\in\operatorname{pa}(v)} \lambda_{wv}^{(I)} X_w^{(I)} + \varepsilon_v^{(I)}, \qquad v \in V. \tag{3}$$

We use $\Lambda^{(I)}$ to denote the matrix of edge coefficients for this model and $\Omega^{(I)} = \operatorname{diag}((\omega_v^{(I)})_{v\in V})$ to denote the covariance matrix of the Gaussian vector $\varepsilon^{(I)} := (\varepsilon_v^{(I)})_{v\in V}$. To encode the invariances of the structural equations of nodes that are not intervened on, i.e., are not in I, we impose that $\lambda_{wv}^{(I)} = \lambda_{wv}^{(\emptyset)}$ and $\omega_v^{(I)} = \omega_v^{(\emptyset)}$ whenever $v \notin I$. Each I induces a covariance model

$$\mathcal{M}(G,I) = \{\Sigma^{(I)} \colon \Lambda^{(I)} \in \mathbb{R}^E, \Omega^{(I)} \in D_+\}$$

where $\Sigma^{(I)} = (\mathrm{I}-(\Lambda^{(I)})^T)^{-1}\Omega^{(I)}(\mathrm{I}-\Lambda^{(I)})^{-1}$ and $\Lambda^{(I)}, \Omega^{(I)}$ satisfy the invariances, on edge coefficients and error variances, of nodes that are not intervened on.

The interventional DAG model specified by G and the set of intervention targets $\mathcal{I}$ is the set

$$\mathcal{M}_{\mathcal{I}}(G) := \left\{(\Sigma^{(I)})_{I\in\mathcal{I}} : \Sigma^{(I)} \in \mathcal{M}(G,I), I \in \mathcal{I}\right\};$$

it consists of tuples of covariance matrices of length $|\mathcal{I}|$ that may arise by performing interventions according to $\mathcal{I}$.

Two DAGs G_1, G_2 are in the same *$\mathcal{I}$-Markov equivalence class*, $\mathcal{I}$-MEC, if and only if $\mathcal{M}_{\mathcal{I}}(G_1) = \mathcal{M}_{\mathcal{I}}(G_2)$. To decide if two DAGs are in the same $\mathcal{I}$-MEC, Yang et al. [11] introduce the notion of an $\mathcal{I}$-DAG:

Definition 1 ($\mathcal{I}$-DAG): Fix a collection of interventions $\mathcal{I}$ and a DAG G, the $\mathcal{I}$-DAG $G^{\mathcal{I}}$ is the graph G augmented with $\mathcal{I}$-vertices $\{\zeta_I\}_{\emptyset\neq I\in\mathcal{I}}$, and the $\mathcal{I}$-edges $\{\zeta_I \to u\}_{u\in I, I\in\mathcal{I}}$.

The following theorem provides a concise graphical representation of the *$\mathcal{I}$-Markov equivalence classes*.

Theorem 1 [11, Th. 3.14]: Let $\mathcal{I}$ be a conservative set of intervention targets. Two DAGs G_1, G_2 are in the same $\mathcal{I}$-MEC iff for all $I \in \mathcal{I}$ the $\mathcal{I}$-DAGs $G_1^{\tilde{\mathcal{I}}_I}$ and $G_2^{\tilde{\mathcal{I}}_I}$ have the same skeleton and the same v-structures, where

$$\tilde{\mathcal{I}}_I = \{\emptyset\} \cup \{I \cup J\}_{J\in\mathcal{I}, I\neq J}.$$

The $\mathcal{I}$-MEC of a DAG G can be represented uniquely by its $\mathcal{I}$-CPDAG, this provides a combinatorial representation of the causal information that we can extract from the interventional

data available, often also referred to as the $\mathcal{I}$-essential graph in the literature. This is a mixed graph with the same skeleton as G, a directed edge, (u, v), if every member of the $\mathcal{I}$-MEC has that edge with the same orientation, and an undirected edge, $\{u, v\}$, if there are two DAGs in the $\mathcal{I}$-MEC for which that edge has opposite orientations. See Hauser and Bühlmann [12, Th. 18] for details in the setting of perfect interventions, the same construction carries over to the setting of general interventions [11].

IV. Learning Causal Models on Polytrees With Interventions

From now on we assume that the DAG G is a polytree, this means that the skeleton of G is a tree, i.e., a graph in which there is exactly one path between any two nodes.

Our procedure first learns the skeleton. For this purpose we create a novel interventional version of the Chow-Liu algorithm [27]. The challenge here lies in constructing a weight matrix that reveals the tree structure, the same way the correlation matrix on a single observational data set does. As seen in Appendix C, Fig. 2, taking simply the correlation matrix of the pooled data does not work. Thus, we introduce the notion of a G-valid weight matrix that suitably captures the tree structure. We will construct G-valid weight matrices by aggregating correlation matrices. The aggregation does not necessarily produce a correlation matrix, yet we show that this approach yields a consistent procedure to learn the skeleton in low- and high-dimensional settings. The approach is detailed in Section IV-A. Its consistency is discussed in Appendix A-A.

To determine the orientation of the edges we propose and compare several methods described in Sections IV-C and IV-D; their consistency is discussed in the Appendix A-B. Importantly, all procedures use only correlations and low-dimensional marginal distributions that are efficiently estimable in a low sample regime, see e.g., the concentration inequality given in [41, Corollary 1] in which it is explicit that the effective sample size decreases as the size of conditioning set increases requiring more samples to achieve the same accuracy in the estimation.

A. Learning a Skeleton

In seminal work, Rebane and Pearl [26] proved that when a distribution satisfies conditional independence constraints induced by a polytree, then the skeleton of the tree can be recovered using the algorithm for the maximum weight spanning tree of Kruskal [42], with weight matrix given by the mutual information between the variables. Notably, the only property of mutual information needed in this algorithm is the data processing inequality [43, Th. 2.8.1]. This implies that Kruskal's algorithm finds the correct polytree skeleton any time the weight matrix that is used respects the following condition.

Definition 2: Given a polytree G, a weight matrix $W \in \mathbb{R}^{p\times p}$ is G-valid if for every triplet $u - v - w$ in G

$$\min\{W(u, v), W(v, w)\} \geq W(u, w). \tag{4}$$

If all inequalities in (4) are strict, then W is strictly G-valid.

Lemma 1 [26, Th. 1]: If W is *strictly* G-valid, then the maximum weight spanning tree of the complete graph over V with weight matrix W is the skeleton of G.

