Structure Learning for Directed Trees

Journal Article

Author(s):
Jakobsen, Martin E.; Shah, Rajen D.; Bühlmann, Peter; Peters, Jonas

Publication date:
2022-05

Permanent link:
https://doi.org/10.3929/ethz-b-000553211

Originally published in:
Journal of Machine Learning Research 23

Funding acknowledgement:
786461 - Statistics, Prediction and Causality for Large-Scale Data (EC)

Journal of Machine Learning Research 23 (2022) 1-97 Submitted 9/21; Revised 3/22; Published 5/22

Structure Learning for Directed Trees

Martin Emil Jakobsen M.JAKOBSEN@MATH.KU.DK
Department of Mathematical Sciences
University of Copenhagen
Copenhagen, Denmark

Rajen D. Shah R.SHAH@STATSLAB.CAM.AC.UK
Statistical Laboratory
University of Cambridge
Cambridge, UK

Peter Bühlmann BUHLMANN@STAT.MATH.ETHZ.CH
Seminar for Statistics
ETH Zurich
Zurich, Switzerland

Jonas Peters JONAS.PETERS@MATH.KU.DK
Department of Mathematical Sciences
University of Copenhagen
Copenhagen, Denmark

Editor: Aapo Hyvarinen

Abstract

Knowing the causal structure of a system is of fundamental interest in many areas of science and can aid the design of prediction algorithms that work well under manipulations to the system. The causal structure becomes identifiable from the observational distribution under certain restrictions. To learn the structure from data, score-based methods evaluate different graphs according to the quality of their fits. However, for large, continuous, and nonlinear models, these rely on heuristic optimization approaches with no general guarantees of recovering the true causal structure. In this paper, we consider structure learning of directed trees. We propose a fast and scalable method based on Chu–Liu–Edmonds' algorithm we call causal additive trees (CAT). For the case of Gaussian errors, we prove consistency in an asymptotic regime with a vanishing identifiability gap. We also introduce two methods for testing substructure hypotheses with asymptotic family-wise error rate control that is valid post-selection and in unidentified settings. Furthermore, we study the identifiability gap, which quantifies how much better the true causal model fits the observational distribution, and prove that it is lower bounded by local properties of the causal model. Simulation studies demonstrate the favorable performance of CAT compared to competing structure learning methods.

Keywords: Causality, restricted causal models, structure learning, directed trees, hypothesis testing.

1. Introduction

Learning the underlying causal structure of a stochastic system involving the random vector $X = (X_1, \dots, X_p)$ is an important problem in economics, industry, and science. Knowing the causal structure allows researchers to understand whether X_i causes X_j (or vice versa) and how a system reacts under an intervention. However, it is not generally possible to learn the causal structure (or parts thereof) from the observational data of a system alone. Without further restrictions on the system of interest there might exist another system with a different causal structure inducing the same observational distribution, i.e., the structure might not be identifiable from observed data.

Common structure learning methods using observational data are constraint-based (e.g., Pearl, 2009; Spirtes et al., 2000), score-based (e.g., Chickering, 2002), or a mix thereof (e.g., Nandy et al., 2018). Each of these approaches requires different assumptions to ensure identifiability of the causal structure and consistency of the approach. In structural causal models, one assumes that there are (causal) functions $f_1, \dots, f_p$ such that for all

$$1 \leq i \leq p: \qquad X_i := f_i(X_{\mathrm{PA}(i)}, N_i),$$

for subsets $\mathrm{PA}(i) \subset \{1, ..., p\}$ and jointly independent noise variables $N = (N_1, ..., N_p) \sim P_N$ (see Definition 1 for a precise definition including further restrictions). The causal graph is constructed as follows: for each variable X_i one adds directed edges from its direct causes or parents $\mathrm{PA}(i)$ into i. For such models, system assumptions concerning the causal functions can make the causal graph identified from the observational distribution. Specific assumptions that guarantee identifiability of the causal graph have been studied for, e.g., linear additive Gaussian noise models with equal noise variance (Peters and Bühlmann, 2014), linear additive non-Gaussian noise models (Shimizu et al., 2006), nonlinear additive noise models (Hoyer et al., 2008; Peters et al., 2014), post-nonlinear additive noise models (Zhang and Hyvärinen, 2009), partially-linear additive Gaussian noise models (Rothenhäusler et al., 2018) and discrete models (Peters et al., 2011).

Score-based structure learning usually starts with a function ℓ assigning a population score to causal structures. Depending on the assumed model class, this function is minimized by the true structure. For example, when considering directed acylic graph (DAGs), the true causal DAG $\mathcal{G}$ satisfy

$$\mathcal{G} \in \operatorname*{arg\,min}_{\tilde{\mathcal{G}} : \tilde{\mathcal{G}} \text{ is a DAG}} \ell(\tilde{\mathcal{G}}). \tag{1}$$

The idea is then to estimate the score from a finite sample and minimize the empirical score over all DAGs. As the cardinality of the space of all DAGs grows super-exponentially in the number of nodes p (Chickering, 2002), brute-force minimization becomes computationally infeasible even for moderately large systems.[1]

For linear additive Gaussian noise models, assuming the Markov conditions and faithfulness, one can recover the correct Markov equivalence class (MEC) of $\mathcal{G}$, which can be represented by a unique completed partially directed acyclic graph (CPDAG) (Pearl, 2009). The optimization can be done greedily over MECs with greedy equivalent search (GES, Chickering, 2002) or over DAGs (Tsamardinos et al., 2006) and in the former case, the

1. For example, there are over 10^{275} distinct directed acyclic graphs over 40 nodes (Sloane, 2021).

method is known to be consistent. More specifically, the output of GES search is not guaranteed, for a fixed sample size, to solve the empirical version of Equation (1) but it solves the problem with probability tending to one in the large sample limit.

Chickering (1996) showed that, in general, solving the problem in Equation (1) is an NP-hard problem, even if we restrict the search to MECs for structures with fixed causal indegree of $K > 2$. Several exact exponential runtime algorithms have been proposed, for example, A* search (Yuan et al., 2011; Yuan and Malone, 2013) and CPBayes (van Beek and Hoffmann, 2015) for discrete systems, algorithms based on integer linear programming (Jaakkola et al., 2010; Cussens et al., 2017; Cussens, 2011), and algorithms based on dynamic programming (Koivisto and Sood, 2004; Silander and Myllymäki, 2006; Parviainen and Koivisto, 2009).

In the nonlinear additive Gaussian noise case, Bühlmann et al. (2014) show that nonparametric maximum-likelihood estimation consistently estimates the correct causal order. However, the greedy search algorithm minimizing the score function does not come with any theoretical guarantees. Other heuristic approaches (for discrete or linear Gaussian systems) include acyclic selection ordering-based search (Scanagatta et al., 2015), memetic insert neighbourhood ordering-based search (Lee and Beek, 2017), and max-min hill-climb (Tsamardinos et al., 2006). Recently, methods have been proposed that perform continuous, non-convex optimization (Zheng et al., 2018) but such methods are without guarantees and it is currently debated whether they exploit some artifacts in simulated data (Reisach et al., 2021). Thus, for nonlinear models, there is currently no score-based method that provably guarantees recovery of the true causal graph with high probability.

In this paper we focus on models of reduced complexity, namely models with directed trees as causal graphs. This complexity reduction allow for polynomial runtime minimization of the score-function using the Chu–Liu–Edmonds' algorithm (proposed independently by Chu and Liu, 1965; Edmonds, 1967) and it allows for the derivation of hypothesis testing theory. As such the structure learning problem remains computationally feasible even for very large systems. Our method is called causal additive trees (CAT). The method is easy to implement and consists of two steps. In the first step, we employ user-specified (univariate) regression methods to estimate the conditional expectations $x \mapsto \mathbb{E}[X_i|X_j = x]$ for all $i \neq j$. We then use these to construct edge weights as inputs to the Chu–Liu–Edmonds' algorithm. This algorithm then outputs a directed tree with minimal edge weight, corresponding to a directed tree minimizing the score in Equation (1).

1.1 Contributions

We now highlight four main contributions of the paper:

(i) Computational feasibility: Assuming an identifiable model class, such as additive noise, allows us to infer the causal DAG by minimizing Equation (1) for a suitable score function. However, even for trees, the cardinality of the search space grows super-exponentially in the number of variables p. Hence, brute-force minimization (exhaustive search) in Equation (1) remains computationally infeasible for large systems. We propose the score-based method CAT (based on Chu–Liu–Edmonds' algorithm) and prove that it recovers the causal tree with a run-time complexity of $\mathcal{O}(p^2)$. This method can be useful even when not restricting onself to the class of directed trees: e.g., when using a heuristic method such as

greedy search for aiming to find an optimal scoring DAG, one can use the score of the optimal scoring tree as a sanity check or the corresponding tree for initialization.

(ii) Consistency: We prove that CAT is pointwise consistent in an identified additive Gaussian noise setup. That is, we recover the causal directed tree with probability tending to one as the sample size increases. Consistency only requires that the regression methods for estimating the conditional mean functions have mean squared prediction error converging to zero in probability. This property that is satisfied by many nonparametric regression methods such as nearest neighbors, neural networks, or kernel methods (see e.g. Györfi et al., 2002). Moreover, the vanishing estimation error is only required for causal edges for which the conditional means coincide with the causal functions. We also derive sufficient conditions that ensure consistency in an asymptotic setup with vanishing identifiability. Specifically, we show that consistency is retained even when the identifiability gap decreases at a rate q_n with $q_n^{-1} = o(\sqrt{n})$ as long as the conditional expectation mean squared prediction error corresponding to the causal edges vanishes at a rate $o_p(q_n)$.

(iii) Hypothesis testing: We provide two algorithms for performing hypothesis tests concerning the presence and absence of substructures, such as particular edges, in the true causal graph. The type I error is controlled asymptotically when the mean squared prediction error of the regression corresponding to the true causal edges decays at a relatively slow $o_p(n^{-1/2})$ rate. The tests are valid post-selection, that is, the hypotheses to be tested may be chosen after the graph has been estimated, and when multiple tests are performed, the family-wise error rate is controlled for any number of tests. Furthermore, one of the two proposed testing procedures is valid in the non-identified setting.

(iv) Identifiability analysis: We analyze the identifiability gap, that is, the smallest population score difference between an alternative graph and the causal graph. The reduced system complexity, due to the restriction to trees, allows us to derive simple yet informative lower bounds. For additive Gaussian noise models, for example, the lower bound can be computed using only local properties of the underlying model: it is based on a first term that considers the minimal score gap between individual edge reversals and a second term involving the minimal mutual information of two neighboring nodes, when conditioning on another neighbor of the parent node.

1.2 Related Constraint-based Approaches

As an alternative to score-based methods, constraint-based methods such as PC or FCI (Spirtes et al., 2000) test for conditional independences statements in P_X and use these results to infer (parts of) the causal structure. Such methods usually assume that P_X is both Markov and faithful with respect to the causal graph $\mathcal{G}$. Under these assumptions, the Markov equivalence class of the causal graph $\mathcal{G}$ is identified. In a jointly Gaussian setting (e.g. linear additive Gaussian noise models), consistency of constraint-based approaches relies on faithfulness, whereas uniform consistency requires strong faithfulness (see, e.g., Zhang and Spirtes, 2002; Kalisch and Bühlman, 2007) – a condition that has been shown to be strong (Uhler et al., 2013). In nonlinear settings, corresponding guarantees do not exist. This may at least partially be due to the fact that conditional independence testing is known to be a hard statistical problem (Shah and Peters, 2020).

Constraint-based methods have also been studied for polytrees. A polytree is a DAG whose undirected graph is a tree. Polytrees, unlike directed trees, allow for multiple root nodes as well as nodes with multiple parents. Rebane and Pearl (1987), inspired by the work of Chow and Liu (1968), propose a constraint-based structure learning method for polytrees over discrete variables that can identify the correct skeleton and causal basins, structures constructed from nodes with at least two parents. More precisely, the skeleton is determined by the maximum weight spanning tree (MWST) algorithm with mutual information measure weights, while the directionality of edges is inferred by conditional independence constraints implied by the observed distribution. In the case of causal trees this constraint-based structure learning method cannot direct any edges because causal basins do not exist (Rebane and Pearl, 1987). Dominguez et al. (2013) and Ouerd (2000) extend the Rebane and Pearl (1987) algorithm for causal discovery to multivariate Gaussian polytree distributions. Friedman et al. (1997) propose a similar algorithm to learn tree Bayesian networks by finding a MWST with mutual information weights. This recovers the skeleton of the causal graph, after which an arbitrary root node is selected and all edges are oriented away from said root node. As such, the method of Friedman et al. (1997) is only guaranteed to recover a directed tree that is Markov equivalent to the causal directed tree.

In this work, we employ Chu–Liu–Edmonds' algorithm, a directed analogue of the MWST algorithm, to not only recover the skeleton but also the direction of all edges in the causal graph. This is possible since we consider restricted causal models, e.g., nonlinear additive Gaussian noise models. More specifically, these restricted causal models allow us to define edge weights that, unlike the mutual information weights, preserve directionality information. In fact, when discarding information that allows us to infer directionality of the edges, one recovers the mutual information weights of Rebane and Pearl (1987), see Remark 1 in Appendix B for details.

1.3 Organization of the Paper

In Section 2, we define the setup and relevant score functions. We further strengthen existing identifiability results for nonlinear additive noise models. In Section 3, we propose CAT, an algorithm solving the score-based structure learning problem that is based on Chu–Liu–Edmonds' algorithm. We prove consistency of CAT for a fixed distribution and for a setup with vanishing identifiability. In Section 4, we provide results on asymptotic normality of the scores, construct confidence regions and propose feasible testing procedures. Section 5, we analyzes the identifiability gap. Section 6 shows the results of various simulation experiments. All proofs can be found in Appendix D.

2. Score-based Learning and Identifiability of Trees

In the remainder of this work we use of the following graph terminology (a more detailed introduction can be found in Appendix A, see also Koller and Friedman, 2009). A directed graph $\mathcal{G} = (V, \mathcal{E})$ consists of $p \in \mathbb{N}_{>0}$ vertices (or nodes) $V = \{1, \dots, p\}$ and a collection of directed edges $\mathcal{E} \subset \{(i \to j) \equiv (i, j) : i, j \in V, i \neq j\}$. A directed acyclic graph (DAG) is a directed graph that does not contain any directed cycles. A directed tree is a connected DAG in which all nodes have at most one parent. The unique node of a directed tree $\mathcal{G}$

with no parents is called the root node and is denoted by $\mathrm{rt}(\mathcal{G})$. We let $\mathcal{T}_p$ denote the set of directed trees over $p \in \mathbb{N}_{>0}$ nodes.

2.1 Identifiability of Causal Additive Tree Models

We now revisit and strengthen known identifiability results on restricted structural causal models. Consider a distribution that is induced by a structural causal model (SCM) with additive noise. Then, there are only special cases (such as linear additive Gaussian noise models) for which alternative models with a different causal structure exist that generate the same distribution (see Peters et al., 2017, for an overview). To state and strengthen these results formally, we introduce the following notation.

For any $k \in \mathbb{N}$ we define the following classes of functions from $\mathbb{R}$ to $\mathbb{R}$: $\mathcal{M}$ denotes all measurable functions, $\mathcal{D}_k$ denotes the set of all k times differentiable functions and $\mathcal{C}_k$ denotes the k times continuously differentiable functions. We let $\mathcal{P}$ denote the set of mean zero probability measures on $\mathbb{R}$ that have a density with respect to Lebesgue measure. $\mathcal{P}_+ \subset \mathcal{P}$ denotes the subset for which a density is strictly positive. For any function class $\mathcal{F} \subseteq \{f | f : \mathbb{R} \to \mathbb{R}\}$, $\mathcal{P}_{\mathcal{F}} \subset \mathcal{P}$ denotes the subset with a density function in $\mathcal{F}$. As a special case, we let $\mathcal{P}_{\mathrm{G}} \subset \mathcal{P}_{+\mathcal{C}_\infty} := \mathcal{P}_+ \cap \mathcal{P}_{\mathcal{C}_\infty}$ denote the subset of Gaussian probability measures. For any set $\mathcal{P}$ of probability measures, $\mathcal{P}^p$ denotes all p-dimensional product measures on $\mathbb{R}^p$ with marginals in $\mathcal{P}$.

We now define structural causal additive tree models (or causal additive tree models, for short) as SCMs with a tree structure.[2]

Definition 1 (Structural causal additive tree models) *Consider a class $\mathcal{T}_p \times \mathcal{M}^p \times \mathcal{P}^p$. Any tuple $(\mathcal{G}, (f_i), P_N) \in \mathcal{T}_p \times \mathcal{M}^p \times \mathcal{P}^p$ induces a structural causal model over $X = (X_1, \ldots, X_p)$ given by the following structural assignments*

$$X_i := f_i(X_{\mathrm{pa}^{\mathcal{G}}(i)}) + N_i, \quad \text{for all } 1 \leq i \leq p,$$

where $f_{\mathrm{rt}(\mathcal{G})} \equiv 0$ and $N = (N_1, \ldots, N_p) \sim P_N$, which we call a structural causal additive tree model. By slight abuse of notation, we write $Q \in \mathcal{T}_p \times \mathcal{M}^p \times \mathcal{P}^p$ for a probability distribution that is induced by a structural causal additive tree model.

Furthermore, we define the set of restricted structural causal additive tree models. We will see later that for these models, the causal graph is identifiable from the observable distribution of the system. When the causal graph of a sufficiently nice additive noise SCM is not identifiable, then certain differential equations must hold (see the proof of Proposition 4 for details). The definition of restricted structural causal additive tree models ensures that this does not happen.

Definition 2 (Restricted structural causal additive tree models) *The collection of restricted structural causal additive tree models $\Theta_R \subset \mathcal{T}_p \times \mathcal{D}_3^p \times \mathcal{P}_{+\mathcal{C}_3}^p$ is given by all models $\theta = (\mathcal{G}, (f_i), P_N) \in \mathcal{T}_p \times \mathcal{D}_3^p \times \mathcal{P}_{+\mathcal{C}_3}^p$ satisfying the following conditions for all $i \in \{1, \ldots, p\} \setminus \{\mathrm{rt}(\mathcal{G})\}$:*

(i) f_i is nowhere constant, i.e., it is not constant on any non-empty open set, and

2. This model class comes with the strong assumption on additive noise, which excludes certain types of hidden confounding, for example.

(ii) the induced log-density ξ of $X_{\mathrm{pa}^{\mathcal{G}}(i)}$, noise log-density ν of N_i and causal function f_i are such that there exists $x, y \in \mathbb{R}$ with $\nu''(y - f_i(x))f_i'(x) \neq 0$ such that

$$\xi''' \neq \xi'' \left(\frac{f_i''}{f_i'} - \frac{\nu''' f_i'}{\nu''} \right) - 2\nu'' f_i'' f_i' + \nu' f_i''' + \frac{\nu'\nu''' f_i'' f_i'}{\nu''} - \frac{\nu'(f_i'')^2}{f'}, \tag{2}$$

where the derivatives of ξ, ν and f_i are evaluated in x, $y - f_i(x)$ and x, respectively.