In a polytree G there is at most one trek between any two vertices. In particular for any triple $u - v - w$, the only possible trek between u and w is the one involving also v, so from Proposition 2 we have that $|\rho_{u,w}| = 0$ if $u - v - w$ is a collider triple and $|\rho_{u,w}| = |\rho_{u,v}||\rho_{v,w}|$ otherwise. In particular, the absolute (observational) correlations define a G-valid weight matrix that is strictly G-valid if G is causally minimal. Similarly, for each $I \in \mathcal{I}$, the associated absolute correlation matrix $|R^I|$ is a G-valid weight matrix.

To efficiently learn from available interventional data, we wish to form a single weight matrix that encodes the information of all the experimental settings. To this end, we use an aggregation function $A : (\mathbb{R}^{p\times p})^{|\mathcal{I}|} \to \mathbb{R}^{p\times p}$ that takes a collection of G-valid weight matrices and outputs a strictly G-valid matrix. One way of obtaining such a function A is to apply the same order-preserving transformation $a : \mathbb{R}^{|\mathcal{I}|} \to \mathbb{R}$ to each vector of correlations $(\rho_{ij}^1, \ldots, \rho_{ij}^{|\mathcal{I}|})$. By order-preserving, we mean that $a(x_1, \ldots, x_{|\mathcal{I}|}) \leq a(y_1, \ldots, y_{|\mathcal{I}|})$ anytime $x_i \leq y_i$ for $i = 1, \ldots, |\mathcal{I}|$. Possible choices for a are:

1) $a(\rho_{ij}^1, \ldots, \rho_{ij}^{|\mathcal{I}|})) = -\sum_{I\in\mathcal{I}} \frac{n_I}{2} \log(1 - (\rho_{ij}^I)^2)$,[1]
2) Weighted mean,
3) Weighted median,

where n_I is the size of the dataset I and the weights we refer to for the mean and the median are $\frac{n_I}{\sum n_I}$.

The output of Kruskal's algorithm is the same for any strictly G-valid matrix W. In practice, however, we work with an estimate $\tilde{W}$ and different aggregation functions can give different results. We discuss the numerical performance of the three aggregation functions in Section V-A.

B. Identifiability of Edge Orientations

The output of our learning methods is an $\mathcal{I}$-CPDAG (Section III-C) whose construction simplifies for polytrees. The next definition singles out the edge directions that can be identified in a polytree. Notice that in a polytree there is only one path between two vertices, thus all the colliders are unshielded.

Definition 3: An edge $u \to v \in G$ is $\mathcal{I}$-directly-identifiable if it is either part of a collider, or there exists $I \in \mathcal{I}$ such that $|I \cap \{u, v\}| = 1$. It is $\mathcal{I}$-identifiable if it is either $\mathcal{I}$-directly-identifiable, or there is an edge $w \to u \in E$ that is $\mathcal{I}$-identifiable.

Proposition 3: The $\mathcal{I}$-CPDAG of G is the partially directed graph that has the same skeleton of G, and whose directed edges are the edges of G that are $\mathcal{I}$-identifiable.

It is clear from Definition 3 that once we have computed the skeleton of the polytree, the $\mathcal{I}$-CPDAG can be identified by searching the $\mathcal{I}$-directly-identifiable edges first. Then, once we have a triple of the form $u \to v - w$ we can orient $v - w$ testing if the triple forms a collider. In the next sections we describe different orientiation strategies: Section IV-C focuses

[1]This is well defined for absolute correlations in $[0, 1)$.

on single edge orientations, and Section IV-D focuses on finding colliders. Searching for one type of $\mathcal{I}$-directly-identifiable edge or the other first does not matter on a population level. However, it can make a difference when working with data due to approximation errors inherent to each possible approach. A comparison of performance between these approaches is given in Appendix C-3.

Hereafter, X_u^I, X_v^I denote vectors of observed values of the variables X_u, X_v in the interventional setting $I \in \mathcal{I}$. Entrywise, $X_u^I = (X_{u,1}^I, \ldots, X_{u,n_I}^I)$ and $X_v^I = (X_{v,1}^I, \ldots, X_{v,n_I}^I)$. The total sample size is then $n := \sum_{I\in\mathcal{I}} n_I$. Fix $\mathcal{I}$ and an edge $\{u, v\}$, we define

$$\begin{aligned}
\mathcal{I}_{\dot{v}} &:= \{I \in \mathcal{I} : v \notin I\},\\
\mathcal{I}_{\dot{u},v} &:= \{I \in \mathcal{I} : v \in I, u \notin I\},\\
\mathcal{I}_{\dot{u},\dot{v}} &:= \{I \in \mathcal{I} : u, v \notin I\},\\
\mathcal{I}_{u,\dot{v}} &:= \{I \in \mathcal{I} : u \in I, v \notin I\}.
\end{aligned}$$

C. Learning Single Edge Orientations

1) Invariance of Regression Coefficients (IRC): To orient the edge $\{u, v\}$, we assume there exists $I_0 \in \mathcal{I}$ where no intervention took place, i.e., $I_0 = \emptyset$. Fix $I \in \mathcal{I}_{u,\dot{v}}$. If the true model is $u \to v$, then an intervention on u does not change the regression coefficient of $X_v^I = \lambda_{uv}^I X_u^I + \epsilon^I$, $\epsilon^I \sim \mathcal{N}(0, \sigma_{v|u})$. Namely, $\lambda_{uv}^I = \lambda_{uv}^{\emptyset}$. Thus to orient the edge $\{u, v\}$ we test the hypothesis $\mathcal{H}$:$\lambda_{uv}^I = \lambda_{uv}^{\emptyset}$. There are different options for the choice of this test and IRC testing has been used before in causal structure learning [e.g., [44], [45]]. Here we use an F-test [46], which is a modified version of the test statistic in [47] for the case of unequal variances. We briefly present it here.

Let e_{I_0} and e_I denote the vectors of residuals of the regressions $X_v^I = \lambda_{uv}^{I_0} X_u^{I_0} + \epsilon^{I_0}$ and $X_v^I = \lambda_{uv}^I X_u^I + \epsilon^I$ respectively, and let e denote the vector of residuals, which is obtained by pooling the data sets $(X_u^{I_0}, X_v^{I_0})$, (X_u^I, X_v^I). Under the null hypothesis $\mathcal{H}_{I,u\to v}$ the statistic is

$$\frac{(e^T e - e_{I_0}^T e_{I_0} - e_I^T e_I)/k}{(e_{I_0}^T e_{I_0} - e_I^T e_I)/f} \sim F(k, f)$$

where $k = 2$ and $f = \frac{((n_{I_0}-k)\hat{\sigma}_{I_0}^2 + (n_I - k)\hat{\sigma}_{I_0})^2}{(n_{I_0}-k)\hat{\sigma}_{n_0}^4 + (n_I-k)\hat{\sigma}_{n_I}^4}$.

For each $I \in \mathcal{I}_{u,\dot{v}}$ we test the hypothesis $\mathcal{H}_{I,u\to v}$ and reject the orientation $u \to v$ in favor of $v \to u$ if the p-value of any of these tests is below $\alpha/|\mathcal{I}_{u,\dot{v}}|$ for some chosen significance level α. This test is possible whenever $\mathcal{I}_{u,\dot{v}} \neq \emptyset$.

In an analogous fashion, changing the roles of u, v, if $\mathcal{I}_{\dot{u},v}$ is nonempty, we may perform a test for the orientation $v \to u$. If only one of the tests $u \to v$ or $v \to u$ is possible, then the decision rule to orient the edge $\{u, v\}$ is clear. In case both tests are possible, we choose to reject the test with the smallest p-value, as long as such p-value is below the corresponding significance ($\alpha/|\mathcal{I}_{u,\dot{v}}|$ or $\alpha/|\mathcal{I}_{\dot{u},v}|$).

2) Single Edge BIC Score Before Collider Search: Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be any two parametric models with parameter spaces $\mathcal{M}_1 \subset \mathbb{R}^{r_1}$ and $\mathcal{M}_2 \subset \mathbb{R}^{r_2}$ and of dimensions d_1 and d_2, respectively. A model selection based on the BIC consists of choosing the minimizer of the following penalized log-likelihood

$$l(\mathcal{H}_i \mid X_n) = \arg\min_{\theta_i\in\mathcal{M}_i}\big(-\log(p_{\mathcal{H}_i}(X_n \mid \theta_i)) + \frac{d_i}{2}\log(n)\big), \quad (5)$$

for $i \in \{1, 2\}$, where X_n is the observed dataset of size n. For more details on model selection and information criteria, we refer the reader to [48, Ch. 2].

In our setting we want to select between the model in which $u \to v \in G$ and the one in which $v \to u \in G$, respectively denoted as $\mathcal{H}_{u\to v}$ and $\mathcal{H}_{v\to u}$. The polytree assumption allows us to compute the log likelihood using only information on the marginal distribution of u and v. We show in detail only the derivation of the penalized log likehood for $\mathcal{H}_{u\to v}$, since the one for the other model is analogous.

Under the model $\mathcal{H}_{u\to v}$ we can write in each dataset $X_v = \lambda_{u,v}^I X_u + \epsilon_{v|u}^I$ where, using Eq. (3), we can see that

$$\epsilon_{v|u}^I = \sum_{w\in \mathrm{pa}(v)\setminus\{u\}} \lambda_{wv}^{(I)} X_w^{(I)} + \varepsilon_v^{(I)} \sim \mathcal{N}(0, \sigma_{v|u}^I), \quad (6)$$

$$\sigma_{v|u}^I = \sum_{w\in \mathrm{pa}(v)\setminus\{u\}} (\lambda_{wv}^2)^{(I)} \sigma_w^{(I)} + \sigma_v^{(I)}, \quad (7)$$

a consequence of the polytree is assumption is that X_u is independent from all the other parents of v, and so it is independent from $\epsilon_{v|u}$ as well. This allows us to rewrite the log-likelihood of a observed pair $(X_{u,i}^I, X_{v,i}^I)$ as

$$\begin{aligned}
&-\log 2\pi - \frac{1}{2}\log {\sigma_u^I}^2 - \frac{1}{2}\log {\sigma_{v|u}^I}^2\\
&-\frac{(X_{u,i}^I)^2}{2{\sigma_u^I}^2} - \frac{(X_{v,i}^I - \lambda_{u,v}^I X_{u,i}^I)^2}{2{\sigma_{v|u}^I}^2}.
\end{aligned}$$

The log-likelihood for the whole dataset is then

$$\begin{aligned}
&-n\log 2\pi + \sum_{I\in\mathcal{I}}\Bigg(-\frac{n_I}{2}\log {\sigma_u^I}^2 - \frac{n_I}{2}\log {\sigma_{v|u}^I}^2\\
&+\sum_{i\in[1,\ldots,n_I]} -\frac{(X_{u,i}^I)^2}{2{\sigma_u^I}^2} - \frac{(X_{v,i}^I - \lambda_{u,v}^I X_{u,i}^I)^2}{2{\sigma_{v|u}^I}^2}\Bigg). \quad (8)
\end{aligned}$$

Straightforward computations show that the MLEs for the variance parameters are the usual variances computed on each dataset individually leading to

$$\hat{\sigma}_u^{I^2} = \frac{1}{n_I}||X_u^I||^2, \hat{\sigma}_{v|u}^{I^2} = \frac{1}{n_I}||X_v^I - \lambda_{u,v}^I X_u^I||^2. \quad (9)$$

For the regression coefficients, since we know that they vary only across the datasets in which v has been intervened upon, it is not possible to find a closed form expression for the estimator. To handle this, we tell apart the datasets in which v has been intervened on from those in which it has not. Define $\lambda_{u,v}$ as the regression coefficient that is shared across the environments in which v has not been intervened on. Substituting (9) into (8), the part of the log-likelihood that depends on the regression coefficients becomes

$$-\left(\sum_{I\in\mathcal{I}_{\dot{v}}}\frac{n_I}{2}\log||X_v^I - \lambda X_u^I||^2 + \sum_{I\in\mathcal{I}_v}\frac{n_I}{2}\log||X_v^I - \lambda^I X_u^I||^2\right).$$

So if $I \in \mathcal{I}_v$ then $\hat{\lambda}^I = (X_u^I \cdot X_v^I)/||X_u^I||^2$, while $\hat{\lambda}$ is given by the solution of the 1-dimensional bounded optimization problem

$$\arg\max_{\lambda \in [-1,1]} - \sum_{I \in \mathcal{I}_v} \frac{n_I}{2} \log \left\| X_v^I - \lambda X_u^I \right\|^2. \tag{10}$$

For solving this problem, we find Brent's method, see, e.g., [49, Ch. 10.3], to work well in practice. Finally, note that the dimension of the model, that is, the number of free parameters we need to optimize over, is $2|\mathcal{I}| + 1 + |\mathcal{I}_v|$. Indeed, we have two variance coefficients for each interventional dataset, one regression coefficient for the datasets in which v has not been intervened upon, and one regression coefficient for each dataset in which v has been intervened on.