The following lemma, due to Hoyer et al. (2008), shows that for causal additive tree models with Gaussian noise, the differential equation constraints of Definition 2 simplify.[3]

Lemma 3 *Let $\theta = (\mathcal{G}, (f_i), P_N) \in \mathcal{T}_p \times \mathcal{D}_3^p \times \mathcal{P}_{\mathrm{G}}^p$. Assume that for all $i \in \{1, \ldots, p\} \setminus \{\mathrm{rt}(\mathcal{G})\}$ the following two conditions hold (a) f_i is nowhere constant and (b) f_i is not linear. Then, $\theta \in \Theta_R$.*

Existing identifiability results for causal graphs in restricted SCMs (Hoyer et al., 2008; Peters et al., 2014) are stated and proven in terms of the ability to distinguish the induced distributions of two restricted structural causal models: For all $\theta = (\mathcal{G}, \ldots) \in \Theta_R$ and $\tilde{\theta} = (\tilde{\mathcal{G}}, \ldots) \in \Theta_R$, if $\mathcal{G} \neq \tilde{\mathcal{G}}$, then $\mathcal{L}(X_\theta) \neq \mathcal{L}(X_{\tilde{\theta}})$ (where $\mathcal{L}$ denotes the distribution of a random variable), that is, X_θ and $X_{\tilde{\theta}}$ do not have the same distribution. We now prove a stronger identifiability result that does not assume that $\tilde{\theta}$ is a restricted causal model.

Proposition 4 (Identifiability of causal additive tree models) *Suppose that X_θ and $X_{\tilde{\theta}}$ are generated by the SCMs $\theta = (\mathcal{G}, (f_i), P_N) \in \Theta_R \subset \mathcal{T}_p \times \mathcal{D}_3^p \times \mathcal{P}_{+\mathcal{C}_3}^p$ and $\tilde{\theta} = (\tilde{\mathcal{G}}, (\tilde{f}_i), \tilde{P}_N) \in \mathcal{T}_p \times \mathcal{D}_1^p \times \mathcal{P}_{\mathcal{C}_0}^p$, respectively. It holds that*

$$\mathcal{L}(X_\theta) = \mathcal{L}(X_{\tilde{\theta}}) \implies \mathcal{G} = \tilde{\mathcal{G}}.$$

We prove Proposition 4 using the techniques of Peters et al. (2014). While we prove the statement only for restricted causal additive tree models, which suffices for this work, we conjecture that a similar extension holds for restricted structural causal DAG models. The extension of Proposition 4 is important for the following reason. Given a finite data set, practical methods usually assume that the true distribution is induced by an underlying restricted SCM. One can then fit different causal structures and output the structure that fits the data best. The above extension accounts for the fact that regression methods hardly represent all such restrictions: e.g., most nonlinear regression techniques can also fit linear models.

2.2 Score Functions

We now define population score functions which are later used to recover the causal tree. We henceforth assume that $X \in \mathbb{R}^p$ is a random vector with distribution P_X generated by a restricted causal additive tree model $\theta = (\mathcal{G}, (f_i), P_N) \in \Theta_R \subset \mathcal{T}_p \times \mathcal{D}_3^p \times \mathcal{P}_{+\mathcal{C}_3}^p$ with $\mathcal{G} = (V, \mathcal{E}) \in \mathcal{T}_p$ such that $\mathbb{E}\|X\|_2^2 < \infty$. Thus, $\mathcal{G}$ denotes the causal tree. We use $\tilde{\mathcal{G}} \in \mathcal{T}_p$ to denote an arbitrary, different (directed) tree. For the remainder of this paper, we assume

3. For completeness, we include the proof of Lemma 3 in Appendix D, using the approach of Zhang and Hyvärinen (2009) but expressed in our notation.

that for any $i \neq j$ it holds that $X_i - \mathbb{E}[X_i|X_j]$ has a density with respect to Lebesgue measure.[4] We often refer to one of the following two scenarios: either, *(i)*, we have limited a priori information that $P_N \in \mathcal{P}^p_{+\mathcal{C}_3}$, or, *(ii)*, we know that the noise innovations are Gaussian, that is, $P_N \in \mathcal{P}^p_{\mathrm{G}}$.

Definition 5 *For any graph $\tilde{\mathcal{G}} \in \mathcal{T}_p$ we define for each node $i \in V$ the*

(i) local Gaussian score as $\ell_{\mathrm{G}}(\tilde{\mathcal{G}}, i) := \log\left(\mathrm{Var}\left(X_i - \mathbb{E}\left[X_i|X_{\mathrm{pa}_{\tilde{\mathcal{G}}}(i)}\right]\right)\right)/2$,

(ii) local entropy score as $\ell_{\mathrm{E}}(\tilde{\mathcal{G}}, i) := h\left(X_i - \mathbb{E}\left[X_i|X_{\mathrm{pa}_{\tilde{\mathcal{G}}}(i)}\right]\right)$,

(iii) local conditional entropy score as $\ell_{\mathrm{CE}}(\tilde{\mathcal{G}}, i) := h\left(X_i|X_{\mathrm{pa}_{\tilde{\mathcal{G}}}(i)}\right)$.

Here, we use the convention that $\mathbb{E}(X_i|\emptyset) = 0$ and $h(X_i|\emptyset) = h(X_i)$; the functions $h(\cdot)$, $h(\cdot|\cdot)$, and $h(\cdot,\cdot)$ (used below) denote the differential entropy, conditional entropy, and cross entropy, respectively. The Gaussian, entropy and conditional entropy score of $\tilde{\mathcal{G}}$ are, respectively, given by the sum of local scores:

$$\ell_{\mathrm{G}}(\tilde{\mathcal{G}}) := \sum_{i=1}^p \ell_{\mathrm{G}}(\tilde{\mathcal{G}}, i), \quad \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) := \sum_{i=1}^p \ell_{\mathrm{E}}(\tilde{\mathcal{G}}, i), \quad \ell_{\mathrm{CE}}(\tilde{\mathcal{G}}) := \sum_{i=1}^p \ell_{\mathrm{CE}}(\tilde{\mathcal{G}}, i).$$

(See Polyanskiy and Wu (2019) or Cover and Thomas (2006) for the basic information-theoretic concepts used in this paper.) Similar scores have been considered by Bühlmann et al. (2014) and Mooij et al. (2016), for example. For linear additive Gaussian noise systems, the Gaussian score of Definition 5 is proportional to the large sample limit of the Gaussian log-likelihood score function commonly used in for Bayesian network learning (see, e.g., Koller and Friedman, 2009).

The following lemma shows that the Gaussian score of the graph $\tilde{\mathcal{G}} \in \mathcal{T}_p$ arises naturally as a translated infimum cross entropy between P_X and all Q induced by causal additive tree models with Gaussian noise. Similarly, the entropy score can be seen as an infimum cross entropy between P_X and all Q induced by another class of SCMs.

Lemma 6 *For any $\tilde{\mathcal{G}} \in \mathcal{T}_p$ it holds that*

$$\ell_{\mathrm{G}}(\tilde{\mathcal{G}}) = \inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{D}_1^p \times \mathcal{P}_{\mathrm{G}}^p} h(P_X, Q) - p \log(\sqrt{2\pi e}).$$

Furthermore, with $\mathcal{F}(\tilde{\mathcal{G}}) := (\mathcal{F}_i(\tilde{\mathcal{G}}))_{1 \leq i \leq p}$, where $\mathcal{F}_i(\tilde{\mathcal{G}}) := \{x \mapsto \mathbb{E}[X_i|X_{\mathrm{pa}_{\tilde{\mathcal{G}}}(i)} = x]\}$ for all $1 \leq i \leq p$, it holds that

$$\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) = \inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p} h(P_X, Q).$$

4. This ensures that the entropy score function introduced in Definition 5 below is well-defined and that the analysis of the identifiability gap in Section 5 is valid.

Score-based methods identify the underlying structure by evaluating the score functions (or estimates thereof) on different graphs and choosing the best scoring graph. The difference between the score $\ell_\cdot(\mathcal{G})$ of the true graph and the score $\ell_\cdot(\tilde{\mathcal{G}})$ of the best scoring alternative graph $\tilde{\mathcal{G}}$ is an important property of the problem: e.g., if it would be zero, we could not identify the true graph from the scores. We, therefore, refer to expressions of the form $\min_{\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}} \ell_\cdot(\tilde{\mathcal{G}}) - \ell_\cdot(\mathcal{G})$ as the identifiability gap. In the remainder of this paper, we refer to strict positivity of the identifiability gap as Assumption 1.

Assumption 1 *If $\theta \in \Theta_R \subset \mathcal{T}_p \times \mathcal{D}_3^p \times \mathcal{P}_{\mathrm{G}}^p$ or $\theta \in \Theta_R \subset \mathcal{T}_p \times \mathcal{D}_3^p \times \mathcal{P}_{+\mathcal{C}_3}^p$ it holds that*

$$\min_{\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}} \ell_{\mathrm{G}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{G}}(\mathcal{G}) > 0 \quad \text{or} \quad \min_{\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) > 0, \tag{3}$$

respectively.

Assumption 1 does not trivially follow from the results further above. By arguments similar to those in Lemma 6 we have that, if the true data-generating model is a restricted causal additive tree model with Gaussian noise, $\theta \in \Theta_R \subset \mathcal{T}_p \times \mathcal{D}_3^p \times \mathcal{P}_{\mathrm{G}}^p$, then $\ell_{\mathrm{G}}(\mathcal{G}) = h(P_X) - p\log(\sqrt{2\pi e})$. Hence, the Gaussian score gap between $\tilde{\mathcal{G}}$ and the causal graph $\mathcal{G}$ equals

$$\ell_{\mathrm{G}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{G}}(\mathcal{G}) = \inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{D}_1^p \times \mathcal{P}_{\mathrm{G}}^p} h(P_X, Q) - h(P_X) = \inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{D}_1^p \times \mathcal{P}_{\mathrm{G}}^p} D_{\mathrm{KL}}(P_X \| Q),$$

where D_{KL} denotes the Kullback-Leibler divergence measure. Proposition 4 implies that

$$\forall \tilde{\mathcal{G}} \neq \mathcal{G}, \quad \forall Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{D}_1^p \times \mathcal{P}_{\mathrm{G}}^p : D_{\mathrm{KL}}(P_X \| Q) > 0.$$

However, this does not immediately imply that the identifiability gap (where we take the infimum over such Q) is strictly positive. Similar considerations[5] hold for the entropy score gap

$$\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) = \inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p} D_{\mathrm{KL}}(P_X \| Q).$$

In Section 5 we derive informative lower bounds on the Gaussian and entropy identifiability gaps (i.e., the infimum KL-divergence) of Equation (3). It is possible to enforce Assumption 1 indirectly by the assumptions and modifications detailed in the following lemma.

Lemma 7 *Assumption 1 holds if one of the following conditions is satisfied.*

(a) We have a restricted causal additive tree model with Gaussian noise $\theta \in \Theta_R \subset \mathcal{T}_p \times \mathcal{D}_3^p \times \mathcal{P}_{\mathrm{G}}^p$ and for all $i \neq j$ it holds that $x \mapsto \mathbb{E}[X_i | X_j = x]$ has a differentiable version.

(b) We have a restricted causal additive tree model with Gaussian noise $\theta \in \Theta_R \subset \mathcal{T}_p \times \mathcal{D}_3^p \times \mathcal{P}_{\mathrm{G}}^p$ and for all $1 \leq i \leq p$ it holds that the causal function f_i is contained within a function class $\mathcal{F}_i \subseteq \mathcal{D}_1$ which satisfies $\arg\min_{f' \in \mathcal{F}_i} \mathbb{E}[(X_i - f'(X_j))^2] \in \mathcal{F}_i$ for all $j \neq i$, and we consider a modified Gaussian score function $\ell_{\mathrm{G.mod}} : \mathcal{T}_p \to \mathbb{R}$ with local score given by $\ell_{\mathrm{G.mod}}(\tilde{\mathcal{G}}, i) := \log(\min_{\tilde{f} \in \mathcal{F}_i} \mathbb{E}[(X_i - \tilde{f}(X_{\mathrm{pa}_{\tilde{\mathcal{G}}}(i)}))^2])/2$.

5. In fact, Proposition 4 does not immediately imply that $D_{\mathrm{KL}}(P_X \| Q) > 0$ for $Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p$ as it does not necessarily hold that the causal functions in $\mathcal{F}(\tilde{\mathcal{G}})$ are differentiable or that the noise innovation densities in $\mathcal{P}^p$ are continuous.

(c) We have a restricted causal additive tree model $\theta \in \Theta_R \subset \mathcal{T}_p \times \mathcal{D}_3^p \times \mathcal{P}_{+\mathcal{C}_3}^p$, for all $i \neq j$ it holds that $x \mapsto \mathbb{E}[X_i|X_j = x]$ has a differentiable version and for all $i \neq j$ it holds that $X_i - \mathbb{E}[X_i|X_j]$ has a continuous density.

The modified Gaussian score function and restrictions of condition (b) in Lemma 7 coincides with the working conditions of Bühlmann et al. (2014). Alternative information-theoretic conditions guaranteeing that Assumption 1 holds are derived in Section 5. If Assumption 1 is satisfied, then we can use the score functions to identify the true causal graph of a restricted structural model: In the Gaussian noise setting, for example, we have

$$\mathcal{G} = \operatorname*{arg\,min}_{\tilde{\mathcal{G}} \in \mathcal{T}_p} \ell_{\mathrm{G}}(\tilde{\mathcal{G}}). \tag{4}$$

In practice, we consider estimates of the above quantities and optimize the corresponding empirical loss function. Solving Equation (4) (or its empirical counterpart) using exhaustive search is computationally intractable already for moderately large choices of p.[6] We now introduce CAT, a computationally efficient method that solves the optimization exactly.

3. Causal Additive Trees (CAT)

We introduce the population version of our algorithm CAT in Section 3.1 and discuss its finite sample version and asymptotic properties in Sections 3.2 and 3.3.

3.1 An Oracle Algorithm

Similarly as for the case of DAGs, the problem in Equation (4) is a combinatorial optimization problem, for which the cardinality of the search space grows super-exponentially with p. Indeed, the number of undirected trees on p labelled nodes is p^{p-2} (Cayley, 1889) and therefore p^{p-1} is the corresponding number of labelled trees. For the class of DAGs (which includes directed trees), existing structure learning such as Bühlmann et al. (2014) propose a greedy search technique that iteratively selects the lowest scoring directed edge under the constraint that no cycles is introduced in the resulting graph. In general, greedy search procedures do not come with any guarantees and there are indeed situations in which they fail. By exploiting the assumption of a tree structure, we will see that the optimization problem of Equation (4) can be solved computationally efficiently without the need for heuristic optimization techniques.

Provided with a connected directed graph with edge weights, Chu–Liu–Edmonds' algorithm finds a minimum edge weight directed spanning tree, given that such a directed tree exists. That is, for a connected directed graph $\mathcal{H} = (V, \mathcal{E}_{\mathcal{H}})$ on the nodes $V = \{1, \dots, p\}$ with edge weights $w := \{w_{ji} : (j \to i) \in \mathcal{E}_{\mathcal{H}}\}$, Chu–Liu–Edmonds' algorithm recovers a minimum edge weight directed spanning tree (MWDST) subgraph of $\mathcal{H}$,

$$\operatorname*{arg\,min}_{\tilde{\mathcal{G}}=(V,\tilde{\mathcal{E}}) \in \mathcal{T}_p \cap \mathcal{H}} \sum_{(j \to i) \in \tilde{\mathcal{E}}} w_{ji},$$

6. In the context of linear Gaussian noise models, Chickering (2002) proves consistency of greedy equivalent search towards the correct Markov equivalence class. This, however, does not imply that the optimization problem in Equation (4) is solved: for a given sample, the method is not guaranteed to find the optimal scoring graph (but the output will converge to the correct graph).

where $\mathcal{T}_p \cap \mathcal{H}$ denotes all directed spanning trees of $\mathcal{H}$. The runtime of the original algorithms of Chu and Liu (1965) and Edmonds (1967) is $\mathcal{O}(|\mathcal{E}_\mathcal{H}| \cdot p) \leq \mathcal{O}(p^3)$. Karp (1971) presented an alternative proof for the correctness of the algorithm of Edmonds (1967). Tarjan (1977) devised a modification (corrected by Camerini et al., 1979) with runtime $\mathcal{O}(\min\{|\mathcal{E}_\mathcal{H}| \log(p), p^2\})$.[7] Gabow et al. (1986) devised yet another modification with runtime $O(p \log p + |\mathcal{E}_H|)$ and noted that no further improvements to the algorithm can be made (since it uses only binary decisions and can be used to sort p numbers). In our experiments, we use the `C++` implementation of Tarjans modification by Tofigh and Sjölund (2007) which is contained in the R-package `RBGL` (Carey et al., 2021) and the Python implementation of Edmonds' version from the Python-package `NetworkX` (Hagberg et al., 2022).

The causal graph recovery problem in Equation (4) is equivalently solved by finding a minimum edge weight directed tree, i.e., a minimum edge weight directed spanning tree of the fully connected graph on the nodes V. For example, finding the minimum of the Gaussian score function is equivalent to minimizing a translated version of the Gaussian score function

$$\begin{aligned}\operatorname*{arg\,min}_{\tilde{\mathcal{G}} \in \mathcal{T}_p} \ell_{\mathrm{G}}(\tilde{\mathcal{G}}) &= \operatorname*{arg\,min}_{\tilde{\mathcal{G}} \in \mathcal{T}_p} \sum_{i=1}^{p} \frac{1}{2} \log(\operatorname{Var}(X_i - \mathbb{E}[X_i | X_{\operatorname{pa}_{\tilde{\mathcal{G}}}(i)}])) - \sum_{i=1}^{p} \frac{1}{2} \log(\operatorname{Var}(X_i)) \\ &= \operatorname*{arg\,min}_{\tilde{\mathcal{G}} \in \mathcal{T}_p} \sum_{i=1}^{p} \frac{1}{2} \log\left(\frac{\operatorname{Var}(X_i - \mathbb{E}[X_i | X_{\operatorname{pa}_{\tilde{\mathcal{G}}}(i)}])}{\operatorname{Var}(X_i)} \right). \end{aligned} \tag{5}$$

Since the summand for the root node in Equation (5) note equals zero, we only need to sum over all nodes with an incoming edge in $\tilde{\mathcal{G}}$. Now define the Gaussian edge weights $w^{\mathrm{G}} := (w^{\mathrm{G}}_{ji})_{j \neq i}$ by

$$w^{\mathrm{G}}_{ji} := \frac{1}{2} \log\left(\frac{\operatorname{Var}(X_i - \mathbb{E}[X_i | X_j])}{\operatorname{Var}(X_i)} \right), \tag{6}$$

for all $j \neq i$. Hence, for a causal additive tree model with Gaussian noise satisfying Assumption 1 it holds that the causal directed tree is given by the MWDST with respect to the Gaussian edge weights,

$$\mathcal{G} = \operatorname*{arg\,min}_{\tilde{\mathcal{G}} \in \mathcal{T}_p} \ell_{\mathrm{G}}(\tilde{\mathcal{G}}) = \operatorname*{arg\,min}_{\tilde{\mathcal{G}} = (V, \tilde{\mathcal{E}}) \in \mathcal{T}_p} \sum_{(j \to i) \in \tilde{\mathcal{E}}} w^{\mathrm{G}}_{ji}.$$

Similarly, the minimum of the entropy score function is given by the MWDST with respect to the entropy edge weights $w^{\mathrm{E}} := (w^{\mathrm{E}}_{ji})_{j \neq i}$ given by $w^{\mathrm{E}}_{ji} := h(X_i - \mathbb{E}[X_i | X_j]) - h(X_i)$, for all $j \neq i$. We will henceforth denote the method where we apply Chu–Liu–Edmonds' algorithm to find the MWDST with respect to the Gaussian and entropy edge weights as CAT.G and CAT.E, respectively.

7. The algorithm presented in both Edmonds (1967) and Tarjan (1977) find minimum branchings of $\mathcal{H}$, i.e., directed forest spanning subgraphs of $\mathcal{H}$ with minimum edge weight. Note that the MWDST problem is invariant to identical translation of all edge weights. If $\mathcal{H}$ is a fully connected graph and we translate all edge weights $w'_{ji} := w_{ji} - \varepsilon \max\{w_{ji} : j \neq i\}$ for $\varepsilon > 1$, then a minimum branching using edge weights (w'_{ji}) is a MWDST subgraph of $\mathcal{H}$. For testing purposes, we also need to be able to find MWDST subgraphs of non-fully connected graphs $\mathcal{H}$, hence, as noted by Edmonds (1967), if we translate all edge weights $w'_{ji} := w_{ji} - \sum_{j \neq i} |w_{ji}|$, then a minimum branching using edge weights (w'_{ji}) is a MWDST subgraph of $\mathcal{H}$.

3.2 Finite Sample Algorithm

Given an $n \times p$ data matrix $\mathbf{X}_n$, representing n i.i.d. copies of $X = (X_1, \ldots, X_p)$, we estimate the edge weights by simple plug-in estimators. Let us denote the conditional expectation function and its estimate by

$$\varphi_{ji}(x) := \mathbb{E}[X_i | X_j = x], \qquad \hat{\varphi}_{ji}(x) := \hat{\mathbb{E}}[X_i | X_j = x], \tag{7}$$

for all $j \neq i$. The empirical Gaussian edge weights $\hat{w}^{\mathrm{G}} = (\hat{w}^{\mathrm{G}}_{ji})_{j \neq i}$ are then given by

$$\hat{w}^{\mathrm{G}}_{ji} := \frac{1}{2} \log \left(\frac{\widehat{\mathrm{Var}}(X_i - \hat{\varphi}_{ji}(X_j))}{\widehat{\mathrm{Var}}(X_i)} \right), \tag{8}$$

for all $i \neq j$, where $\widehat{\mathrm{Var}}(\cdot)$ denotes a variance estimator using the sample $\mathbf{X}_n$. We now propose to combine the Chu–Liu–Edmonds' algorithm described above with the Gaussian score as detailed in Algorithm 1. It is also possible to combine CAT with standard pruning techniques (see, e.g., Bühlmann et al., 2014) that, e.g., based on approximate p-values, remove insignificant edges and output directed forests. An R implementation of CAT with options for cross-fitting and pruning is available on GitHub.[8]

Algorithm 1 Causal additive trees (CAT)

1: **procedure** CAT($\mathbf{X}_n$, regression method)
2: Run regression method to obtain $\hat{\varphi}_{ji}$ for all $j \neq i$.
3: Compute empirical edge weights $\hat{w}^{\mathrm{G}}$, see Equation (8).
4: Apply Chu–Liu–Edmonds' algorithm to find MWDST with respect to $\hat{w}^{\mathrm{G}}$.
5: **return** MWDST $\hat{\mathcal{G}}$.
6: **end procedure**

By default we suggest to use the empirical Gaussian edge weights as described in Algorithm 1. However, it is also possible to run Chu–Liu–Edmonds' algorithm on the empirical entropy edge weights $\hat{w}^{\mathrm{E}} = (\hat{w}^{\mathrm{E}}_{ji})_{j \neq i}$ given by

$$\hat{w}^{\mathrm{E}}_{ji} := \hat{h}(X_i - \hat{\varphi}_{ji}(X_j)) - \hat{h}(X_i),$$

for all $j \neq i$, where $\hat{h}(\cdot)$ denotes a user-specific entropy estimator using the observed data $\mathbf{X}_n$. Estimating differential entropy is a difficult statistical problem but we will later in Section 6 demonstrate by simulation experiments that it can be beneficial to use the estimated entropy edge weights when the additive noise distributions are highly non-Gaussian.

Under suitable conditions on the (possibly nonparametric) regression technique, we now show that the proposed algorithm consistently recovers the true causal graph of a causal additive tree model with Gaussian noise using the empirical Gaussian edge weights.

3.3 Consistency

We study a version of the CAT.G algorithm applied to a causal additive tree model with Gaussian noise where the regression estimates are trained on auxiliary data, simplifying

8. `https://github.com/MartinEmilJakobsen/CAT`

the theoretical analysis. We believe that consistency without sample splitting holds but may require some stronger conditions (in the experimental section, we do not use sample splitting). As such, we only view the sample splitting as a theoretical device for simplifying proofs but we do not recommend it in practical applications. For each n we let $\mathbf{X}_n = ((X_{1,i})_{1\le i\le p}, \dots, (X_{n,i})_{1\le i\le p})$ and $\tilde{\mathbf{X}}_n = ((\tilde{X}_{1,i})_{1\le i\le p}, \dots, (\tilde{X}_{n,i})_{1\le i\le p})$ denote independent datasets each consisting of n i.i.d. random variables with distribution identical to that of $X = (X_1, ..., X_p) \in \mathbb{R}^p$. We suppose that the regression estimates $\hat{\varphi}_{ji}$ have been trained on $\tilde{\mathbf{X}}_n$ and then compute the edge weights using $\mathbf{X}_n$ as in step 3 of Algorithm 1:

$$\hat{w}^{\mathrm{G}}_{ji} := \frac{1}{2}\log\left(\frac{\frac{1}{n}\sum_{k=1}^n (X_{k,i} - \hat{\varphi}_{ji}(X_{k,j}))^2}{\frac{1}{n}\sum_{k=1}^n X_{k,i}^2 - (\frac{1}{n}\sum_{k=1}^n X_{k,i})^2}\right). \tag{9}$$

The consistency results may be extended to cross-fitted edge weight estimators formed as an average of estimators of the form in (9) with the roles of the $\mathbf{X}_n$ and $\tilde{\mathbf{X}}_n$ samples interchanged, which would make full use of the available data. The following result shows pointwise consistency of CAT.G whenever the conditional mean estimation is weakly consistent.

Theorem 8 (Pointwise consistency) *Suppose that for all $j \neq i$ the following two conditions hold:*

(a) if $(j \to i) \in \mathcal{E}$, $\mathbb{E}[(\hat{\varphi}_{ji}(X_j) - \varphi_{ji}(X_j))^2 | \tilde{\mathbf{X}}_n] \xrightarrow{P}_n 0$;

(b) if $(j \to i) \notin \mathcal{E}$, $\mathbb{E}[(\hat{\varphi}_{ji}(X_j) - \tilde{\varphi}_{ji}(X_j))^2 | \tilde{\mathbf{X}}_n] \xrightarrow{P}_n 0$ for some fixed $\tilde{\varphi}_{ji} : \mathbb{R} \to \mathbb{R}$,

where φ_{ji} and $\hat{\varphi}_{ji}$ are defined in Equation (7). Furthermore, suppose that Assumption 1 holds. In the large sample limit, we recover the causal graph with probability one, that is

$$P(\hat{\mathcal{G}} = \mathcal{G}) \to_n 1,$$

where $\hat{\mathcal{G}}$ is the output of Algorithm 1 using weights $\hat{w}^{\mathrm{G}}$ given by Equation (9).

Theorem 8 states that under the given assumptions, the estimated graph will converge to the true causal graph with probability tending to one. In fact, the assumptions are fairly week: we only require weakly consistent estimation of the conditional means for edges that are present in the causal graph; these represent causal relationships and are often assumed to be smooth. This distinction allow us to employ regression techniques that are consistent only for those function classes that we consider reasonable for modeling the causal mechanisms. For non-causal edges, $(j \to i) \notin \mathcal{E}$, the estimator $\hat{\varphi}_{ji}$ only needs to converge to a function $\tilde{\varphi}_{ji}$, which does not necessarily need to be the conditional mean.

3.3.1 Consistency under Vanishing Identifiability

We now consider an asymptotic regime involving a sequence $(\theta_n)_{n\in\mathbb{N}}$ of SCMs with potentially changing conditional mean functions φ_{ji} and a vanishing identifiability gap. We have the following result.

Theorem 9 (Consistency under vanishing identifiability) *Let $(\theta_n)_{n\in\mathbb{N}}$ be a sequence of SCMs on $p \in \mathbb{N}$ nodes all with the same causal directed tree $\mathcal{G} = (V, \mathcal{E})$ such that*

(i) for $q_n := \min_{\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}} \ell_{\mathrm{G}}(\mathcal{G}) - \ell_{\mathrm{G}}(\tilde{\mathcal{G}})$ *(the gap of model* θ_n*), we have* $q_n^{-1} = o(\sqrt{n})$*;*

(ii) for all $(j \to i) \in \mathcal{E}$ *and* $\varepsilon > 0$, $P_{\theta_n}\left(q_n^{-1}\mathbb{E}_{\theta_n}\left[(\varphi_{ji}(X_j) - \hat{\varphi}_{ji}(X_j))^2 | \tilde{\mathbf{X}}_n\right] > \varepsilon\right) \to_n 0$*;*

(iii) for all $j \neq i$ *and* $\varepsilon > 0$, $P_{\theta_n}\left(\frac{q_n^{-2}}{n}\mathbb{E}_{\theta_n}\left[(\varphi_{ji}(X_j) - \hat{\varphi}_{ji}(X_j))^4 | \tilde{\mathbf{X}}_n\right] > \varepsilon\right) \to_n 0$*; and*

(iv) there exists $C > 0$ *such that for all* $j \neq i$ $\inf_n P_{\theta_n}(\mathrm{Var}_{\theta_n}(X_i|X_j) \leq C) = 1$ *and* $\sup_n \mathbb{E}_{\theta_n}\|X\|_2^4 < \infty$*.*

Then it holds that

$$P(\hat{\mathcal{G}} = \mathcal{G}) \to_n 1.$$

Condition (i) asks that the identifiability gap q_n goes to zero more slowly than the standard convergence rate $1/\sqrt{n}$ of estimators in regular parametric models. Such a requirement would be necessary in almost any structure identification problem. Condition (ii) requires the mean squared error of the regression estimates corresponding to true causal edges to be $o_P(q_n)$. We regard this as a fairly mild assumption: indeed, the minimax rate of estimation of regression functions in Hölder balls with smoothness β is $n^{-2\beta/(2\beta+1)}$ (Tsybakov, 2009). Thus, we can expect that if the causal regression functions have smoothness $\beta \geq 1/2$ and all lie in a Hölder ball, (ii) can be satisfied for any q_n satisfying (i). Condition (iii) allows the fourth moments of the estimation errors to increase at any rate slower than $nq_n^2 \to \infty$; of course, we would typically expect this error to decay, at least for the causal edges.

4. Hypothesis Testing

This section presents two procedures to test any substructure hypothesis regarding the causal directed tree of a causal additive tree model with Gaussian noise. We continue our analysis using the sample split estimators of Equation (9), where the conditional expectations are estimated on an auxiliary dataset. Our approach makes use of the fact that the estimated weights in Equation (9) are logarithms of ratios of i.i.d. quantities, and thus the joint distribution of the estimated edge weights should, with appropriate centering and scaling, be asymptotically Gaussian; see Lemma D.4 in Appendix D for the precise statement. This allows us to create a (biased) confidence region of the true edge weights, which in turn gives a confidence set for the true graph. This confidence set of graphs is not necessarily straightforward to compute and list. However, we show that it can be queried to test hypotheses of interest, such as the presence or absence of a particular edge. As these hypothesis tests are derived from a confidence region, they are valid even when the hypothesis to test has been chosen after examining the data.

Similar to the results in the previous sections, we avoid making assumptions on the performance of regressions corresponding to non-causal edges. Unlike the consistency analysis, however, here we do not, in general, require identifiability of the true graph.

In order to state our results and assumptions, we introduce the following notation. For a collection of variables $(K_{ji})_{j \neq i}$, we let $K_i := (K_{1i}, \ldots, K_{(i-1)i}, K_{(i+1)i}, \ldots, K_{pi})^\intercal \in \mathbb{R}^{p-1}$, furthermore, for any collection $(K_i)_{1 \leq i \leq p}$, we let $K := (K_1, \ldots, K_p)^\intercal$. With this notation,

let, for all $k \in \{1, \dots, n\}$, the vectors of squared residuals and squared centered observations be given by

$$\hat{M}_k := \{(X_{k,i} - \hat{\varphi}_{ji}(X_{k,j}))^2\}_{j \neq i} \in \mathbb{R}^{p(p-1)}, \qquad \hat{V}_k = \left\{ \left(X_{k,i} - \frac{1}{n} \sum_{m=1}^{n} X_{m,i} \right)^2 \right\}_{1 \leq i \leq p} \in \mathbb{R}^p.$$

Further let

$$\hat{\mu} := \frac{1}{n} \sum_{k=1}^{n} \hat{M}_k, \qquad \hat{\nu} =: \frac{1}{n} \sum_{k=1}^{n} \hat{V}_k.$$

Note that with this notation, the empirical Gaussian edge weight for $j \to i$ is given by $\log(\hat{\mu}_{ji}/\hat{\nu}_i)/2$. Let us denote by $\widehat{\Sigma}_M \in \mathbb{R}^{p(p-1) \cdot p(p-1)}$, $\widehat{\Sigma}_V \in \mathbb{R}^{p \cdot p}$ and $\widehat{\Sigma}_{MV} \in \mathbb{R}^{p(p-1) \cdot p}$, the empirical variances of the $\hat{M}_k$ and $\hat{V}_k$ and their empirical covariance respectively, so

$$\hat{\Sigma} := \begin{pmatrix} \widehat{\Sigma}_M & \widehat{\Sigma}_{MV} \\ \widehat{\Sigma}_{MV}^\intercal & \widehat{\Sigma}_V \end{pmatrix} := \frac{1}{n} \sum_{k=1}^{n} \begin{pmatrix} \hat{M}_k \hat{M}_k^\intercal - \hat{\mu}\hat{\mu}^\intercal & \hat{M}_k \hat{V}_k^\intercal - \hat{\mu}\hat{\nu}^\intercal \\ \hat{V}_k \hat{M}_k^\intercal - \hat{\nu}\hat{\mu}^\intercal & V_k V_k^\intercal - \hat{\nu}\hat{\nu}^\intercal \end{pmatrix}.$$

With this, we may now present our construction of confidence intervals for the edge weights. (For simplicity, all proofs in this section assume the variables to have mean zero.)

4.1 Confidence Region for the Causal Tree

We use the delta method to estimate the variances of the $\hat{w}_{ji}^{\mathrm{G}}$, and a simple Bonferroni correction to ensure simultaneous coverage of the confidence intervals we develop. Writing z_α for the upper $\alpha/\{2p(p-1)\}$ quantile of a standard normal distribution, we set

$$\hat{u}_{ji},\, \hat{l}_{ji} := \frac{1}{2} \log\left(\frac{\hat{\mu}_{ji}}{\hat{\nu}_i} \right) \pm z_\alpha \frac{\hat{\sigma}_{ji}}{2\sqrt{n}} = \hat{w}_{ji}^{\mathrm{G}} \pm z_\alpha \frac{\hat{\sigma}_{ji}}{2\sqrt{n}}, \tag{10}$$

where

$$\hat{\sigma}_{ji}^2 := \frac{\widehat{\Sigma}_{M,ji,ji}}{\hat{\mu}_{ji}^2} + \frac{\widehat{\Sigma}_{V,i,i}}{\hat{\nu}_i^2} - 2\frac{\widehat{\Sigma}_{MV,ji,i}}{\hat{\mu}_{ji}\hat{\nu}_i}.$$

We treat $[\hat{l}_{ji}, \hat{u}_{ji}]$ as a confidence interval for the true edge weight w_{ji}^{G} and define the following region of directed trees formed of minimizers of the score with edge weights in the confidence hyperrectangle:

$$\hat{C}_{\mathrm{Bon}} := \hat{C}\left(\hat{l}, \hat{u}\right) := \left\{ \operatorname*{arg\,min}_{\tilde{\mathcal{G}}=(V,\tilde{\mathcal{E}}) \in \mathcal{T}_p} \sum_{(j \to i) \in \tilde{\mathcal{E}}} w'_{ji}, : \forall j \neq i, w'_{ji} \in [\hat{l}_{ji}, \hat{u}_{ji}] \right\}.$$

We have the following coverage guarantee for $\hat{C}_{\mathrm{Bon}}$.

Theorem 10 (Confidence region) *Suppose the following conditions hold:*

(i) there exists $\xi > 0$ such that $\mathbb{E}\|X\|^{4+\xi} < \infty$;

(ii) there exists $\xi > 0$ such that for all $j \neq i$, $\mathbb{E}[|\hat{\varphi}_{ji}(X_j) - \varphi_{ji}(X_j)|^{4+\xi} | \tilde{\mathbf{X}}_n] = O_p(1)$;

(iii) $\mathrm{Var}((\hat{M}_1^\intercal, \hat{V}_1^\intercal)^\intercal | \tilde{\mathbf{X}}_n) \xrightarrow{P}_n \Sigma$, *where* Σ *is constant with strictly positive diagonal;*

(iv) for $(j \to i) \in \mathcal{E}$, $\sqrt{n}\mathbb{E}[(\hat{\varphi}_{ji}(X_{k,j}) - \varphi_{ji}(X_{k,j}))^2 | \tilde{\mathbf{X}}_n] \xrightarrow{P}_n 0$.

Then

$$\liminf_{n\to\infty} P\left(\mathcal{G} \in \hat{C}_{\mathrm{Bon}}\right) \geq 1 - \alpha.$$

The second condition requires little more than 4th moments for the absolute errors in the regression (they do not need to converge to zero). Condition (iv) requires that the mean squared prediction errors corresponding to the true causal edges decay faster than a relatively slow $1/\sqrt{n}$ rate. If the causal graph is unidentifiable, then when (iv) holds for all edges corresponding to population score minimizing graphs, $\hat{C}_{\mathrm{Bon}}$ covers every such graph with a probability of at least $1 - \alpha$.

4.2 Testing of Substructures

Whilst the confidence region $\hat{C}_{\mathrm{Bon}}$ has attractive coverage properties, it will typically not be possible to compute it in practice (due to the ranges of w'_{ji} one would need to try). We now introduce two computationally feasible schemes for querying whether $\hat{C}_{\mathrm{Bon}}$ satisfies certain constraints such as containing or not containing a given substructure. More precisely, we propose a conservative exact query scheme called CheckC (for 'check confidence region'), and an asymptotically valid query scheme called ConvB (for 'converging bounds'), which we will see in the simulation experiments is less conservative. The ConvB test gains power at the expense of generality. While the CheckC test works in both the identified and the non-identified setup, the ConvB test needs both identifiability and stronger assumptions in order to hold level.

The idea is as follows: by Theorem 10 the confidence region for the causal graph $\hat{C}_{\mathrm{Bon}}$ contains the causal graph with probability tending to at least $1 - \alpha$. Thus, if we can verify that no graph in $\hat{C}_{\mathrm{Bon}}$ contains a certain substructure, we are able to test the hypothesis that the causal graph satisfies said substructure with asymptotically valid $1-\alpha$ level control.