Remark 1: In this section we allow the marginal variances σ_u and $\sigma_{v|u}$ to vary also in the datasets in which u and v are not intervened upon. This is because the variance can change also as a consequence of an intervention on an ancestor. A precise likelihood computation would require to check for each experiment if there are possible ancestors in the graph that can affect the marginal variances, but this would be too computationally expensive while giving little benefit in terms of likelihood comparison. We discuss this aspect further in the Appendix, Prop. A.4, where we also clarify that consistency is not affected by the relaxation of the variance constraints.

3) Single Edge BIC Score After Collider Search: Let $G^{CP} = (V, E^{CP})$ be the CPDAG associated to G and G^U be graph obtained from G^{CP} after removing all the oriented edges. We can then work on each connected component of G^U independently. Let U be one of these components, and $u - v$ be one of its $\mathcal{I}$-directly-identifiable edges. If we orient the edge from u to v then we would orient the remaining edges in U in such a way that no other colliders appear. We now propose a simple likelihood computation that takes this into account.

The choice of an orientation for the edges in U reduces to the choice of a root vertex in it. If u is a vertex in U, then we let $\mathcal{H}_u^U$ denote the model in which u is the root. The likelihood function for this model factorizes in a simple way since each vertex has at most one parent. We can write the log-likelihood of a datapoint as

$$\log f(X_U^I) = \log f(X_u^I) + \sum_{v \neq u} \log f\left(X_v^I | X_{pa(v)}^I\right),$$

where each of the summands can be maximized independently from the others. This time the maximization is easier than the one in Section IV-C2 because the variance parameters are shared across all environments in which a node has not been intervened upon. A closed form solution for this problem is provided in [50] and is given by $\hat{\sigma}^2_{v|\,\mathrm{pa}(v)} = \sum_{I_{\hat{v}}} \frac{n_I}{n_{\hat{v}}} \hat{\sigma}^{I^2}_{v|\,\mathrm{pa}(v)}$, where $n_{\hat{v}} = \sum_{I_{\hat{v}}} n_I$ and $\hat{\sigma}^{I^2}_{v|\,\mathrm{pa}(v)}$ is the MLE of the conditional variance in the dataset I, while $\hat{\lambda} = \frac{\hat{\sigma}_{v,\mathrm{pa}(v)}}{\hat{\sigma}^2_{\mathrm{pa}(v)}}$. The dimension of the model $\mathcal{H}_u$ is $1 + |\mathcal{I}_u| + 2\sum_{v \neq u}(1 + |\mathcal{I}_v|)$.

D. Collider Search

1) BIC Score for Collider Search: If the skeleton has a triplet $u - v - w$, we can decide whether it forms a collider or not by testing for the independence of u and w in all the environments. Log-likelihood ratio statistics for a test are provided by function 1 defined in Section (IV-A), applied to $x_I = \rho^I_{u,w}$. Here the difference in the dimension between the model $\mathcal{H}_{u \to v \leftarrow w}$ and the other 3 models in which u and w are not independent is $|\mathcal{I}|$.

2) BIC Score for Collider Completion: Even though testing for the independence of u and w as in Section IV-D1 would give a consistent procedure also in the case in which we have already oriented $u \to v$, in this case a more refined analysis of the likelihood is possible. Indeed, here we can test the model $u \to v \to w$ against $u \to v \leftarrow w$. In the former case the likelihood factorizes as $f(X_u^I) f(X_v^I | X_u^I) f(X_w^I | X_v^I)$ and can be maximized as in Section IV-C2. In the latter case it factorizes as $f(X_u^I) f(X_w^I) f(X_v^I | X_u^I, X_w^I)$, and the MLE for $\lambda_v^I = (\lambda^I_{v|u}, \lambda^I_{v|w})$ in the environments in which v has been intervened upon is given by the usual estimator $\hat{\Sigma}^{I^{-1}}_{u,w} \hat{\Sigma}^I_{v|u,w}$, while the common regression coefficients for the environments in which v hasn't been intervened on $\hat{\lambda}_{v|u}, \hat{\lambda}_{v|u}$, is given by the solution of the following 2 dimensional optimization problem:

$$arg\,max_{\lambda_{v|u}, \lambda_{v|u} \in [0,1]^2} \sum_{I \in \mathcal{I}_{\hat{v}}} -\frac{n_I}{2} \log ||X_v^I - \lambda_{v|u} X_u^I - \lambda_{v|w} X_w^I||^2$$

The dimension of the model $\mathcal{H}_{u \to v \to w}$ is $3|\mathcal{I}| + |\mathcal{I}_v| + |\mathcal{I}_w| + 2$ while that of the model $\mathcal{H}_{u \to v \to w}$ is $2(2|\mathcal{I}| + |\mathcal{I}_v| + 1)$.

E. Complete Orientation Procedures

We propose two procedures for the orientation. The pseudocode for each, including subroutines, is in Appendix B:

- P.1 Compute the CPDAG using the collider search, then use single edge orientation.
- P.2 Use single edge orientation first, then orient the rest of the $\mathcal{I}$-CPDAG using the collider search.

These procedures differ in the amount of statistical vs. causal information they extrapolate from data, with P.1 being the more statistically oriented, while P.2 the more causally oriented. Note that for the single edge orientation we can choose between BIC (Section IV-C2) and IRC (Section IV-C1), and to find colliders we can use the more general setting of Section IV-D1 hereon referred to as "simple", or the more refined analysis of Section IV-D2 hereon referred to as "refined".

Whenever we use the orientation procedures involving a BIC score, it is intended that we solve a local model selection problem in the following way: we compute the MLE estimator for the possible models (e.g., $\mathcal{H}_{u \to v}$ and $\mathcal{H}_{v \to u}$ for the single edge orientation) and then plug it into the likelihood function to compute the maximum likelihood $\hat{L}$, the BIC score of the model $\mathcal{H}$ will be $\log(n) \dim(\mathcal{H}) - 2\log(\hat{L})$, where n is the sample size. Finally, we select the model with the highest BIC score.

V. Simulation Studies

In this section we first assess the performance of the different skeleton and orientation recovery procedures, then we construct full versions of our algorithms to compare to other methods. The setup for our simulations is explained in Appendix C-A, the code is available at [51].