4.2.1 Substructure Hypotheses

A substructure restriction $\mathcal{R} = (\mathcal{E}_{\mathcal{R}}, \mathcal{E}_{\mathcal{R}}^{\mathrm{miss}}, r)$ on the nodes V contains specified sets $\mathcal{E}_{\mathcal{R}}$ of existing edges, $\mathcal{E}_{\mathcal{R}}^{\mathrm{miss}}$ of missing edges, and a specific root node r (any of such restrictions may be void, too). For example, a substructure restriction could be that a single edge is present (such as $X_1 \to X_2$), or that a single edge is not present (such as $X_1 \not\to X_2$); the restriction can also specify a directed tree. Our approach allows us to conclude that at least one of the constraints in $\mathcal{R}$ does *not* hold for the true graph $\mathcal{G} = (V, \mathcal{E})$. More precisely, we propose a test for the null hypothesis

$$\mathcal{H}_0(\mathcal{R}) : \mathcal{E}_{\mathcal{R}} \setminus \mathcal{E} = \emptyset,\ \mathcal{E} \setminus \mathcal{E}_{\mathcal{R}}^{\mathrm{miss}} = \mathcal{E},\ r = \mathrm{rt}(\mathcal{G}),$$

i.e, that all constraints in a substructure restriction $\mathcal{R}$ are satisfied in the causal graph. We henceforth assume that a proposed substructure $\mathcal{R}$ has no internal inconsistencies, i.e., that there exists at least one directed tree over the nodes V satisfying all conditions of $\mathcal{H}_0(\mathcal{R})$. Example 1 illustrates how substructure restrictions allow us to test various hypotheses about the causal graph.

Example 1 *In Figure 1 we illustrate a true causal graph and five examples of substructure hypotheses that we can test.*

- *Hypothesis 1 (true) consists of the restriction* $\mathcal{R} = \mathcal{E}_{\mathcal{R}}$*, where* $\mathcal{E}_{\mathcal{R}} := \{(X_4 \to X_5)\}$*. This substructure restriction specifies that* $(X_4 \to X_5)$ *is present in the causal graph.*
- *Hypothesis 2 (false) consists of the restriction* $\mathcal{R} = \mathcal{E}_{\mathcal{R}}^{\text{miss}}$*, where* $\mathcal{E}_{\mathcal{R}}^{\text{miss}} := \{(X_6 \to X_3)\}$*. This restriction specifies that* $(X_6 \to X_3)$ *is not in the causal graph.*
- *Hypothesis 3 (true) consists of the restriction* $\mathcal{R} := (\mathcal{E}_{\mathcal{R}}, \mathcal{E}_{\mathcal{R}}^{\text{miss}})$ *with multiple present edges and a single missing edge. Here, the substructure restriction specifies that all edges in* $\mathcal{E}_{\mathcal{R}} := \{(X_3 \to X_2), (X_4 \to X_5), (X_4 \to X_7), (X_6 \to X_3)\}$ *are present, and that the edge in* $\mathcal{E}_{\mathcal{R}}^{\text{miss}} := \{(X_8 \to X_9)\}$ *is not present in the causal graph.*
- *Hypothesis 4 (false) consists of the restriction* $\mathcal{R} := (\mathcal{E}_{\mathcal{R}}, \mathcal{E}_{\mathcal{R}}^{\text{miss}})$ *with multiple present edges and multiple missing edges. This substructure restriction specifies that all edges in* $\mathcal{E}_{\mathcal{R}} := \{(X_1 \to X_2), (X_1 \to X_4), (X_5 \to X_5)\}$ *are present, and that all edges in* $\mathcal{E}_{\mathcal{R}}^{\text{miss}} := \{(X_3 \to X_6), (X_8 \to X_7)\}$ *are not present in the causal graph.*
- *Hypothesis 5 (false) contains the substructure* $\mathcal{R} := \mathcal{E}_{\mathcal{R}}$ *with multiple present edges, specifying that all edges in* $\mathcal{E}_{\mathcal{R}} := \{(X_1 \to X_2), (X_2 \to X_3), (X_1 \to X_4), (X_4 \to X_5), (X_4 \to X_7), (X_5 \to X_6), (X_5 \to X_8), (X_6 \to X_9)\}$ *are present in the causal graph. This substructure restriction uniquely specifies a specific complete directed tree.*

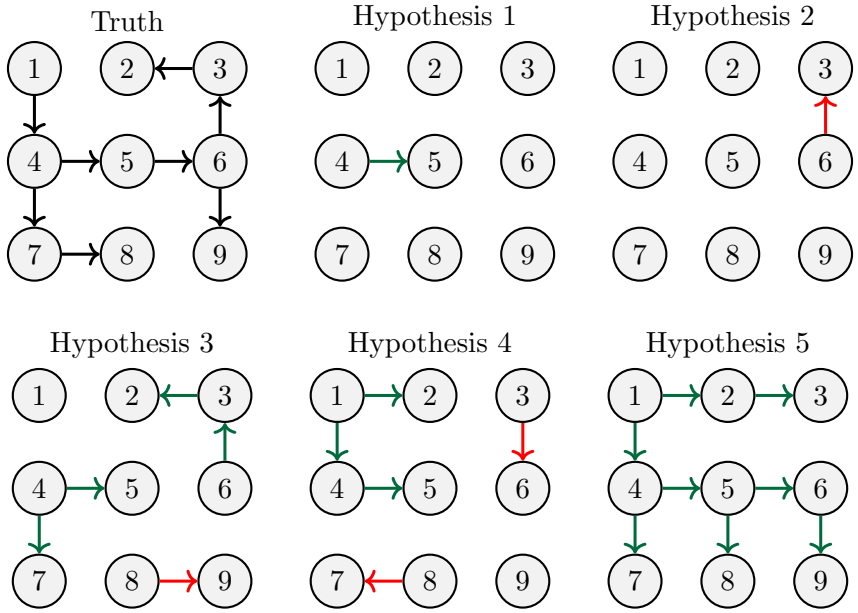

Figure 1: Illustration of six graphs, see Example 1. Colored edges represent testing the presence of edges (green, $\mathcal{E}_{\mathcal{R}}$) or whether edges are missing (red, $\mathcal{E}_{\mathcal{R}}^{\text{miss}}$).

4.2.2 Checking the Confidence Region

In order to present the first method, we introduce some notation. For any non-empty subset of directed trees $\mathcal{T} \subset \mathcal{T}_p$, let $S_{\mathcal{T}}(w)$ be the score attained by the minimum edge weight directed tree recovered by Chu–Liu–Edmonds' algorithm with input edge weights $w := (w_{ji})_{j \neq i}$, when restricting the search to all directed trees in $\mathcal{T}$. That is, if we denote the minimum edge weight directed spanning tree (MWDST) as recovered by Chu–Liu–Edmonds' algorithm, when searching over all directed trees in $\mathcal{T}$ by

$$\mathcal{G}^*_{\mathcal{T}}(w) := \operatorname*{arg\,min}_{\tilde{\mathcal{G}}=(V,\tilde{\mathcal{E}})\in\mathcal{T}} \sum_{(j\to i)\in\tilde{\mathcal{E}}} w_{ji}, \tag{11}$$

then with $\mathcal{G}^*_{\mathcal{T}}(w) = (V, \mathcal{E}^*_{\mathcal{T}}(w))$ the associated score is given by

$$S_{\mathcal{T}}(w) := \sum_{(j\to i)\in\mathcal{E}^*_{\mathcal{T}}(w)} w_{ji}. \tag{12}$$

Now let $\mathcal{T}_p(\mathcal{R}) \subset \mathcal{T}_p$ be the set of all directed trees satisfying the substructure restriction $\mathcal{R}$ and suppose that the causal directed tree $\mathcal{G}$ satisfies $\mathcal{R}$, i.e., $\mathcal{G} \in \mathcal{T}_p(\mathcal{R})$. Hence with probability tending to at least $1-\alpha$ we know that there exists a graph in $\hat{C}_{\text{Bon}}$ satisfying the substructure restriction $\mathcal{R}$. That is, there exist edge weights $w' = (w'_{ji})_{j\neq i}$, with $\hat{l}_{ji} \leq w' \leq \hat{u}_{ji}$ for all $j \neq i$, such that $\mathcal{G}^*_{\mathcal{T}_p}(w')$ satisfies the substructure restriction $\mathcal{R}$. Hence, it must hold that $S_{\mathcal{T}_p(\mathcal{R})}(w') = S_{\mathcal{T}_p}(w')$. Since the score function is weakly monotone, we have, with probability tending to at least $1-\alpha$, that

$$S_{\mathcal{T}_p(\mathcal{R})}(\hat{l}) \leq S_{\mathcal{T}_p(\mathcal{R})}(w') = S_{\mathcal{T}_p}(w') \leq S_{\mathcal{T}_p}(\hat{u}).$$

On the other hand, if $S_{\mathcal{T}_p(\mathcal{R})}(\hat{l}) > S_{\mathcal{T}_p}(\hat{u})$, then we know for certain that $\hat{C}_{\text{Bon}}$ does not contain any graph satisfying the substructure restriction $\mathcal{R}$. We thus define our CheckC test function as

$$\psi^{\text{CheckC}}_{\mathcal{R}} := \begin{cases} 0 & \text{if } S_{\mathcal{T}_p(\mathcal{R})}(\hat{l}) \leq S_{\mathcal{T}_p}(\hat{u}) \\ 1 & \text{otherwise.} \end{cases} \tag{13}$$

Recall that Chu–Liu–Edmonds' algorithm recovers a minimum edge weight directed spanning tree subgraph of a connected graph $\mathcal{H}$. We can construct a specific connected graph $\mathcal{H}$ for which the set of directed spanning tree subgraphs coincides with $\mathcal{T}_p(\mathcal{R})$. In pseudo-algorithm of Algorithm 2 we detail how to test substructure hypotheses with CheckC test.

This testing procedure is conservative as seen by the simulation experiments in Section 6.3. While Theorem 11 proves that hypothesis testing using the CheckC test achieves pointwise asymptotic level, the simulation experiments show that the finite sample power of the test is low for small to moderately large sample sizes. For example, if $\max\{\hat{l}_{ji} : j \neq i\} \leq \min\{\hat{u}_{ji} : j \neq i\}$, then no false substructure hypothesis can be rejected. In Section 4.2.3 we propose an alternative test which exhibits improved finite sample power.

4.2.3 Converging Bounds

We now present the ConvB test which is based on an asymptotically valid query scheme, that is, with probability increasing to one (in the large sample limit) it makes a valid choice on

Algorithm 2 Hypothesis testing of $\mathcal{H}_0(\mathcal{R})$ using the CheckC test

procedure CheckC($\mathcal{R} = (\mathcal{E}_\mathcal{R}, \mathcal{E}_\mathcal{R}^{\text{miss}}, r)$, $\hat{l} = (\hat{l}_{ji})_{j\neq i}$, $\hat{u} = (\hat{u}_{ji})_{j\neq i}$)
  Initialize fully connected graph $\mathcal{H} := \{(j \to i) : i, j \in V, j \neq i\}$.
  For each $(j \to i) \in \mathcal{E}_\mathcal{R}$, delete from $\mathcal{H}$ the edges $\{(k \to i) : k \in V \setminus \{j\}\} \cup \{i \to j\}$.
  For each $(j \to i) \in \mathcal{E}_\mathcal{R}^{\text{miss}}$, delete from $\mathcal{H}$ the edge $(j \to i)$.
  If root $r \in \mathcal{R}$, delete from $\mathcal{H}$ the edges $\{(j \to r) : j \in V\}$.
  Apply Chu–Liu–Edmonds' algorithm to find $S_{\mathcal{T}_p(\mathcal{R})}(\hat{l})$ and $\mathcal{G}^*_{\mathcal{T}_p(\mathcal{R})}(\hat{l})$, the minimum $\hat{u}$-weighted directed spanning subtree of $\mathcal{H}$.
  Apply Chu–Liu–Edmonds' algorithm to find $S_{\mathcal{T}_p}(\hat{u})$ and $\mathcal{G}^*_{\mathcal{T}_p}(\hat{u})$, the minimum $\hat{l}$-weighted directed spanning subtree of the fully connected graph.
  If $S_{\mathcal{T}_p(\mathcal{R})}(\hat{l}) \leq S_{\mathcal{T}_p}(\hat{u})$, then set $\psi_\mathcal{R}^{\text{CheckC}} := 0$, otherwise set $\psi_\mathcal{R}^{\text{CheckC}} := 1$.
  return $\psi_\mathcal{R}^{\text{CheckC}}$.
end procedure

whether any graph in $\hat{C}_{\text{Bon}}$ satisfies a substructure restriction $\mathcal{R}$. We call this test ConvB for 'converging bounds' because it requires that all lower edge weight bounds converge towards the Gaussian population edge weights. Consider a true null hypothesis $H_0(\mathcal{R})$, i.e., a substructure restriction $\mathcal{R}$ which is satisfied by the causal graph $\mathcal{G}$. Suppose that $\mathcal{G} \in \hat{C}_{\text{Bon}}$, which implies the existence of edge weights $w' = (w'_{ji})_{j\neq i}$, with $\hat{l}_{ji} \leq w'_{ji} \leq \hat{u}_{ji}$ for all $j \neq i$, such that the minimum edge weight directed spanning tree,

$$\mathcal{G}^*_{\mathcal{T}_p}(w') := \underset{\tilde{\mathcal{G}}=(V,\tilde{\mathcal{E}})\in\mathcal{T}_p}{\arg\min} \sum_{(j\to i)\in\tilde{\mathcal{E}}} w'_{ji},$$

satisfies the restrictions $\mathcal{R}$. The intuition for our approach is as follows: We propose a method that 'helps' all edge weights that are not in direct disagreement with $\mathcal{R}$, and 'penalizes' all edge weights that are in disagreement with $\mathcal{R}$, more precisely, we define the edge weights $\check{w} = (\check{w}_{ji})_{j\neq i}$ by

$$\check{w}_{ji} = \begin{cases} \hat{u}_{ji} & \text{if } [\exists k \neq j : (k \to i) \in \mathcal{E}_\mathcal{R}] \vee [(i \to j) \in \mathcal{E}_\mathcal{R}] \vee [(j \to i) \in \mathcal{E}_\mathcal{R}^{\text{miss}}] \vee [i = r], \\ \hat{l}_{ji} & \text{otherwise,} \end{cases}.$$

We can then expect that $\mathcal{G}^*_{\mathcal{T}_p}(\check{w})$ still satisfies the restriction $\mathcal{R}$ (with probability tending to one, see Theorem 11). Conversely, the probability that $\mathcal{G}^*_{\mathcal{T}_p}(\check{w})$ does not satisfy the restriction $\mathcal{R}$ is, in the large sample limit, bounded by the probability that $\mathcal{G}$ is not in the confidence region $\hat{C}_{\text{Bon}}$. We may set our test function

$$\psi_\mathcal{R}^{\text{ConvB}} = \begin{cases} 0, & \text{if } \mathcal{G}^*_{\mathcal{T}_p}(\check{w}) \text{ satisfies } \mathcal{R} \\ 1, & \text{otherwise.} \end{cases}$$

The pseudo-algorithm in Algorithm 3 details how to test any substructure hypothesis $\mathcal{H}_0(\mathcal{R})$ using the asymptotic query scheme of the ConvB test.

Our GitHub repository (see Footnote 8) contains R implementations of both testing procedures. The following theorem shows that both substructure hypothesis tests achieve pointwise asymptotic level. Any number of null hypotheses may be tested simultaneously,

Algorithm 3 Hypothesis testing of $\mathcal{H}_0(\mathcal{R})$ using the ConvB test

procedure ConvB($\mathcal{R} = (\mathcal{E}_\mathcal{R}, \mathcal{E}^{\text{miss}}_\mathcal{R}, r)$, $\hat{l} = (\hat{l}_{ji})_{j \neq i}$, $\hat{u} = (\hat{u}_{ji})_{j \neq i}$)
  Initialize $\check{w} := \hat{l}$.
  For each $(j \to i) \in \mathcal{E}_\mathcal{R}$ and all $k \in V \setminus \{j\}$, set $\check{w}_{ki} := \hat{u}_{ki}$.
  For each $(j \to i) \in \mathcal{E}^{\text{miss}}_\mathcal{R}$, set $\check{w}_{ji} := \hat{u}_{ji}$.
  If root $r \in \mathcal{R}$, then for all $j \in V$, set $\check{w}_{jr} := \hat{u}_{lr}$.
  Apply Chu–Liu–Edmonds' algorithm to find $\mathcal{G}^*_{\mathcal{T}_p}(\check{w})$.
  If $\mathcal{G}^*_{\mathcal{T}_p}(\check{w})$ satisfies $\mathcal{R}$, then set $\psi^{\text{ConvB}}_\mathcal{R} := 0$, otherwise set $\psi^{\text{ConvB}}_\mathcal{R} := 1$.
  return $\psi^{\text{ConvB}}_\mathcal{R}$.
end procedure

without the need for any multiple testing correction. This is because the tests may be viewed as simply querying the properties of the single confidence region of Theorem 10, which has coverage of the truth with probability at least $1 - \alpha$.

Theorem 11 (Pointwise asymptotic level) *Let $\alpha \in (0, 1)$ and let $\mathcal{R}_1, \mathcal{R}_2, \ldots$ be any collection of potentially data-dependent substructure restrictions. Suppose that conditions of Theorem 10 are satisfied. If either*

(a) $\psi_{\mathcal{R}_k} = \psi^{\text{CheckC}}_{\mathcal{R}_k}$ *for all* $k \geq 1$, *or*

(b) $\psi_{\mathcal{R}_k} = \psi^{\text{ConvB}}_{\mathcal{R}_k}$ *for all* $k \geq 1$, *Assumption 1 holds, and for all* $(j \to i) \notin \mathcal{E}$ *it holds that* $\sqrt{n}\mathbb{E}[(\hat{\varphi}_{ji}(X_{k,j}) - \varphi_{ji}(X_{k,j}))^2 | \tilde{\mathbf{X}}_n] \xrightarrow{P} 0$,

then it holds that

$$\limsup_{n \to \infty} P\left(\bigcup_{k \,:\, \mathcal{H}_0(\mathcal{R}_k) \text{ is true}} (\psi_{\mathcal{R}_k} = 1) \right) \leq \alpha.$$

The ConvB test requires stronger conditions than the CheckC test. Additionally to the assumptions made by the CheckC test, it requires identifiability of the causal graph and $\sqrt{n}$-convergence of the mean squared estimation error for the non-causal edges. On the other hand, it would be possible to give uniform asymptotic level guarantees for the CheckC test as it only relies on the coverage properties of confidence intervals for the true weights.

5. Bounding the Identifiability Gap

We have seen in Section 3.3 that the identifiability gap, that is, the smallest score difference between the causal tree $\mathcal{G}$ and any alternative graph $\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}$, plays an important role when identifying causal trees from observational data. It provides information about whether the causal graph is identifiable through the corresponding score function, for example, if we can establish that the smallest Gaussian score gap is strictly positive, i.e.,

$$\min_{\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}} \ell_{\text{G}}(\tilde{\mathcal{G}}) - \ell_{\text{G}}(\mathcal{G}) = \inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p} D_{\text{KL}}(P_X \| Q) > 0, \tag{14}$$

then $\mathcal{G}$ is identified by the Gaussian score function. Lemma 7 lists conditions guaranteeing that Assumption 1, i.e., Equation (14) holds. However, postitivity of the identifiability gap for a single model is not sufficient for uniform consistency or consistency under vanishing identifiability.

For consistency under vanishing identifiability we need to ensure that the identifiability gap vanishes at a slower rate than $1/\sqrt{n}$; see Theorem 9. Similarly, for uniform consistency over a class of causal additive noise models $\Theta \subset \mathcal{T}_p \times \mathcal{M}^p \times \mathcal{P}^p$, one needs the existence of a strictly positive constant $c > 0$ uniformly lower bounding the identifiability gap, i.e.,

$$\inf_{\theta \in \Theta} \min_{\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}} \ell_{\mathrm{G}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{G}}(\mathcal{G}) > c. \tag{15}$$

The identifiability gap is an involved quantity. In this section, we derive a lower bound that is based on local properties of the underlying structural causal models (such as the ability to reverse edges), using information-theoretic quantities.

We first consider the special cases of bivariate models (Section 5.1) and multivariate Markov equivalent trees (Section 5.2) and then turn to general trees (Section 5.3). However, before we venture into the derivation of the specific lower bounds we first examine the connection between the identifiability gaps associated with the different score functions. In this section, we assume that $X \sim P_X$ is generated by a structural causal additive tree model with $\mathbb{E}\|X\|^2 < \infty$ such that the local Gaussian, entropy and conditional entropy scores are well-defined. We neither assume that θ is a restricted structural causal additive model, i.e., $\theta \in \Theta_R$, nor strict positivity of the identifiability gap, i.e., Assumption 1. The following result shows that the local node-wise score gaps associated with the different score functions are ordered.

Lemma 12 *For any $\tilde{\mathcal{G}} \in \mathcal{T}_p$ and for all $i \in V$*

$$\ell_{\mathrm{CE}}(\tilde{\mathcal{G}}, i) - \ell_{\mathrm{CE}}(\mathcal{G}, i) \leq \ell_{\mathrm{E}}(\tilde{\mathcal{G}}, i) - \ell_{\mathrm{E}}(\mathcal{G}, i).$$

If the underlying model is an causal additive tree model with Gaussian noise, then

$$\ell_{\mathrm{E}}(\tilde{\mathcal{G}}, i) - \ell_{\mathrm{E}}(\mathcal{G}, i) \leq \ell_{\mathrm{G}}(\tilde{\mathcal{G}}, i) - \ell_{\mathrm{G}}(\mathcal{G}, i).$$

It follows that the full graph score gaps and identifiability gaps associated with the different score functions satisfy a similar ordering. Thus, given that the underlying model is an causal additive tree model with Gaussian noise, a strictly positive entropy identifiability gap implies that the Gaussian identifiability gap is strictly positive. It is, however, not possible to establish strict positivity of the conditional entropy identifiability gap; see Remark 1 in Appendix B. Therefore, we focus on establishing a lower bound for the entropy identifiability gap that is tighter than that given by the conditional entropy identifiability gap.

In general, we cannot use node-wise comparisons of the scores of two graphs to bound the identifiability gap (the reason is that in general a node receives a better score in a graph, where it has a parent, compared to a graph, where it does not; see Example 3 in Appendix B for a formal argument). We start by analyzing the identifiability gap in models with two variables.

5.1 Bivariate Models

We now consider two nodes $V = \{X, Y\}$, and graphs $\mathcal{T}_2 = \{(X \to Y), (Y \to X)\}$. Without loss of generality assume that $(X, Y) \in \mathcal{L}^2(P)$ is generated by an additive noise SCM $\theta = (\mathcal{G}, (f_i), P_N)$ with causal graph $\mathcal{G} = (X \to Y) \in \mathcal{T}_2$ to which the only alternative graph is $\tilde{\mathcal{G}} = (Y \to X)$. That is,

$$X := N_X, \quad Y := f(X) + N_Y, \tag{16}$$

where $(N_X, N_Y) \sim P_N \in \mathcal{P}^2$. The bivariate entropy identifiability gap, which we will later refer to as the edge reversal entropy score gap, is defined as

$$\begin{aligned}\Delta\ell_{\mathrm{E}}(X \overset{\leftarrow\!\!-\!\!\rightarrow}{} Y) :&= \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) \\ &= h(Y) + h(X - \mathbb{E}[X|Y]) - h(X) - h(Y - \mathbb{E}[Y|X]),\end{aligned}$$

where the fully drawn arrow symbolizes the true causal relationship and the dashed arrow the alternative. The following lemma simplifies the bivariate entropy identifiability gap to a single mutual information between the effect and the residual of the minimum mean squared prediction error regression of cause on the effect.

Lemma 13 *Consider the bivariate setup of Equation* (16) *and assume that $f(X)$ has density. It holds that*

$$\Delta\ell_{\mathrm{E}}(X \overset{\leftarrow\!\!-\!\!\rightarrow}{} Y) = I(X - \mathbb{E}[X|Y]; Y) \geq 0.$$

Thus, the causal graph is identified in a bivariate setting if one maintains dependence between the predictor and minimum mean squared error regression residual in the anti-causal direction. This result is in accordance with the previous identifiability results. For example, in the linear additive Gaussian noise case, $I(X - \mathbb{E}[X|Y]; Y) = 0$. Consequently, the causal graph is not identified from the entropy score function.

Whenever the conditional mean in the anti-causal direction vanishes, e.g., with symmetric causal function and symmetric noise distribution, it is possible to derive a more explicit lower bound with more intuitive sufficient conditions for identifiability of the causal graph.

Proposition 14 *Consider the bivariate setup of Equation* (16) *and assume that $f(X)$ has density. If the reversed direction conditional mean $\mathbb{E}[X|Y]$ almost surely vanishes (e.g., because f, X and N_Y are symmetric), then*

$$\Delta\ell_{\mathrm{E}}(X \overset{\leftarrow\!\!-\!\!\rightarrow}{} Y) = I(X; f(X) + N_Y),$$

which is strictly positive if and only if $X \not\perp\!\!\!\perp f(X) + N_Y$. In addition, we have the following statements.

(a) Let $f(X)^{\mathrm{G}}$ and N_Y^{G} be independently normally distributed with the same mean and variance as $f(X)$ and N_Y, respectively. If $D_{\mathrm{KL}}(f(X) \| f(X)^{\mathrm{G}}) \leq D_{\mathrm{KL}}(N_Y \| N_Y^{\mathrm{G}})$, then

$$\Delta\ell_{\mathrm{E}}(X \overset{\leftarrow\!\!-\!\!\rightarrow}{} Y) \geq \frac{1}{2} \log\left(1 + \frac{\mathrm{Var}(f(X))}{\mathrm{Var}(N_Y)}\right).$$

(b) If the density of $f(X) + N_Y$ is log-concave, then

$$\Delta\ell_{\mathrm{E}}(X \overset{\leftarrow\!-\!-}{\longrightarrow} Y) \geq \frac{1}{2}\log\left(\frac{2}{\pi e} + \frac{2}{\pi e}\frac{\mathrm{Var}(f(X))}{\mathrm{Var}(N_Y)}\right).$$

This lower bound is non-trivial only if $\mathrm{Var}(f(X)) > (\pi e/2{-}1)\mathrm{Var}(N_Y) \approx 3.27\mathrm{Var}(N_Y)$.

Thus, if the conditional mean $\mathbb{E}[X|Y]$ in the anti-causal direction vanishes, then under certain conditions, the causal direction is identified by the entropy score function (as long as $\mathrm{Var}(f(X))$ is sufficiently large relative to $\mathrm{Var}(N_Y)$). The edge reversal score gap for the Gaussian score is given by

$$\begin{aligned}\Delta\ell_{\mathrm{G}}(X \overset{\leftarrow\!-\!-}{\longrightarrow} Y) &:= \frac{1}{2}\log\left(\frac{\mathrm{Var}(X - \mathbb{E}[X|Y])}{\mathrm{Var}(X)}\right) - \frac{1}{2}\log\left(\frac{\mathrm{Var}(Y - \mathbb{E}[Y|X])}{\mathrm{Var}(Y)}\right) \\ &= \frac{1}{2}\log\left(\frac{\mathrm{Var}(X - \mathbb{E}[X|Y])}{\mathrm{Var}(X)}\right) + \frac{1}{2}\log\left(1 + \frac{\mathrm{Var}(f(X))}{\mathrm{Var}(N_Y)}\right),\end{aligned}$$

which reduces to the lower bound in point (a) of Proposition 14 if the conditional mean $\mathbb{E}[X|Y]$ in the anti-causal direction vanishes.

Example 2 *Consider the bivariate setup of Equation* (16)*. Suppose that the causal function f is a quadratic function $f(x) = \alpha x^2 + \beta$ for some $\alpha, \beta, \in \mathbb{R}$ and that $N_X \sim \mathcal{N}(0, \sigma_X^2)$ and $N_Y \sim \mathcal{N}(0, \sigma_Y^2)$. It holds that $E[X|Y]$ vanishes, and the bivariate Gaussian identifiability gap reduces to*

$$\Delta\ell_{\mathrm{G}}(X \overset{\leftarrow\!-\!-}{\longrightarrow} Y) = \frac{1}{2}\log\left(1 + 2\alpha^2\frac{\sigma_X^4}{\sigma_Y^2}\right).$$

5.2 Multivariate Markov Equivalent Trees

Two Markov equivalent trees differ in precisely one directed path that is reversed in one graph relative to the other.[9] The entropy score gap of Markov equivalent trees therefore reduces to the binary case.

Proposition 15 *Consider any $\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}$ that is Markov equivalent to the causal tree $\mathcal{G}$. Let $c_1 \to \cdots \to c_r$ be the unique directed path in $\mathcal{G}$ that is reversed in $\tilde{\mathcal{G}}$. Then*

$$\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) = \sum_{i=1}^{r-1}\Delta\ell_{\mathrm{E}}(c_i \overset{\leftarrow\!-\!-}{\longrightarrow} c_{i+1}) \geq \min_{1\leq i\leq r-1}\Delta\ell_{\mathrm{E}}(c_i \overset{\leftarrow\!-\!-}{\longrightarrow} c_{i+1}).$$

Thus, a lower bound of the entropy score gap that holds uniformly over the Markov equivalence class is given by the smallest possible edge reversal in the causal directed graph:

$$\min_{\tilde{\mathcal{G}}\in\mathrm{MEC}(\mathcal{G})\setminus\{\mathcal{G}\}}\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) \geq \min_{(j\to i)\in\mathcal{E}}\Delta\ell_{\mathrm{E}}(j \overset{\leftarrow\!-\!-}{\longrightarrow} i).$$

9. To see this, note that any two directed trees are Markov equivalent if and only if they satisfy the exact same d-separations or equivalently they share the same skeleton (there are no v-structures in directed trees). Distinct directed trees sharing the same skeleton must have distinct root nodes. Consequently, there exist exactly one directed path in $\mathcal{G}$ from $\mathrm{rt}(\mathcal{G})$ to $\mathrm{rt}(\tilde{\mathcal{G}})$ that is reversed in $\tilde{\mathcal{G}}$; see Lemma D.6

5.3 General Multivariate Trees

We now derive a lower bound of the entropy identifiability gap, i.e., a lower bound of the entropy score gap that holds uniformly over all alternative trees $\mathcal{T}_p \setminus \{\mathcal{G}\}$. To do so, we exploit a graph reduction technique (introduced by Peters et al., 2014) which enables us to reduce the analysis to three distinct scenarios. This graph reduction works as follows. Fix any alternative graph $\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}$, and iteratively remove any node (from both $\mathcal{G}$ and $\tilde{\mathcal{G}}$) that has no children and the same parents in both $\mathcal{G}$ and $\tilde{\mathcal{G}}$. The score gap is unaffected by the graph reduction.[10]

Applying this iteration scheme, until no such node can be found, results in two reduced graphs $\mathcal{G}_R = (V_R, \mathcal{E}_R)$ and $\tilde{\mathcal{G}}_R = (V_R, \tilde{\mathcal{E}}_R)$. These reduced graphs cannot be empty, for that would only happen if $\tilde{\mathcal{G}} = \mathcal{G}$. Further, they have identical vertices but different edges. And they can be categorized into one of three cases. To do so, consider a node L that is a sink node, i.e., a node without children, in $\mathcal{G}_R$ and consider its parent in $\mathcal{G}_R$. Now, considering $\tilde{\mathcal{G}}_R$, one of the following conditions must hold: the parent is also a parent of L in $\tilde{\mathcal{G}}_R$ (we then call it Z), the parent is not connected to L in $\tilde{\mathcal{G}}_R$ (we then call it W), or the parent is a child of L in $\tilde{\mathcal{G}}_R$ (we then call it Y). Figure 2 visualizes these three scenarios.

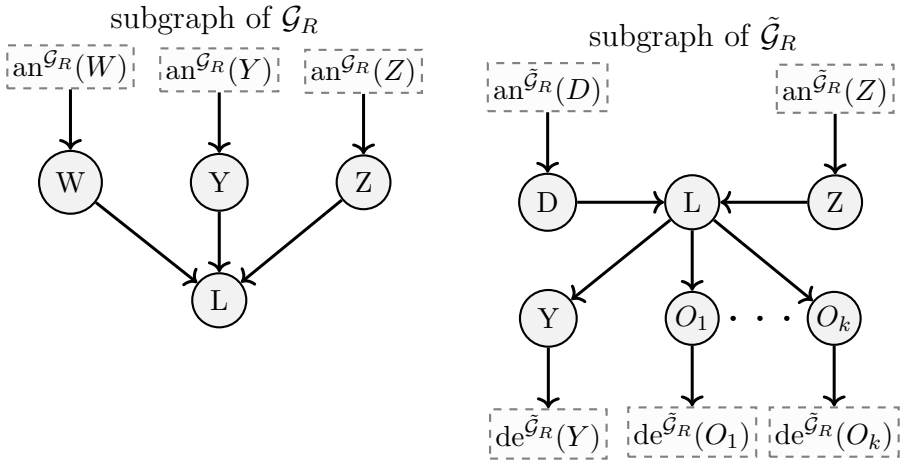

Figure 2: Schematic illustration of parts of two reduced graphs produced by the graph reduction technique described in Section 5.3. Consider a sink node L in $\mathcal{G}_R$. Its parent (in $\mathcal{G}_R$) must either be a parent in $\tilde{\mathcal{G}}_R$, too, it must be a child in $\tilde{\mathcal{G}}_R$, or it is unconnected to L in $\tilde{\mathcal{G}}_R$. Thus, exactly one of the sets Z, Y, and W is non-empty. This case distinction is used to compute the three bounds in Theorem 16. D, $O_1, \ldots, O_k$ denote further (possibly existing) nodes in $\tilde{\mathcal{G}}_R$.

10. All removed nodes $V \setminus V_R$ have identical incoming edges in both graphs and therefore have identical local scores. That is, for any loss function $l \in \{\ell_{\mathrm{CE}}, \ell_{\mathrm{E}}, \ell_{\mathrm{G}}\}$ we have that $l(\tilde{\mathcal{G}}) - \ell(\mathcal{G}) = \sum_{i \in V_R} \ell(\tilde{\mathcal{G}}, i) - \ell(\mathcal{G}, i) + \sum_{i \in V \setminus V_R} \ell(\tilde{\mathcal{G}}, i) - \ell(\mathcal{G}, i) = \sum_{i \in V_R} \ell(\tilde{\mathcal{G}}, i) - \ell(\mathcal{G}, i) = \ell(\tilde{\mathcal{G}}_R) - l(\mathcal{G}_R)$.

We can now obtain bounds for each of the three case individually. For the case with a node Z (a 'staying parent'), define

$$\Pi_Z(\mathcal{G}) := \left\{(z,l,o) \in V^3 \text{ s.t. } (z \to l) \in \mathcal{E} \text{ and } o \in \mathrm{nd}^{\mathcal{G}}(l) \setminus \{z,l\}\right\}.$$

The score gap can then be lower bounded by $\min_{(z,l,o)\in\Pi_Z(\mathcal{G})} I(X_z; X_o|X_l)$ (see Lemma D.7). Intuitively, $I(X_z; X_o|X_l)$ quantifies the strength of the connection between z and o, when conditioning on l (which does not lie on the path between z and o). This is a non-local bound in that it does not constrain the length of the path connecting z and o. Analyzing or bounding this term might be difficult. We will see in Section 5.4 that this part is not needed for causal additive tree models with Gaussian noise.

For the case with a node W ('removing parent'), define

$$\Pi_W(\mathcal{G}) := \left\{(w,l,o) \in V^3 \text{ s.t. } (w \to l) \in \mathcal{E} \text{ and } o \in (\mathrm{ch}^{\mathcal{G}}(w) \setminus \{l\}) \cup \mathrm{pa}^{\mathcal{G}}(w)\right\}.$$

This case results in the lower bound $\min_{(w,l,o)\in\Pi_W(\mathcal{G})} I(X_w; X_l|X_o)$ (see Lemma D.8). Here, w is a parent of l and o is directly connected to w. Intuitively, $I(X_w; X_l|X_o)$ quantifies the strength of the edge $w \to l$. We condition on o but that node is not directly connected to l (only via w). For the first two cases, faithfulness (Spirtes et al., 2000) implies that these terms are non-zero and bounding them away from zero reminds of strong faithfulness (Zhang and Spirtes, 2002). However, in the second case, one considers individual edges, which reminds more of a strong version of causal minimality (Spirtes et al., 2000; Peters et al., 2017).

For the case with a node Y ('parent to child'), a lower bound is given by the minimal edge reversal score gap $\min_{(j\to i)\in\mathcal{E}} \Delta\ell_{\mathrm{E}}(j \xleftarrow{\!\!-\!\!\to} i)$ (see Lemma D.9). The term $\Delta\ell_{\mathrm{E}}(j \xleftarrow{\!\!-\!\!\to} i)$ measures the identifiability of the direction of an individual edge. It is zero in the linear additive Gaussian noise case, for example. We provide more details on the reduced graphs and on the arguments in the three cases in Section D.4.2 of Appendix D.

Combining the three bounds from above, we obtain the following theorem.

Theorem 16 *It holds that*

$$\min_{\tilde{\mathcal{G}}\in\mathcal{T}_p\setminus\{\mathcal{G}\}} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) \geq \min\Bigg\{ \min_{(z,l,o)\in\Pi_Z(\mathcal{G})} I(X_z; X_o|X_l), \min_{(w,l,o)\in\Pi_W(\mathcal{G})} I(X_w; X_l|X_o), \min_{(j\to i)\in\mathcal{E}} \Delta\ell_{\mathrm{E}}(j \xleftarrow{\!\!-\!\!\to} i)\Bigg\}. \tag{17}$$

This result lower bounds the identifiability gap using information-theoretic quantities. Corresponding results for the Gaussian score follow immediately by Lemma 12. The last two terms are local properties of the underlying structural causal model; the first term is not. As seen in Section 5.2, the last term on the right-hand side is required when considering only Markov equivalent trees; if it is non-zero, it allows us to orient all edges in the skeleton. The first two terms (non-zero under faithfulness) are additionally required when the considered trees are not Markov equivalent.

We now turn to the case of causal additive tree models with Gaussian noise innovations. Here, the first term is not needed; the bound then depends only on local properties of the structural causal model.

5.4 Gaussian Multivariate Trees

The score gap lower bound in Equation (17) consists of local dependence properties except for the node tuples $\Pi_Z(\mathcal{G})$ (Lemma D.7) that arise when considering alternative graphs that result in reduced graphs with a node Z ('staying parents'). However, we show that for additive Gaussian noise models, the score gap for such alternative graphs can be lower bounded by the score gaps already considered in alternative graphs with a node Y ('parent to child') and a node W ('removing parent'). Thus, we have the following theorem, with a bound consisting only of local properties of the model.