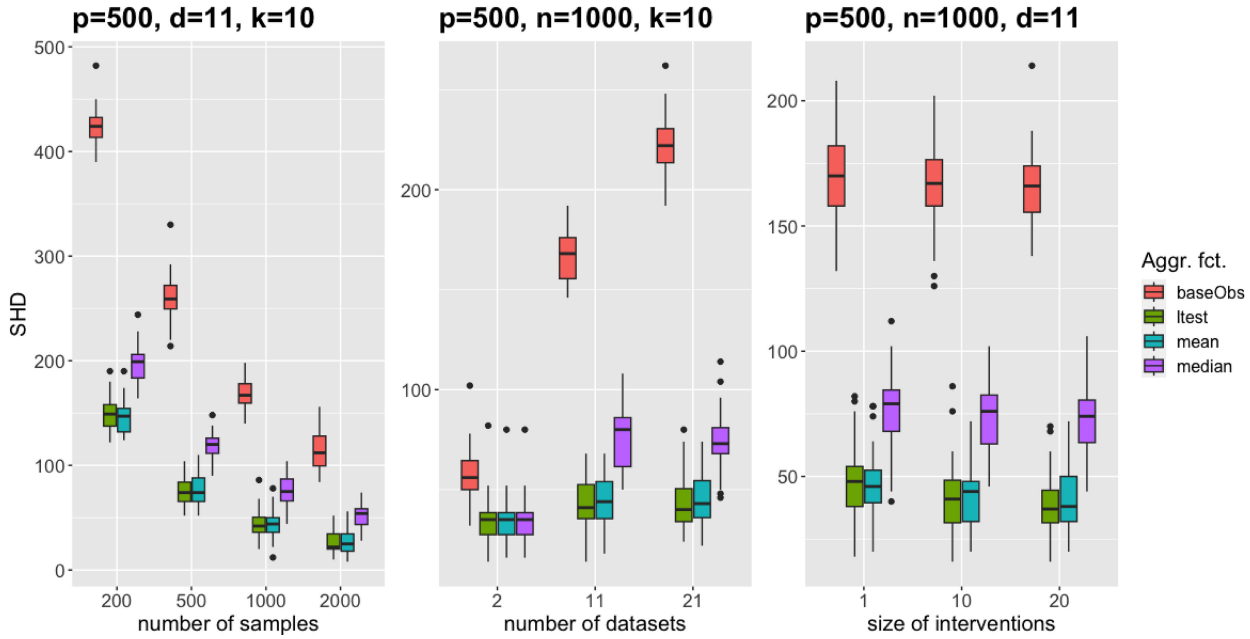

Fig. 1. Skeleton recovery with $p = 500$ nodes, $N = 200, 500, 1000, 2000$ samples, $d = 2, 11, 21$ datasets (1 interventional and $d - 1$ observational) and $k = 1, 10, 20$ nodes targeted per intervention, respectively. The aggregation function baseObs denotes a baseline with observational data only. The labels Itest, mean, median indicate each of the procedures in the list in Section IV-A respectively.

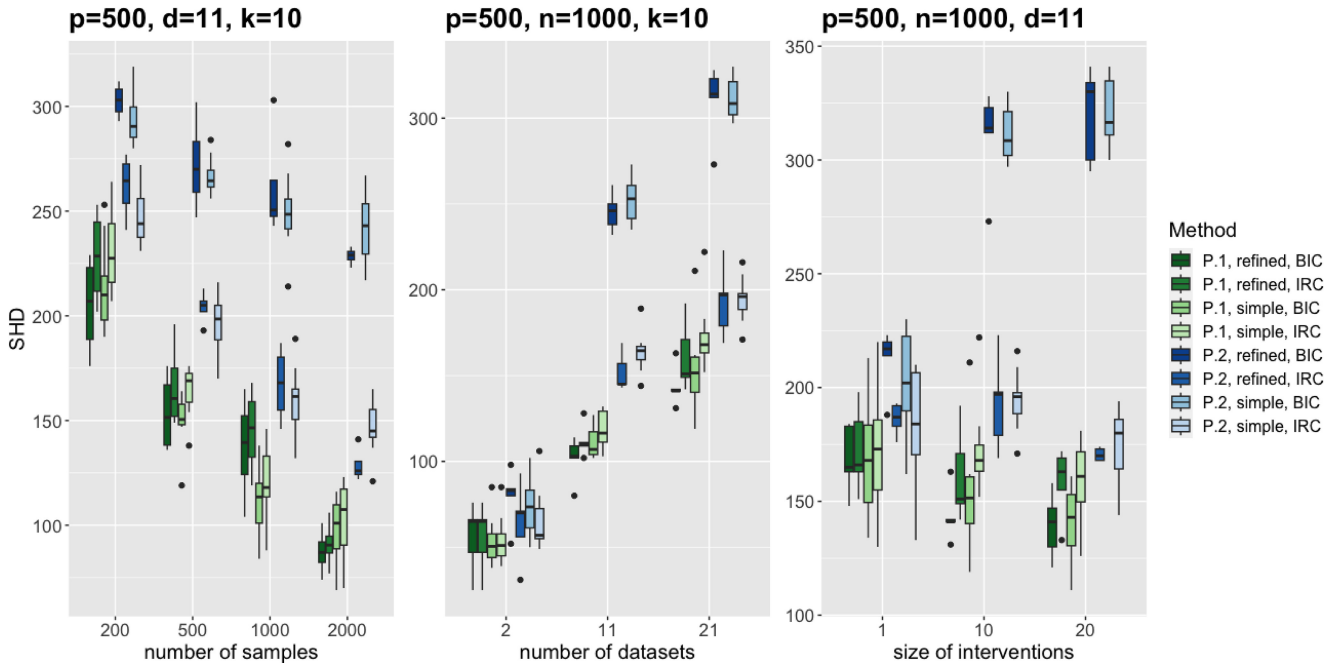

Fig. 2. Orientation recovery with base parameters 500 nodes, 1000 samples and 10 intervention targets with 10 intervened nodes each. For all proposed methods we see the convergence when the sample size increases. In this case increasing the number of datasets negatively affects the performance, this point is disccused in Rem. 2.

A. Skeleton and Orientation Recovery

Fig. 1 shows the structural Hamming distance (SHD) in skeleton recovery for the three aggregation functions introduced in Section IV-A. The SHD is the minimum number of edge additions, deletions and reversals necessary to transform one graph into another. We see that the SHD decreases for increasing number of samples, with little effect of the number of datasets and the size of the intervention. Fig. 3 in Appendix C-C. shows that the weighted mean is the fastest. A more detailed analysis is contained in Appendix C-B.

Fig. 2 depicts the SHD in orientation recovery, conditional on the skeleton being correct. Interestingly, the more refined test proposed in Section IV-D2 gives no benefits compared to the ones in Section IV-D1, and it is slower to compute. P.1 also performs better than P.2 in general. More details on the orientation procedure in Appendix C-C.