Theorem 17 (Gaussian localization of the identifiability gap) *For causal additive tree models with Gaussian noise, we have that*

$$\min_{\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}} \ell_{\mathrm{G}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{G}}(\mathcal{G}) \geq \min \left\{ \min_{(w,l,o) \in \Pi_W(\mathcal{G})} I(X_w; X_l \,|\, X_o), \min_{(j \to i) \in \mathcal{E}} \Delta \ell_{\mathrm{E}}(j \xleftarrow{\cdot\cdot} i) \right\}.$$

6. Simulation Experiments

In this section, we investigate the finite-sample performance of CAT and perform simulation experiments investigating the identifiability gap and its lower bound. In Section 6.1 we compare the performance of CAT to CAM of Bühlmann et al. (2014) for causal additive tree models with Gaussian and non-Gaussian noise. In Section 6.2 we compare the CAT and CAM for causal discovery on non-tree DAG models (CAT always outputs a directed tree). In Section 6.3 we investigate the finite sample power and level of the proposed hypothesis testing procedures. In Section 6.4 we perform simulation experiments that highlight the behavior of the identifiability gap and its corresponding lower bound derived in Section 5. The code scripts (R) for the simulation experiments, empirical applications and the implementation of CAT and the two testing procedures are available on GitHub (see Footnote 8).

6.1 Causal Structure Learning for Trees

In this section, we compare the performance of the structure learning methods CAT and CAM when employed on additive noise models with causal graphs given by directed trees.

6.1.1 Tree Generation Schemes

We employ two different random directed tree generation schemes: Type 1 (many leaf nodes) and Type 2 (many branch nodes). Figure 3 illustrates two directed trees generated in accordance with the two generation schemes. For more details, see Algorithms 4 and 5 in Section C.1 of Appendix C.

6.1.2 Gaussian Experiment

In this experiment, we generate data similarly to the experimental setup of Bühlmann et al. (2014). For any given directed tree we generate causal functions by sample paths of Gaussian processes with radial basis function (RBF) kernel and bandwidth parameter of one. Sample paths of Gaussian processes with radial basis function kernels are almost

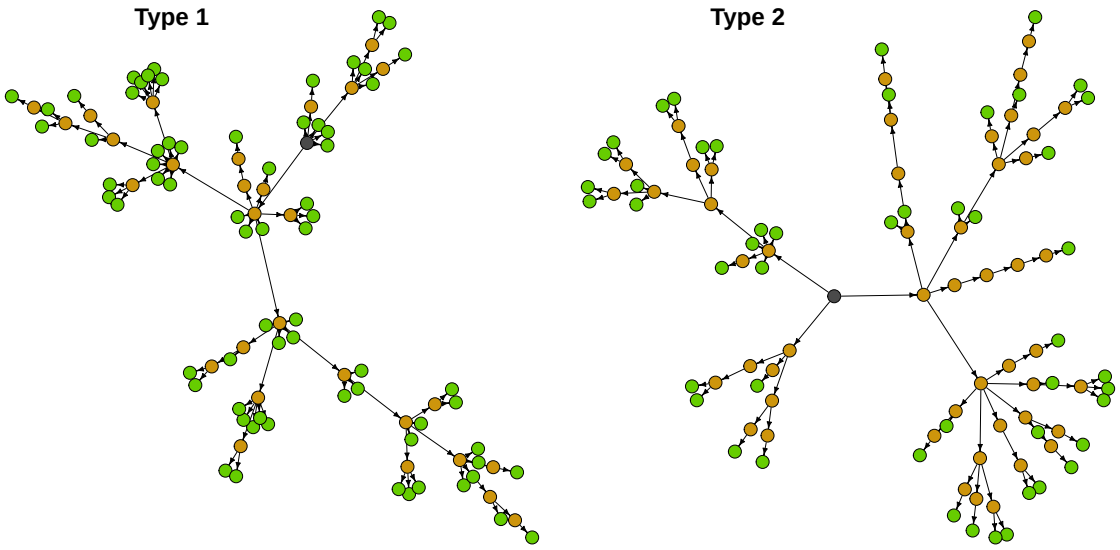

Figure 3: Illustration of Type 1 (many leaf nodes) and Type 2 (many branch nodes) directed trees over $p = 100$ nodes. The green nodes are leaf nodes, the brown nodes are branch nodes, and the black nodes are root nodes. The Type 1 tree contains 70 leaf nodes, while the Type 2 tree only contains 49 leaf nodes.

surely infinitely continuous differentiable (e.g., Kanagawa et al., 2018), non-constant and nonlinear, so they satisfy the requirements of Lemma 3. See Figure 10 in Section C.2 of Appendix C for illustrations of random draws of such functions. Root nodes are mean zero Gaussian variables with standard deviation sampled uniformly on $(1, 2)$. Furthermore, for each fixed tree and set of causal functions, we introduce at each non-root node additive Gaussian noise with mean zero and standard deviation sampled uniformly on $(1/5, \sqrt{2}/5)$.

We first compare our method CAT with Gaussian score function (CAT.G) against the method CAM of Bühlmann et al. (2014) on the previously detailed nonlinear additive Gaussian noise tree setup. We use CAT.G without both cross-fitting and pruning. Note that with cross-fitting the results do not change much but, as expected, cross-fitting yields slightly worse results for small sample sizes (see Figure 11 in Appendix C). We use the R-package `mgcv` (Mixed GAM Computation Vehicle, Wood, 2022) with default settings to construct a thin plate regression spline estimate of the conditional expectations (Wood, 2003). We use the implementation of Chu–Liu–Edmonds' algorithm from the R-package `RBGL`.[11] CAM is employed with a maximum number of parents set to one (restricting the output to directed trees), without preliminary neighborhood selection and subsequent pruning. We measure the performance of the methods by computing the Structural Hamming Distance (SHD, Tsamardinos et al., 2006) and Structural Intervention Distance (SID, Peters and Bühlmann, 2015) to the causal tree.

11. The RBGL implementation finds maximum edge weight directed trees and requires all positive edge weights. As such, we take the negative of our edge weights and shift them all by the absolute value of smallest edge weight. If an edge weight is set to zero this edge can not be chosen.

For each system size $p \in \{16, 32, 64, 128\}$ we generate a causal tree, corresponding causal functions and noise variances and sample $n \in \{50, 100, 200, 500\}$ observations. This is repeated 200 times and the SHD results are summarized in the boxplot of Figure 4.

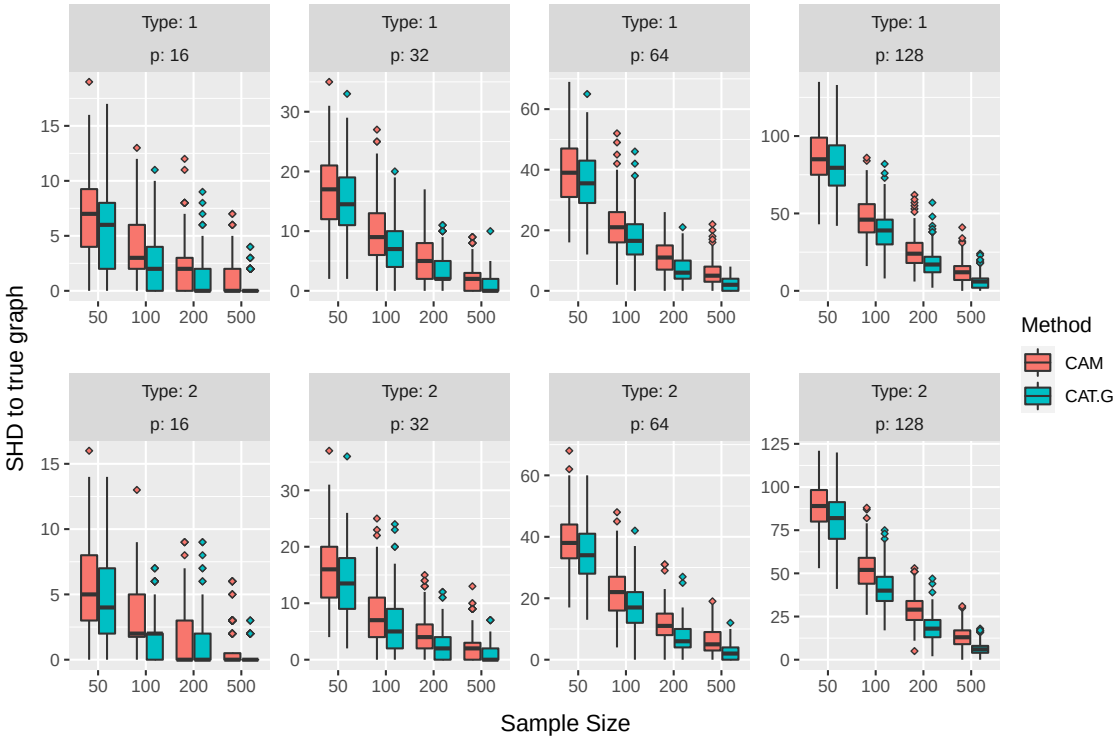

Figure 4: Causal additive tree models with Gaussian noise: Boxplots of the SHD performance of CAM and CAT.G (Gaussian score) for varying sample sizes, system sizes, and tree types. CAT.G outperforms CAM in a wide range of scenarios.

Both methods perform better on trees of Type 2 than on trees of Type 1. CAT.G outperforms CAM in terms of SHD to the true graph both in median distance and IQR length and position for all sample sizes, system sizes and tree types. Considering the SID to the causal tree yields similar conclusions; see Figure 12 in Section C.2 of Appendix C. In their default versions, CAM and CAT.G use different estimation techniques of the conditional expectations, but this does not seem to be the source of the performance difference: Figure 13 in Section C.2 of Appendix C illustrates a similar SHD performance difference when forcing CAT.G to use the edge weights produced by the CAM implementation.

6.1.3 Non-Gaussian Experiment

We now compare the performance of CAM and CAT with Gaussian (CAT.G) and entropy (CAT.E) score functions in a setup with varying noise distributions. The entropy edge weights used by CAT.E are estimated with the differential entropy estimator of Berrett et al. (2019) as implemented in the CRAN R-package `IndepTest` (Berrett et al., 2018). We

use the same simulation setup as in Section 6.1.2 but now we only consider trees of Type 1 and parameterize the setup by $\alpha > 0$, which controls the deviation of the additive noise innovations from a Gaussian distribution. More precisely, we generate the additive noise variables $N_i(\alpha)$ as

$$N_i(\alpha) = \text{sign}(Z_i)|Z_i|^\alpha,$$

where $Z_i \sim \mathcal{N}(0, \sigma_i^2)$ with σ_i sampled uniformly on $(1/5, \sqrt{2}/5)$ or uniformly on $(1, 2)$ if $i = \text{rt}(\mathcal{G})$. For $\alpha = 1$ this yields Gaussian noise, while for alpha $\alpha \neq 1$ the noise is non-Gaussian. We conduct the experiment for all combinations of $\alpha \in \{0.1, 0.2, \ldots, 2, 2.5, 3, 3.5, 4\}$ and sample sizes $n \in \{50, 500\}$ for a fixed system size of $p = 32$. Each setting is repeated 500 times and the results are illustrated in Figure 5.

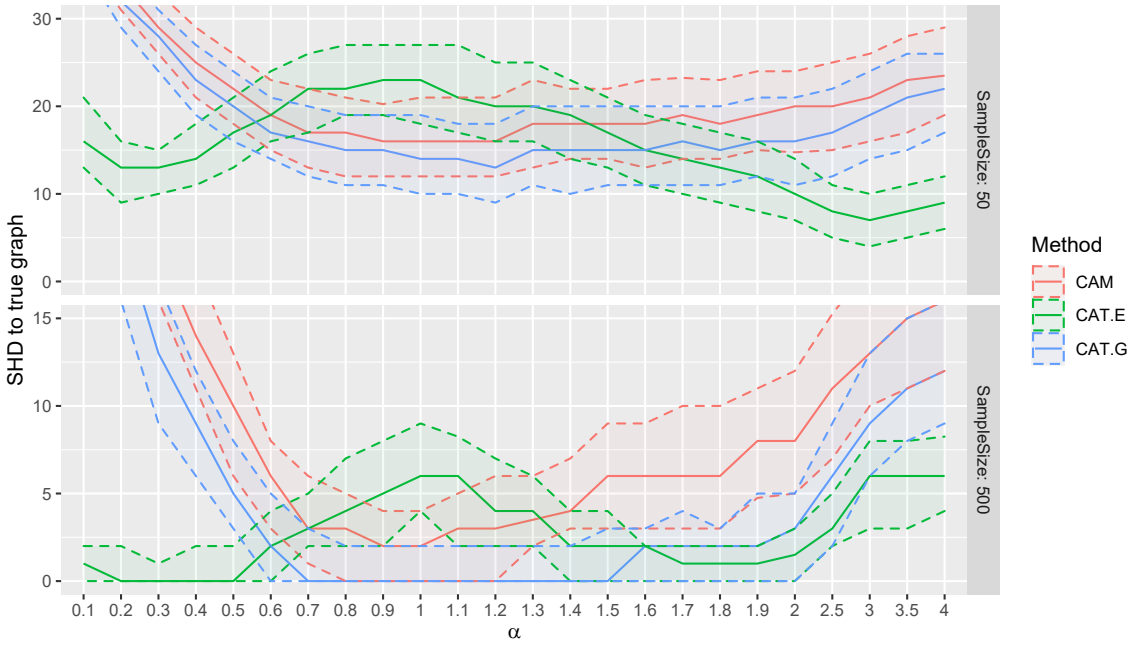

Figure 5: Deviations from Gaussianity: The parameter α controls the noise deviation from the Gaussian distribution. CAT.G and CAT.E are instances of CAT with edge weights derived from Gaussian and entropy score functions, respectively. The solid lines represent the median SHD and the shaded (dashed) region represents the interquartile range. Using the entropy score yields better results for noise distributions that deviate strongly from Gaussian noise.

For Gaussian noise, both CAM and CAT.G outperform CAT.E. This can (at least) be attributed to two factors: (i) CAT.E does not, unlike CAM and CAT.G, explicitly use the Gaussian noise specification and (ii) differential entropy estimation is a difficult statistical problem (see, e.g., Paninski, 2003; Han et al., 2020) For small and moderate deviations from Gaussianity, CAT.G outperforms both CAM and CAT.E. For larger deviations, CAT.E outperforms both CAT.G and CAM in terms of median SHD. Finally, we note that CAT.G always outperforms CAM in terms of median SHD.

6.2 Robustness: CAT on DAGs

This experiment analyzes how CAT performs compared to CAM and the max-min hill-climbing (MMHC, Tsamardinos et al., 2006) structure learning method using the Bayesian Gaussian equivalent score (BGe, Geiger and Heckerman, 1994; Heckerman and Geiger, 1995) (the latter method is not expected to work well in our setting, as it does not exploit the additional identifiability). We compare the performance of these structure learning methods when applied to data generated from an additive Gaussian noise model with a non-tree DAG as a causal graph. More specifically, we analyze the behavior on single-rooted DAGs.

For any fixed $p \in \mathbb{N}$ we generate a directed tree of Type 1 and for each zero in the upper triangular part of the adjacency matrix we add an edge with 5% probability. The causal functions and Gaussian noise innovations are generated according to the specifications given in the experiment of Section 6.4.2. The structural assignment for each node is additive in each causal parent, i.e., for all $i \in \{1, \dots, p\}$, $X_i := \sum_{j \in \mathrm{pa}^{\mathcal{G}}(i)} f_{ji}(X_j) + N_i$, with $(N_1, \dots, N_p)$ mutually independent Gaussian distributed noise innovations. For each $p \in \{16, 32, 64\}$ and sample size $n \in \{50, 250, 500\}$ we randomly generate 200 single-rooted Gaussian additive models according to the above specifications. For this experiment, we employ CAM with preliminary neighborhood selection and subsequent pruning.

As CAT.G outputs trees, we do not expect it to output the correct graph. In Figure 15 of Section C.2 of Appendix C we have illustrated boxplot comparisons of the SHD between the estimated and true graph for CAM, CAT.G and the MMHC with BGe score (MMHC.BGe). We see a clear ranking of the methods in terms of SHD performance. The best performance is seen for CAM, followed by CAT.G, and finally the worst performing method is that of MMHC.BGe. Note that the BGe score (and various other Bayesian network learning scores) is only suitable for jointly Gaussian data, e.g., for linear additive Gaussian noise systems.

Figure 6 illustrates the performance in terms of ancestor relations. For small to moderately sized systems ($p \in \{16, 32\}$) CAM slightly outperforms CAT.G in terms of median precision (TP/(TP + FP)) when classifying causal ancestors. However, for large systems ($p = 64$) CAT.G outperforms CAM for median precision. On the other hand, CAM is not limited to trees which allows it to find a more significant proportion of the true ancestor, as seen by median recall (TP/P) performance. MMHC.BGe shows subpar performance in terms of ancestor classification, except for large systems and sample sizes when considering recall. CAT.G seems to be a viable alternative for practical non-tree applications where precision more important than recall for classifying causal ancestor relations.

Figure 14 in Section C.2 of Appendix C illustrates similar comparisons when focusing on causal edges. The precision of CAT.G is larger than that of CAM only for small sample sizes, while the opposite is true for large sample sizes. As expected, and as seen for ancestor relations, CAM outperforms both CAT.G and MMHC.BGe in terms of recall.

Finally, while both methods are computationally efficient, CAT has a slightly lower runtime than the greedy search algorithm of CAM. The average runtime of CAM and CAT.G in this experiment for $p = 64$ and $n = 500$ was 288 and 199 seconds, respectively. For both methods, the most time consuming part is estimating the conditional expectations that are used to compute the edge weights.

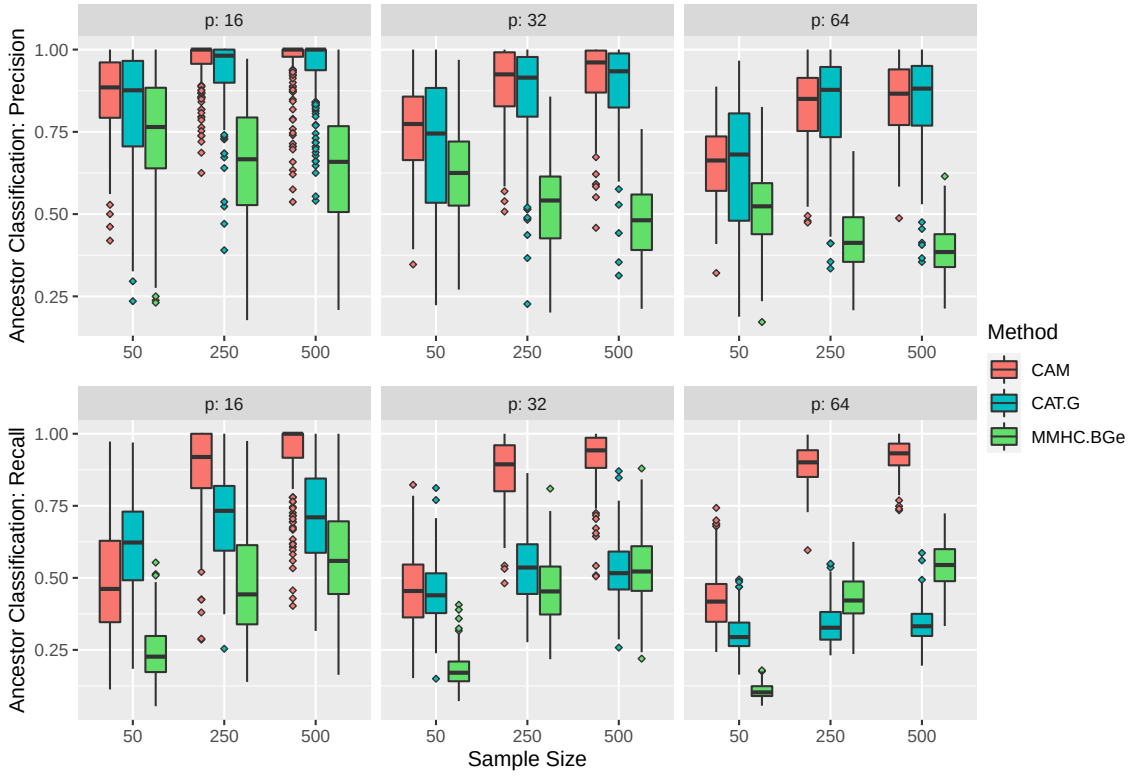

Figure 6: Evaluating the robustness of CAT by estimating ancestor relations in non-tree DAGs, see Section 6.2. CAT.G slightly outperforms CAM in terms of true positive rates for large graphs (top) but finds less ancestor relationships (bottom) due to fitting a tree. As expected, CAT.G and CAM outperform MMHC in terms of precision.