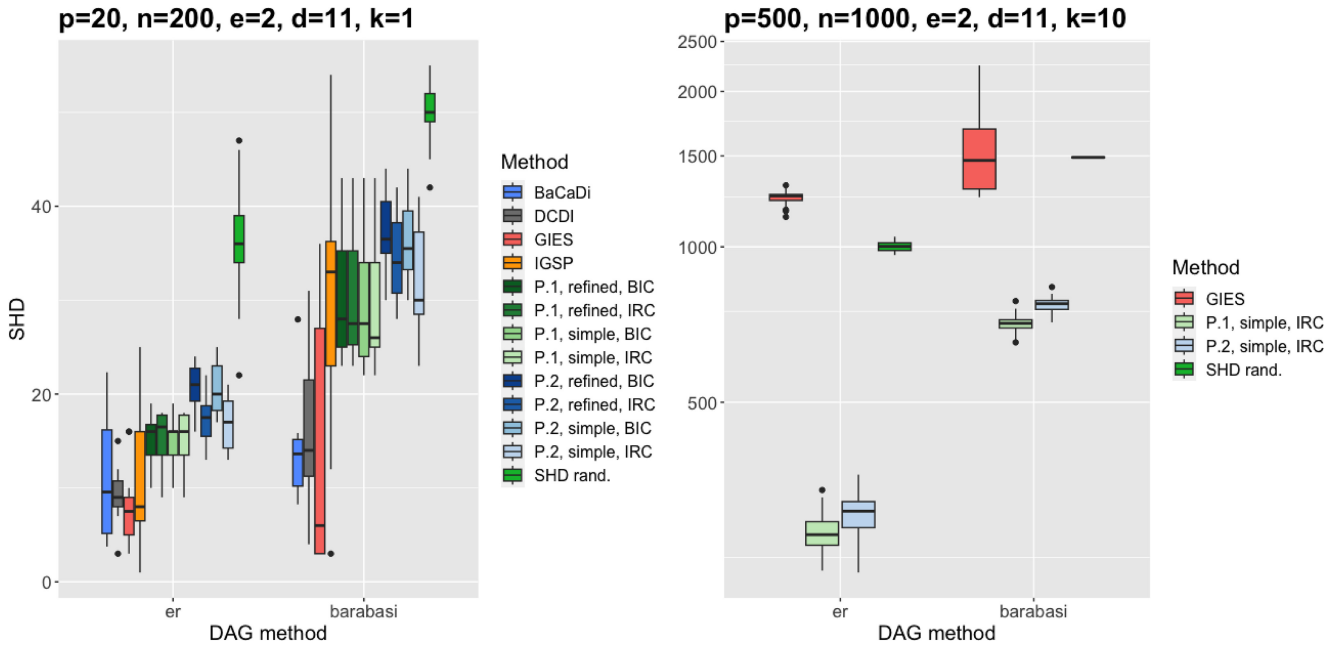

Fig. 3. Performance of our algorithm against different baselines in low- and high-dimensional settings on DAGs for random Erdős-Rényi or Barabasi-Albert graphs with $p = 20, 500$ nodes and expected number of edges per node $e = 2$. The observational sample sizes are respectively 100 and 48, the other samples are evenly distributed in the interventional datasets. Notice that the y-axis for the plot on the right is in log-scale.

Remark 2: Notice that in the experiment shown in the middle panels of Figs. 1 and 2, the overall sample size is fixed, so increasing d significantly reduces the sample size in each one of the datasets, resulting in a much more difficult learning problem. This explains why the orientation procedures seem to be negatively affected by the introduction of interventional datasets.

TABLE I
MEAN RUNTIME ON DAGS WITH 500 NODES AND DIFFERENT NUMBER OF SAMPLES n CORRESPONDING TO FIG. 3

Method	Runtime in seconds					
	$n = 500$		$n = 1000$		$n = 2000$	
	mean	max	mean	max	mean	max
P.1	8	12	8	13	8	13
P.2	8	14	7.8	13	8	15
GIES	7960	14715	449	909	284	553

B. Complete Algorithm

Although our focus is on causal discovery in high-dimensional settings, for completeness, we compare the performance of our full skeleton and orientation recovery procedures (Section IV-E) to the performance of GIES [12], DCDI [15], BaCaDI [17] and IGSP [13] in low dimension with simulations on DAGs with $p = 20$ nodes. The results are summarized in Fig. 3 (left). We see that general algorithms achieve a better accuracy, but this comes with a high price in terms of running time that makes them infeasible for larger graphs. We refer to Table II in Appendix C-E for a runtime analysis.

For high-dimensional settings, to leverage accuracy and fast computation, we construct our full algorithm using either P.1 or P.2 with mean as aggregation function, simple for collider search and IRC for single edge orientations. Among the four algorithms from the literature considered in Fig. 3 (left), the only one that would terminate in less than 24h for DAGs with $p = 500$ nodes is GIES [50]. Thus, we compare our two high-dimensional versions of the algorithm with GIES only. The Fig. 3 (right) shows the results of the simulations. We see in this setting that our algorithm provides better accuracy in terms of structural Hamming distance in addition to being considerably faster, as the computation times in Table I indicate. Appendix C-E contains an extensive comparison of our algorithm against GIES.

To emphasize the point that our algorithms are able to retain structural information of the original DAG, in both the plots in Fig. 3 we added "SHD rand." as baseline. This shows the SHD between the $\mathcal{I}$-CPDAG of the true DAG and the $\mathcal{I}$-CPDAG of randomly generated polytree. This baseline makes sure to highlight that the better performance of our algorithms in terms of SHD is not given by the sparsity structure of the polytree. The distance from the result of our algorithm to this baseline is a strong indication that even in a misspecified setting where the generating graph is not a polytree, our algorithms are still able to infer information about the original DAG.

C. Protein Expression Data

We illustrate our algorithm on the well-known protein expression dataset from [52]. Similar to [13], we processed the dataset into one observational and five one-node interventional datasets with 5846 samples of 11 nodes in total. The CPDAG of the conventionally accepted model of

TABLE II
STRUCTURAL HAMMING DISTANCE TO THE CONSENSUS $\mathcal{I}$-CPDAG OF THE PROTEIN NETWORK FROM [52]

	Method					
	P.1/P.2	GIES	IGSP	CAM	DCDI-G	DCDI-DSF
SHD	15	38	18	35	36	33

the interactions between nodes serves as ground truth; this consensus $\mathcal{I}$-CPDAG and the one estimated by our algorithm can be found in the Appendix. Table II shows the SHD to the consensus $\mathcal{I}$-CPDAG; the statistics for the other algorithms are taken from [15, Table 3]. Although the ground truth is a DAG with 18 edges, our algorithm's performance is comparable to that of more general methods.

VI. CONCLUSION

We proposed an approach to learn linear Gaussian polytree models from interventional data where the interventions have known targets. Our two-step approach of first learning the skeleton and then orienting the edges exploits the availability of interventional data at each step, all the while using only local computations with low-dimensional marginals. We emphasize that our methods are especially well suited for the high-dimensional setting, with large graphs but moderate sample size.

To conclude, we highlight topics that emerge as future directions.

Unknown interventions: A natural extension of our work is to allow for unknown intervention targets. Although the skeleton recovery and collider search procedures apply in this case, the remaining orientation subroutines do not because they operate by searching $\mathcal{I}$-identifiable edges; this, in turn, depends on the characterization of $\mathcal{I}$-Markov equivalence which requires explicit knowledge of $\mathcal{I}$. The notion of Ψ-Markov equivalence, introduced in [53], for unknown intervention targets appears to be a reasonable candidate to replace the use of the $\mathcal{I}$-MEC. The reason being that the Ψ-Markov equivalence class (Ψ-Markov EC) is a generalization of the $\mathcal{I}$-MEC in this case [53, Appendix D].

Forests: We focused on connected polytrees, but a generalization to disconnected forests of polytrees would be of interest. Undirected forests have been studied by [2], but their construction does not carry over naturally to the directed case, and new research is needed.

Hidden variables: The algorithms we propose do not address the case where some variables remain hidden. For purely observational data from a polytree model, the hidden variable setting was considered by Sepehr and Materassi [34]. These authors provide necessary and sufficient conditions for causal structure recovery of the polytree with hidden nodes. It would be interesting to investigate to what extent interventional data would improve the identifiability of the polytree structure when such conditions are not met. Similar to the unknown interventions case, a possible approach would be to replace the $\mathcal{I}$-MEC by the Ψ-Markov EC because its level of generality encompasses hidden variables also.

REFERENCES

[1] A. Dixit et al., "Perturb-seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens," *Cell*, vol. 167, no. 7, pp. 1853–1866. e17, 2016.

[2] D. Edwards, G. Abreu, and R. Labouriau, "Selecting high-dimensional mixed graphical models using minimal AIC or BIC forests," *BMC Bioinf.*, vol. 11, no. 1, p. 18, 2010.

[3] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search* (Adaptive Computation and Machine Learning), 2nd ed. Cambridge, MA, USA: MIT Press, 2000.

[4] D. M. Chickering, "Optimal structure identification with greedy search," *J. Mach. Learn. Res.*, vol. 3, pp. 507–554, Nov. 2002.

[5] L. Solus, Y. Wang, and C. Uhler, "Consistency guarantees for greedy permutation-based causal inference algorithms," *Biometrika*, vol. 108, no. 4, pp. 795–814, Jan. 2021.

[6] S. Akbari, E. Mokhtarian, A. Ghassami, and N. Kiyavash, "Recursive causal structure learning in the presence of latent variables and selection bias," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 10119–10130.

[7] E. Mokhtarian, S. Akbari, A. Ghassami, and N. Kiyavash, "A recursive Markov boundary-based approach to causal structure learning," in *Proc. KDD Workshop Causal Discov.*, 2021, pp. 26–54.

[8] E. Mokhtarian, M. Khorasani, J. Etesami, and N. Kiyavash, "Novel ordering-based approaches for causal structure learning in the presence of unobserved variables," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, Jun. 2023, pp. 12260–12268.

[9] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen, "A linear non-Gaussian acyclic model for causal discovery," *J. Mach. Learn. Res.*, vol. 7, no. 72, pp. 2003–2030, 2006.

[10] P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, "Nonlinear causal discovery with additive noise models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 21, 2008, pp. 689–696.

[11] K. D. Yang, A. Katcoff, and C. Uhler, "Characterizing and learning equivalence classes of causal DAGs under interventions," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 5541–5550.

[12] A. Hauser and P. Bühlmann, "Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 2409–2464, Aug. 2012.

[13] Y. Wang, L. Solus, K. Yang, and C. Uhler, "Permutation-based causal inference algorithms with interventions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5822–5831.

[14] J. M. Mooij, S. Magliacane, and T. Claassen, "Joint causal inference from multiple contexts," *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 1–108, Jan. 2020.

[15] P. Brouillard, S. Lachapelle, A. Lacoste, S. Lacoste-Julien, and A. Drouin, "Differentiable causal discovery from interventional data," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21865–21877.

[16] C. Squires, Y. Wang, and C. Uhler, "Permutation-based causal structure learning with unknown intervention targets," in *Proc. Conf. Uncertain. Artif. Intell.*, 2020, pp. 1039–1048.

[17] A. Hägele, J. Rothfuss, L. Lorch, V. R. Somnath, B. Schölkopf, and A. Krause, "BaCaDI: Bayesian causal discovery with unknown interventions," 2022, *arXiv:2206.01665*.

[18] C. Squires and C. Uhler, "Causal structure learning: A combinatorial perspective," *Found. Comput. Math.*, vol. 23, pp. 1781–1815, Oct. 2023.

[19] A. Zanga, E. Ozkirimli, and F. Stella, "A survey on causal discovery: Theory and practice," *Int. J. Approx. Reason.*, vol. 151, pp. 101–129, Dec. 2022.

[20] S. Acid and L. M. de Campos, "Approximations of causal networks by Polytrees: An empirical study," in *Proc. 5th Int. Conf. Process. Manag. Uncertain. Knowl.-Based Syst.*, 1994, pp. 149–158.

[21] H. Fehri, A. Gooya, Y. Lu, E. Meijering, S. Johnston, and A. Frangi, "Bayesian polytrees with learned deep features for multi-class cell segmentation," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3246–3260, Jul. 2019.

[22] M. S. Zaveri and D. Hammerstrom, "CMOL/CMOS implementations of Bayesian polytree inference: Digital and mixed-signal architectures and performance/price," *IEEE Trans. Nanotechnol.*, vol. 9, no. 2, pp. 194–211, Mar. 2010.

[23] L. E. Sucar, J. Pérez-Brito, J. C. Ruiz-Suárez, and E. Morales, "Learning structure from data and its application to ozone prediction," *Appl. Intell.*, vol. 7, pp. 327–338, Nov. 1997.