6.3 Hypothesis Testing

In this experiment, we experimentally analyze the finite sample size and power properties of the two substructure hypothesis testing procedures proposed in Section 4.2. We generate the underlying models and data similarly to the experimental setup of the Gaussian noise experiment of Section 6.1.2. We generate a random tree of Type 2 (see Section 6.1.1) of size p with Gaussian process causal functions and Gaussian noise innovations generated in accordance with the description in Section 6.1.2.

Given a finite sample of size n we use the first $\lfloor n/2 \rfloor$ observations to estimate all possible conditional mean functions $x \mapsto \mathbb{E}[X_i | X_j = x]$ for $j \neq i$ with thin plate regression splines (R-package `mgcv` with default settings). The remaining $n - \lfloor n/2 \rfloor$ observations are used to estimate the upper and lower Bonferroni corrected confidence bounds $\hat{l} = (\hat{l}_{ji})_{j \neq i}$ and $\hat{u} = (\hat{u}_{ji})_{j \neq i}$ as defined in Equation (10) of Section 4. Using the two testing procedures proposed in Algorithms 2 and 3 of Section 4.2, with a significance level of 5%, we test all simple hypotheses, i.e., all hypotheses of the form $\mathcal{H}_0 : (j \to i)$ and $\mathcal{H}_0 : (j \not\to i)$ for all $j \neq i$. We repeat this procedure 400 times to observe the average behavior of the testing procedure

for the previously mentioned system generation scheme. We do this for all combinations of sample sizes $n \in \{500, 1000, 5000, 10000, 20000\}$ and system sizes $p \in \{2, 4, 6, 8, 16\}$.

Figure 7 illustrates the resulting power properties of the two tests CheckC and ConvB. Both testing procedures have better small sample power when testing a false hypothesis of the form $\mathcal{H}_0 : (j \to i)$ compared to testing a false hypothesis of the form $\mathcal{H}_0 : (j \not\to i)$. Furthermore, the finite sample power of CheckC is inferior to the ConvB. The power of ConvB is only slightly negatively affected by an increase in system size p, when testing a false hypothesis of the form $\mathcal{H}_0 : (j \not\to i)$. On the other hand, CheckC suffers for both types of hypotheses when increasing the system size. For example, the CheckC method has almost zero power when the system size is 16 and the samplesize is 20000.

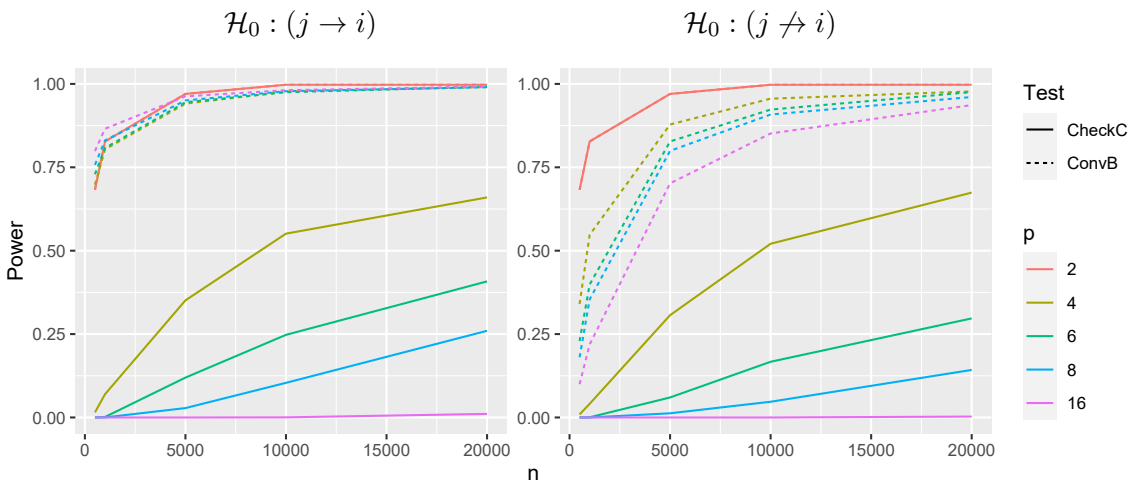

Figure 7: This figure illustrates the power of the proposed testing procedure for simple hypotheses. The left plot shows the empirical probability of rejecting a false hypothesis of the form $\mathcal{H}_0 : (j \to i)$ as a function of the sample size n. Similarly, the right plot shows the empirical probability of rejecting a false hypothesis of the form $\mathcal{H}_0(j \not\to i)$. For both tests the power increases with growing sample size with ConvB outperforming CheckC.

In Table 1, we further detail the power and level achieved by the ConvB test in the above experiment. Both tests seems to hold level in all settings. For false hypotheses of the form $\mathcal{H}_0 : (j \to i)$ we have split the hypotheses into three groups based on Distance(j, i) being 'negative', 'positive' or 'no path'. If j is a descendant of i, then Distance(j, i) is 'negative', if j is a non-parent ancestor of i, then Distance(j, i) is 'positive', and if there is no directed path between j and i, then Distance(j, i) equals 'no path'. For moderately large sample sizes the test exhibits high power. However, the power of the test for a false hypothesis of the form $(j \to i)$ when j is a non-parent ancestor of i is relatively low.

6.4 Identifiability Gap

We now investigate the behavior of the identifiability gap in bivariate models (Section 6.4.1) and evalute the lower bound derived in Section 5 empirically for multivariate models (Section 6.4.2).

Property:		Power of test					Size of test	
$\mathcal{H}_0$:		$(j \to i)$				$(j \not\to i)$	$(j \to i)$	$(j \not\to i)$
		Distance(j,i)						
p	n	Negative	Positive	No Path	Total	Total	Total	Total
2	500	0.68	—	—	0.68	0.68	0.00	0.00
2	1000	0.82	—	—	0.82	0.82	0.00	0.00
2	5000	0.97	—	—	0.97	0.97	0.00	0.00
2	10000	0.99	—	—	0.99	0.99	0.00	0.00
2	20000	0.99	—	—	0.99	0.99	0.00	0.00
4	500	0.69	0.32	0.85	0.69	0.34	0.01	0.01
4	1000	0.79	0.54	0.92	0.80	0.54	0.00	0.00
4	5000	0.93	0.86	0.98	0.94	0.87	0.00	0.01
4	10000	0.97	0.93	0.99	0.97	0.95	0.00	0.00
4	20000	0.99	0.96	0.99	0.99	0.97	0.00	0.00
8	500	0.73	0.38	0.85	0.75	0.18	0.01	0.01
8	1000	0.78	0.53	0.91	0.83	0.35	0.01	0.01
8	5000	0.92	0.86	0.98	0.95	0.79	0.01	0.01
8	10000	0.96	0.94	0.99	0.97	0.90	0.00	0.01
8	20000	0.98	0.97	0.99	0.99	0.96	0.00	0.00
16	500	0.77	0.40	0.86	0.79	0.09	0.01	0.01
16	1000	0.80	0.54	0.92	0.86	0.21	0.01	0.01
16	5000	0.91	0.85	0.98	0.96	0.70	0.01	0.01
16	10000	0.95	0.93	0.99	0.98	0.85	0.01	0.01
16	20000	0.97	0.97	0.99	0.99	0.93	0.00	0.00

Table 1: This table contains further details on the average power and size of the ConvB hypothesis test under the data generation described in Section 6.3.

6.4.1 Bivariate Identifiability Gap

In this experiment, we investigate the behavior of the bivariate identifiability gap and analyze setups with both Gaussian and non-Gaussian noise innovations. Let us consider an additive noise model over (X, Y) with causal graph $X \to Y$. The causal functions will be chosen from the following function class. For any $\lambda \in [0, 1]$, define $f_\lambda : \mathbb{R} \to \mathbb{R}$ as

$$f_\lambda(x) = (1 - \lambda)x^3 + \lambda x.$$

That is, $\lambda \mapsto f_\lambda$ interpolates between a cubic function $x \mapsto x^3$ and a linear function $x \mapsto x$. For any $(\alpha, \lambda) \in (0, \infty) \times [0, 1]$ we consider the following bivariate structural causal additive model

$$X := \operatorname{sign}(N_X)|N_X|^\alpha, \quad Y := f_\lambda(X) + N_Y,$$

where N_X, N_Y are independent standard normal distributed random variables. Recall that the bivariate identifiability gap is given by

$$\begin{aligned}\ell_{\mathrm{E}}(Y \to X) - \ell_{\mathrm{E}}(X \to Y) &= h(X - \mathbb{E}[X|Y]) + h(Y) - h(X - \mathbb{E}[X|Y], Y) \\ &= I(X - \mathbb{E}[X|Y]; Y),\end{aligned}$$

by Lemma 13. Thus, the causal graph $X \to Y$ is identified by the entropy score function if $I(X - \mathbb{E}[X|Y]; Y) > 0$.

For any fixed λ and α we now estimate the identifiability gap; we also calculate the p-value associated with the null hypothesis that the identifiability gap is zero (based on 50000 observations). Similarly to the previous experiment, we estimate the conditional expectations using thin-plate spline regression. We estimate (without sample splitting) the identifiability gap and construct p-values using the CRAN R-package `IndepTest` (Berrett et al., 2018). More specifically, we use the differential entropy estimator of Berrett et al. (2019) and the mutual information based independence test of Berrett and Samworth (2019), respectively.

The heatmap of Figure 8 illustrates the behavior of the identifiability gap for all combinations of $\lambda \in \{0, 0.05, \ldots, 1\}$ and $\alpha \in \{0.3, 0.4, \ldots, 1.7\}$. It suggests that the identifiability gap only tends to zero when we approach the linear additive Gaussian noise setup. Only in the models closest to the linear additive Gaussian noise setup are we unable to reject the null-hypothesis of a vanishing identifiability gap.

This is also what the theory predicts, namely that for bivariate linear additive Gaussian noise models, the causal direction is not identified. It is known that for linear models, non-Gaussianity is helpful for identifiability. The empirical results indicate that the same holds for nonlinear models, i.e., that the identifiability gap increases with the degree of non-Gaussianity of the noise innovations.

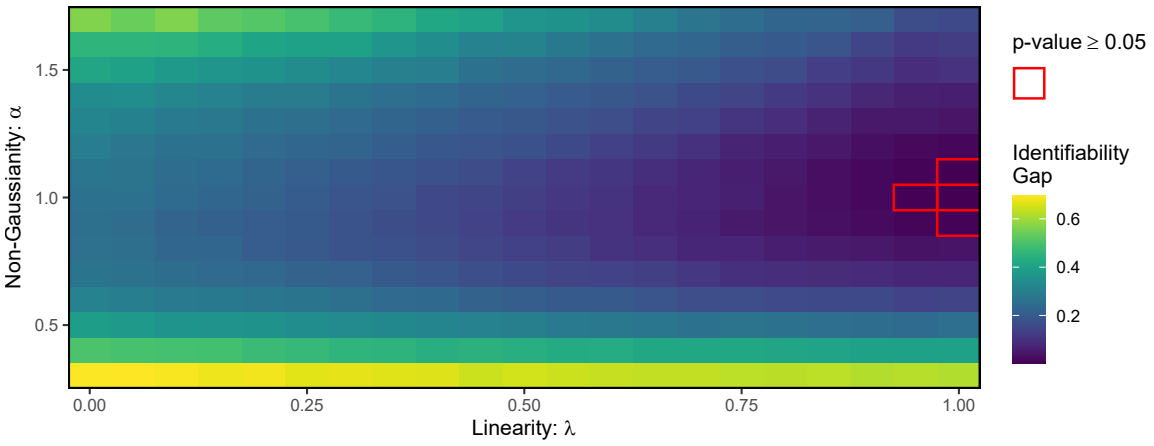

Figure 8: Heatmap of the identifiability gap for varying λ and α. Tiles with a red boundary correspond to the models for which the mutual information based independence test cannot reject the null hypothesis of a vanishing identifiability gap.

6.4.2 Multivariate Identifiability Gap

In this experiment, we investigate the identifiability gap and its relation to the lower bounds established in Theorem 17. For a causal additive tree model with Gaussian noise, it holds that

$$\min_{\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}} \ell_{\mathrm{G}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{G}}(\mathcal{G}) \geq \min \left\{ \min_{(w,l,o) \in \Pi_W(\mathcal{G})} I(X_w; X_l \mid X_o), \min_{i \to j \in \mathcal{E}} \Delta \ell_{\mathrm{E}}(i \xleftarrow{} j) \right\}.$$

In other words, the identifiability gap is lower bounded by the minimum of the smallest local faithfulness measures and the smallest edge-reversal score difference. We now investigate empirically how important the first term is for the inequality to hold. More specifically, for a given model generation scheme, we quantify how often the minimum edge reversal is sufficiently small to establish the lower bound without the conditional mutual information term, that is, how often the identifiability constant $\min_{\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}} \ell_{\mathrm{G}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{G}}(\mathcal{G})$ is larger than the minimum edge reversal.

The minimum edge reversal can be estimated using the same conditional expectation and entropy estimators of the experiment in Section 6.4.1. However, estimating the identifiability gap between the second-best scoring tree and the causal tree needs further elaboration. We know that the best scoring (causal) tree can be found by Chu–Liu–Edmonds' (a directed MWST) algorithm. The second-best scoring tree differs from the best scoring tree in at least one edge. Thus, given the best scoring graph, we remove one of the $p-1$ edges of the best scoring tree from the pool of possible edges and rerun Chu–Liu–Edmonds' algorithm. We do this for each of the $p-1$ edges in the best scoring tree which leaves us with $p-1$ possibly different sub-optimal trees of which the minimum score is attained by the second-best scoring graph.

For the experiment, we randomly sample data generating models similarly to the experiment in Section 6.1.2. However, we change the causal functions from explicit sample paths of a Gaussian process to a thin-plate spline regression model estimating the sample paths due to memory constraints when generating large sample sizes. Figure 9 illustrates, for $p \in \{8, 16\}$, boxplots of the difference between the identifiability gap and the minimum edge reversal for 100 randomly generated causal additive tree models with Gaussian noise. For each model, the identifiability gap and corresponding minimum edge reversal is estimated from 200000 independent and identically distributed observations. The illustration suggests that it is in general necessary to also consider the conditional mutual information term in order to establish a lower bound. However, it also shows that in the majority (90%) of the models, the minimum edge reversal is indeed a lower bound for the identifiability gap.

7. Empirical Application

We consider the well-known non-synthetic bio-informatics data set considered by Sachs et al. (2005). The data set contains simultaneous measurements of expression levels of 11 different phosphorylated proteins and phospholipids of human immune system cells under both observational and interventional experimental settings. Sachs et al. (2005) present (based on expert consensus and experiments) a causal directed acyclic graph with 11 nodes and 20 edges for the 11 phosphorylated proteins and phospholipids.

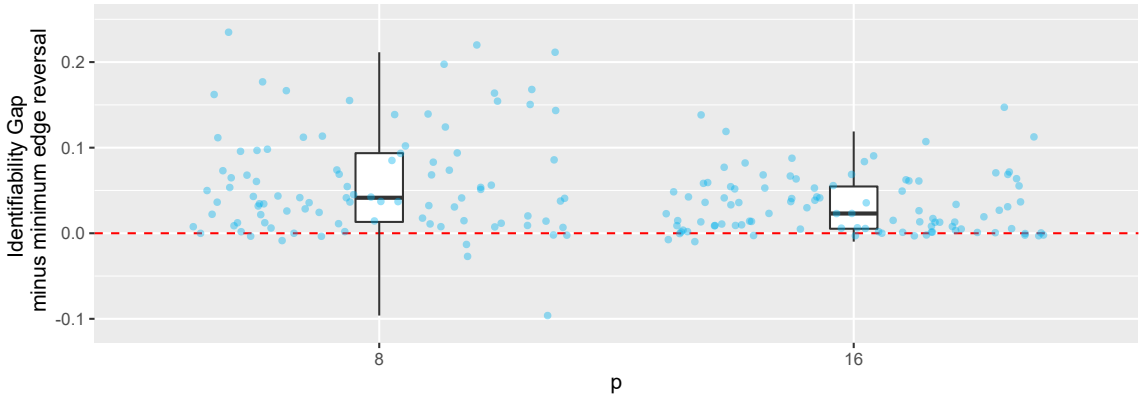

Figure 9: Empirical analysis of the lower bound on the identifiability gap, see Section 6.4.2. In most of the simulated settings, we see that the estimated identifiability gap is larger than the smallest edge-reversal score difference. This suggests that in many cases, the latter term is sufficient for establishing a lower bound on the identifiability gap. We have also implemented CAT.G and CAT.E with the heuristic pruning procedure introduced in Bühlmann et al. (2014).

We compare our structure learning methods CAT.G and CAT.E with the score-based methods of CAM (Bühlmann et al., 2014), GES (Chickering, 2002), NoTears (Zheng et al., 2018) and the mixed method MMHC (Tsamardinos et al., 2006). The structure learning methods are applied to observational data (853 observations using reagents anti-CD3 and anti-CD28). The results of the structure learning methods can be seen in Table 2. Learning causal structure from observational data is a difficult problem but several methods seem to outperform estimating an empty graph or a random graph. CAM is superior in terms of SHD, SID, and recall of edge and root predictions, suggesting that in this data set, one may indeed exploit nonlinearities for indentifying causal structure. However, we also see that CAT.G shows competitive performance and ranks in first or second place with respect to all reported performance measures. Interestingly, even though CAT.G approximates the non-tree causal DAG by a directed tree, it outperforms various DAG structure learning methods such as classical approaches of GES and MMHC and the more recent continuous optimization approach of NoTears. CAT.E does not perform well on these data, witnessing that estimating entropies is a difficult statistical problem.

Finally, we also evaluate the proposed hypothesis testing procedures on this data set, even though the asymptotic guarantees of the hypothesis tests derived in Section 4 are not guaranteed to hold as the true underlying graph is not a directed tree. We test every possible simple hypothesis of the form $\mathcal{H}_0(j \to i)$ and $\mathcal{H}_0(j \not\to i)$. The results can be seen in Table 3 (the CheckC test holds level but has zero power). The ConvB test shows reasonable power against false hypothesis of the form $\mathcal{H}_0(j \to i)$, however, it has no power against the false hypotheses of the form $\mathcal{H}_0(j \not\to i)$. Rejection rates of the true hypotheses of the form $\mathcal{H}_0(j \to i)$ are larger than the asymptotically guaranteed rate of 0.05, possibly because of

the model violation; this phenomenon is not as expressed for true hypotheses of the form $\mathcal{H}_0(j \not\to i)$.