[24] M. E. Jakobsen, R. D. Shah, P. Bühlmann, and J. Peters, "Structure learning for directed trees," *J. Mach. Learn. Res.*, vol. 23, no. 159, pp. 1–97, 2022.

[25] E. Mokhtarian, S. Akbari, F. Jamshidi, J. Etesami, and N. Kiyavash, "Learning Bayesian networks in the presence of structural side information," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, Jun. 2022, pp. 7814–7822.
[26] G. Rebane and J. Pearl, "The recovery of causal poly-trees from statistical data," in *Proc. UAI*, 1987, pp. 222–228.
[27] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Trans. Inf. Theory*, vol. 14, no. 3, pp. 462–467, May 1968.
[28] X. Lou, Y. Hu, and X. Li, "Linear polytree structural equation models: Structural learning and inverse correlation estimation," 2021, *arXiv:2107.10955*.
[29] D. Tramontano, A. Monod, and M. Drton, "Learning linear non-Gaussian polytree models," in *Proc. 38th Conf. Uncertain. Artif. Intell.*, vol. 180, 2022, pp. 1960–1969.
[30] S. Chatterjee and M. Vidyasagar, "Estimating large causal polytree skeletons from small samples," 2022, *arXiv:2209.07028*.
[31] S. Linusson, P. Restadh, and L. Solus, "On the edges of characteristic imset polytopes," 2022, *arXiv:2209.07579*.
[32] M. Azadkia, A. Taeb, and P. Bühlmann, "A fast non-parametric approach for local causal structure learning," 2021, *arXiv:2111.14969*.
[33] J. Etesami, N. Kiyavash, and T. Coleman, "Learning minimal latent directed information polytrees," *Neural Comput.*, vol. 28, no. 9, pp. 1723–1768, Sep. 2016.
[34] F. Sepehr and D. Materassi, "An algorithm to learn polytree networks with hidden nodes," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 15084–15093.
[35] S. Dasgupta, "Learning polytrees," in *Proc. 15th Conf. Uncertain. Artif. Intell.*, 1999, pp. 134–141.
[36] N. Grüttemeier, C. Komusiewicz, and N. Morawietz, "On the parameterized complexity of polytree learning," in *Proc. 30th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2021, pp. 4221–4227.
[37] C. Améndola, M. Drton, A. Grosdos, R. Homs, and E. Robeva, "Third-order moment varieties of linear non-gaussian graphical models," *Inf. Inference J.*, vol. 12, no. 3, pp. 1405–1436, 2023.
[38] M. Maathuis, M. Drton, S. Lauritzen, and M. Wainwright, Eds., *Handbook of Graphical Models* (Handbooks of Modern Statistical Methods). Boca Raton, FL, USA: CRC Press, 2019.
[39] T. Richardson and P. Spirtes, "Ancestral graph Markov models," *Ann. Statist.*, vol. 30, no. 4, pp. 962–1030, 2002.
[40] S. Sullivant, K. Talaska, and J. Draisma, "Trek separation for Gaussian graphical models," *Ann. Statist.*, vol. 38, no. 3, pp. 1665–1685, 2010.
[41] M. Kalisch and P. Bühlmann, "Estimating high-dimensional directed acyclic graphs with the PC-algorithm," *J. Mach. Learn. Res.*, vol. 8, no. 22, pp. 613–636, 2007.
[42] J. B. Kruskal, Jr., "On the shortest spanning subtree of a graph and the traveling salesman problem," *Proc. Amer. Math. Soc.*, vol. 7, pp. 48–50, Feb. 1956.
[43] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Hoboken, NJ, USA: Wiley-Intersci., 2006.
[44] C. Heinze-Deml, J. Peters, and N. Meinshausen, "Invariant causal prediction for nonlinear models," *J. Causal Inference*, vol. 6, no. 2, 2018, Art. no. 20170016.
[45] A. Ghassami, S. Salehkaleybar, N. Kiyavash, and K. Zhang, "Learning causal structures using regression invariance," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 3011–3021.
[46] T. Toyoda, "Use of the chow test under heteroscedasticity," *Econometrica*, vol. 42, no. 3, pp. 601–608, 1974.
[47] G. C. Chow, "Tests of equality between sets of coefficients in two linear regressions," *Econometrica*, vol. 28, no. 3, pp. 591–605, 1960.
[48] D. Anderson and K. Burnham, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, vol. 63. New York, NY, USA: Springer, 2004, p. 10.
[49] W. H. Press, *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge, U.K.: Cambridge Univ. Press, 2007.
[50] A. Hauser and P. Bühlmann, "Jointly interventional and observational data: Estimation of interventional Markov equivalence classes of directed acyclic graphs," *J. Roy. Stat. Soc. B. Stat. Methodol.*, vol. 77, no. 1, pp. 291–318, 2015.
[51] D. Tramontano, L. Waldman, and E. Duarte. "Learning linear Gaussian polytrees with interventions." 2023. [Online]. Available: https://github.com/emduart2/Polytrees
[52] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan, "Causal protein-signaling networks derived from Multiparameter single-cell data," *Science*, vol. 308, no. 5721, pp. 523–529, 2005.
[53] A. Jaber, M. Kocaoglu, K. Shanmugam, and E. Bareinboim, "Causal discovery from soft interventions with unknown targets: Characterization and learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1–11.

Daniele Tramontano was born in Naples, Italy. He received the bachelor's degree in mathematics from the University of Naples "Federico II" in 2019, and the master's degree in mathematics from the University of Padua in 2020. He is currently pursuing the Ph.D. degree in mathematical statistics with the Technical University of Munich, as part of the Imperial–TUM Joint Academy of Doctoral Studies. His research is supervised jointly by M. Drton from TUM and A. Monod from Imperial College London.

L. Waldmann, photograph and biography not available at the time of publication.

M. Drton, photograph and biography not available at the time of publication.

Eliana Duarte was born in Bogotá, Colombia. She received the Ph.D. degree in mathematics from the University of Illinois Urbana–Champaign in 2017, and held postdoctoral positions with the Nonlinear Algebra Group, Max Planck Institute for Mathematics in the Sciences (MPI-MIS), Leipzig, Germany, and the Research Training Group Mathematical Complexity Reduction, Otto Von Guericke Universität Magdeburg. In 2021, she moved to Portugal funded by the Fundação da Ciência e a Tecnologia. Since 2023, she led the research group in Algebraic Statistics with MPI-MIS and since May 2023, she has been a Assistant Professor of Probability and Statistics with the Faculdade de Ciências at Universidade do Porto, Portugal.