Method	Prune	Score	SHD	SHD-C	SID	Precision	Recall
CAM	Yes	ℓ_G	14.00	15.00	72.0	0.571	0.381
CAT	No	ℓ_G	14.00	14.00	79.0	0.636	0.333
CAT	Yes	ℓ_G	15.00	16.00	83.0	0.545	0.286
MMHC	—	BGe	15.00	14.00	84.0	0.417	0.238
MMHC	—	BIC	15.00	14.00	84.0	0.417	0.238
GES	—	BIC	17.00	16.00	107.0	0.231	0.143
CAT	Yes	ℓ_E	18.00	19.00	92.0	0.273	0.143
NoTears	—	—	19.00	17.00	99.0	0.182	0.095
EmptyGraph	—	—	20.00	20.00	94.0	0.091	0.048
RandomGraph	—	—	22.32	21.93	94.7	0.271	0.170
CAT	No	ℓ_E	24.00	25.00	104.0	0.273	0.143

Table 2: Results of the empirical application of various structure learning methods to the data set of Sachs et al. (2005). Here we report the structural hamming distance (SHD), structural hamming distance of the respective CPDAGs (SHD-C), and structural intervention distance (SID) between the causal graph and the estimated graph. The latter two columns show the precision and recall for edge and root classification. The methods EmptyGraph always outputs the empty graph and the method RandomGraph outputs a random single-rooted tree generated according to the generation scheme outlined in Section 6.2. We have implemented CAT.G and CAT.E both with and without the heuristic pruning technique introduced in Bühlmann et al. (2014)

	Nulls incorrect					Nulls correct	
$\mathcal{H}_0$:	$(j \to i)$				$(j \not\to i)$	$(j \to i)$	$(j \not\to i)$
	Distance(j, i)						
	Negative	Positive	No Path	Total	Total	Total	Total
Rejection rates:	0.58	0.53	0.66	0.58	0.00	0.30	0.02
N:	46	26	18	90	20	20	90

Table 3: Further details on the average power and level of the ConvB test with a significance level of 0.05. Here, we have tested every simple hypothesis of the Sachs et al. (2005) data set; see Section 6.3 for further explanations of the distance metric. N denotes the number of hypothesis tests that have been averaged.

8. Summary and Future Work

This paper shows that exact structure learning is possible for systems of lesser complexity, i.e., for restricted structural causal models with additive noise and causal graphs given by directed trees. We propose the method CAT, which is guaranteed to consistently recover the causal directed tree of a causal additive tree model with Gaussian noise under mild assumptions on the regression methods used to estimate conditional means. Furthermore, we argue that CAT is consistent in an asymptotic setup with vanishing identifiability. We present a computationally feasible procedure to test substructure hypotheses and provide an analysis of the identifiability gap. Simulation experiments show that CAT outperforms other (more general) structure learning methods for the specific task of recovering the causal graph in additive noise structural causal models when the causal structure is given by directed trees.

The proof of Proposition 4 is based on the fact that the causal functions of alternative models are differentiable and that the noise densities are continuous. We conjecture that it is possible to get even stronger identifiability statements under weaker assumptions; proving such a result necessitates new proof strategies. We believe that it could be possible to prove uniform consistency under suitable conditions when requiring that the infimum of the identifiability gap is strictly positive and that the mean squared errors of the regression estimates converge uniformly. Furthermore, we believe that it could inspire future research on more general identifiability conditions (e.g., relaxing the smoothness assumptions) for directed trees and DAGs under the assumption of additive noise. Furthermore, it should be possible to use a wild bootstrap approach to construct a simultaneous hyperrectangle confidence region for the Gaussian edge weights. This would, however, require a sufficiently fast convergence rate of the estimation error of the conditional expectations corresponding to non-causal edges. Compared to the Bonferroni correction, this approach could increase the power of the test. We hypothesize that the ConvB test holds level even in many generic, non-identifiable settings.

Acknowledgments

We thank Phillip Bredahl Mogensen and Thomas Berrett for helpful discussions on the entropy score and its estimation. JP thanks Martin Wainwright for discussions on greedy search methods (and when they fail) and inequalities in additive noise models during his visit at UC Berkeley in 2013. PB and JP thank David Bürge and Jan Ernest for helpful discussions on exploiting Chu–Liu–Edmonds' algorithm for causal discovery during the early stages of this project. MEJ and JP were supported by the Carlsberg Foundation; JP was, in addition, supported by a research grant (18968) from VILLUM FONDEN. RDS was supported by EPSRC grant EP/N031938/1. PB received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 786461).

Appendix A. Graph Terminology

A *directed graph* $\mathcal{G} = (V, \mathcal{E})$ consists of $p \in \mathbb{N}_{>0}$ vertices (nodes) $V = \{1, \dots, p\}$ and a collection of directed edges $\mathcal{E} \subset \{(j \to i) \equiv (j, i) : i, j \in V, i \neq j\}$. For any graph $\mathcal{G} = (V, \mathcal{E})$

we let $\mathrm{pa}^{\mathcal{G}}(i) := \{v \in V : \exists (v,i) \in \mathcal{E}\}$ and $\mathrm{ch}^{\mathcal{G}}(i) := \{v \in V : \exists (j,v) \in \mathcal{E}\}$ denote the *parents* and *children* of node $i \in V$ and we define root nodes $\mathrm{rt}(\mathcal{G}) := \{v \in V : \mathrm{pa}^{\mathcal{G}}(i) = \emptyset\}$ as nodes with no parents (that is, no incoming edges). A *path* in $\mathcal{G}$ between two nodes $i_1, i_k \in V$ consists of a sequence $(i_1, i_2), \ldots, (i_{k-1}, i_k)$ of pairs of nodes such that for all $j \in \{1, \ldots, k-1\}$, we have either $(i_j \to i_{j+1}) \in \mathcal{E}$ or $(i_{j+1} \to i_j) \in \mathcal{E}$. A *directed path* in $\mathcal{G}$ between two nodes $i_1, i_k \in V$ consists of a sequence $(i_1, i_2), \ldots, (i_{k-1}, i_k)$ of pairs of nodes such that for all $j \in \{1, \ldots, k-1\}$, we have $(i_j \to i_{j+1}) \in \mathcal{E}$. Furthermore, we let $\mathrm{an}^{\mathcal{G}}(i)$ and $\mathrm{de}^{\mathcal{G}}(i)$ denote the *ancestors* and *descendants* of node $i \in V$, consisting of all nodes $j \in V$ for which there exists a directed path to and from i, respectively. We let $\mathrm{nd}^{\mathcal{G}}(i)$ denote the *non-descendants* of i.

A *directed acyclic graph* (DAG) is a directed graph that does not contain any directed cycles, i.e., directed paths visiting the same node twice. We say that a graph is *connected* if a (possibly undirected) path exists between any two nodes. A *directed tree* is a connected DAG in which all nodes have at most one parent. More specifically, every node has a unique parent except the root node, which has no parent. The root node $\mathrm{rt}(\mathcal{G})$ is the unique node such there exists a directed path from $\mathrm{rt}(\mathcal{G})$ to any other node in the directed tree. In graph theory, a directed tree is also called an *arborescence*, a *directed rooted tree*, and a *rooted out-tree*. A graph $\mathcal{G} = (V', \mathcal{E}')$ is a *subgraph* of another graph $\mathcal{G} = (V, \mathcal{E})$ if $V' \subseteq V$, $\mathcal{E}' \subseteq \mathcal{E}$ and for all $(j \to i) \in \mathcal{E}'$ it holds that $j, i \in V'$. A subgraph is *spanning* if $V' = V$. For any DAG $\mathcal{G} = (V, \mathcal{E})$ and three mutually distinct subsets $A, B, C \subset V$ we let $A \perp\!\!\!\perp_{\mathcal{G}} B \,|\, C$ denote that A and B are d-separated by C in $\mathcal{G}$ (see, e.g., Pearl, 2009).

Appendix B. Further Details on Section 5

Remark 1 *The conditional entropy score gap is not strictly positive when considering the alternative graphs $\tilde{\mathcal{G}}$ that are Markov equivalent to the causal graph $\mathcal{G}$, $\tilde{\mathcal{G}} \in \mathrm{MEC}(\mathcal{G})$. A simple translation of the conditional entropy score function reveals that*

$$\ell_{\mathrm{CE}}(\tilde{\mathcal{G}}) + C = \sum_{(j \to i) \in \tilde{\mathcal{E}}} h(X_i | X_j) - h(X_i) = - \sum_{(j \to i) \in \tilde{\mathcal{E}}} I(X_i; X_j),$$

for a constant $C \in \mathbb{R}$. By symmetry of the mutual information, it holds that $\ell_{\mathrm{CE}}(\tilde{\mathcal{G}}) = \ell_{\mathrm{CE}}(\mathcal{G})$, for any $\tilde{\mathcal{G}} \in \mathrm{MEC}(\mathcal{G})$, since $\tilde{\mathcal{G}}$ and $\mathcal{G}$ share the same skeleton. Thus, the conditional entropy score function can, at most, identify the Markov equivalence class of the causal graph. In fact, the polytree causal structure learning method of Rebane and Pearl (1987) uses the above translated conditional entropy score function to recover the skeleton of the causal graph.

Example 3 (Negative local Gaussian score gap) *Consider two graphs $\mathcal{G}$ and $\tilde{\mathcal{G}}$ with different root nodes, i.e., $\mathrm{rt}(\mathcal{G}) \neq \mathrm{rt}(\tilde{\mathcal{G}})$. If $x \mapsto \mathbb{E}[X_{\mathrm{rt}(\mathcal{G})} | X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(\mathrm{rt}(\mathcal{G}))} = x]$ is not almost surely constant, then it holds that*

$$\begin{aligned}
\ell_{\mathrm{G}}(\tilde{\mathcal{G}}, \mathrm{rt}(\mathcal{G})) - \ell_{\mathrm{G}}(\mathcal{G}, \mathrm{rt}(\mathcal{G})) &= \mathbb{E}[(X_{\mathrm{rt}(\mathcal{G})} - \mathbb{E}[X_{\mathrm{rt}(\mathcal{G})} | X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(\mathrm{rt}(\mathcal{G}))}])^2] - \mathrm{Var}(X_{\mathrm{rt}(\mathcal{G})}) \\
&= \mathbb{E}[\mathrm{Var}(X_{\mathrm{rt}(\mathcal{G})} | X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(\mathrm{rt}(\mathcal{G}))})] - \mathrm{Var}(X_{\mathrm{rt}(\mathcal{G})}) \\
&= -\mathrm{Var}(\mathbb{E}[X_{\mathrm{rt}(\mathcal{G})} | X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(\mathrm{rt}(\mathcal{G}))}]) < 0.
\end{aligned}$$

Appendix C. Further Details on the Simulation Experiments

This section contains further details on the simulation experiments.

C.1 Tree Generation Algorithms

The following two algorithms, Algorithm 4 (many leaf nodes) and Algorithm 5 (many branch nodes), details how the Type 1 and Type 2 trees are generated, respectively.

Algorithm 4 Generating type 1 trees

```
procedure TYPE1(p)
    A := 0 ∈ R^{p×p}
    for j ∈ {1, ..., p} do
        for i ∈ {j + 1, ..., p} do
            if Σ_{k=1}^{p} A_{ki} = 0 then
                if i = j + 1 then
                    A_{ji} := 1
                else
                    A_{ji} := Binomial(success = 0.1)
                end if
            else
                A_{ji} := 0
            end if
        end for
    end for
    return A
end procedure
```

Algorithm 5 Generating type 2 trees

```
procedure TYPE2(p)
    for i ∈ {2, ..., p} do
        j := sample({1, ..., i − 1})
        A_{ji} := 1
    end for
    return A
end procedure
```

C.2 Additional Illustrations

This section contains some additional illustrations of the simulation experiments.

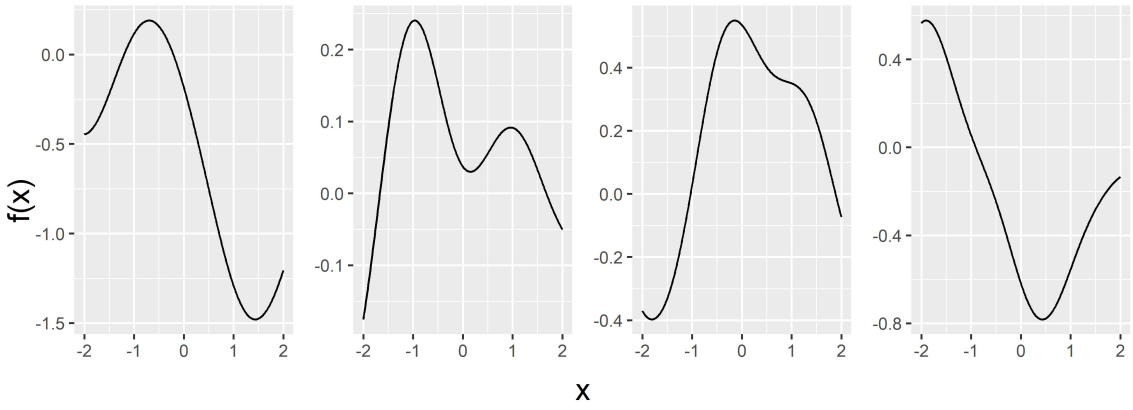

Figure 10: Four causal functions as modeled by the RBF kernel Gaussian Process.

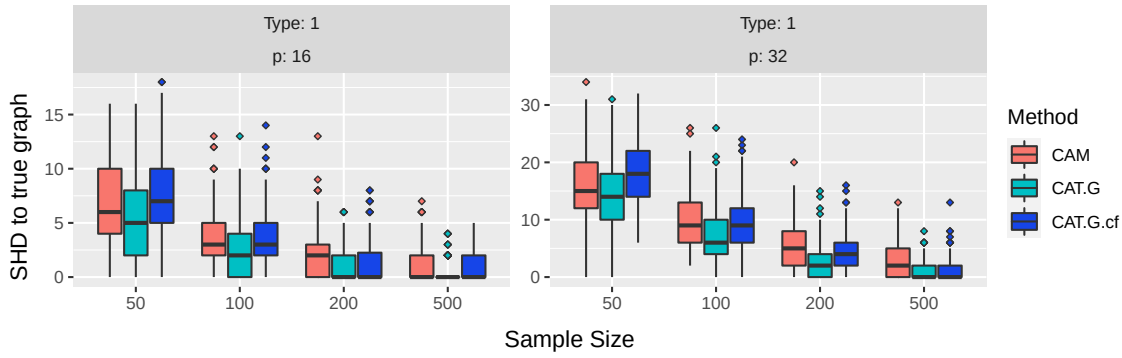

Figure 11: Boxplot illustrating the SHD performance of CAM and CAT for varying sample sizes, system sizes and tree types in the experiment of Section 6.1.2 with 200 repetitions. CAT.G.cf is the CAT.G method with cross-fitted edge weights. We see that cross-fitting has no positive impact on the performance. It seems to worsen the performance for small sample sizes.

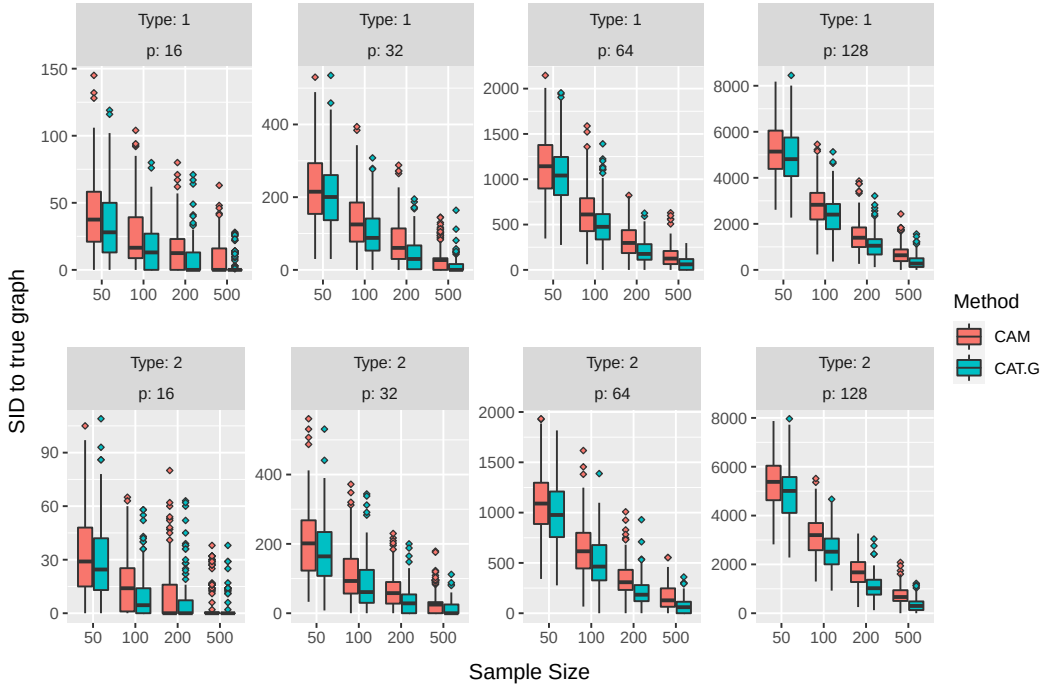

Figure 12: Boxplot illustrating the SID performance of CAM and CAT for varying sample sizes, system sizes and tree types in the experiment of Section 6.1.2. CAT.G is CAT with edge weights derived from the Gaussian score function.

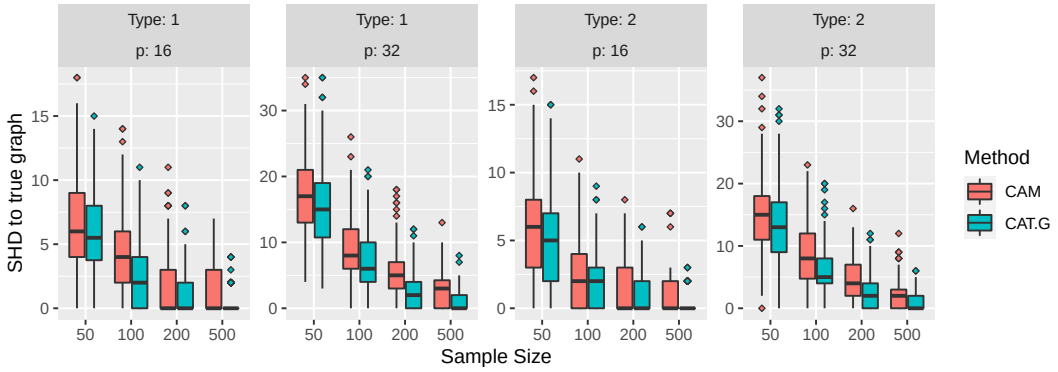

Figure 13: Boxplot illustrating the SHD performance of CAM and CAT for varying sample sizes, system sizes and tree types in the experiment of Section 6.1.2. Here CAT.G is run on the CAM edge weights , so that any difference in nonparametric regression technique is ruled out as the source of the performance difference.

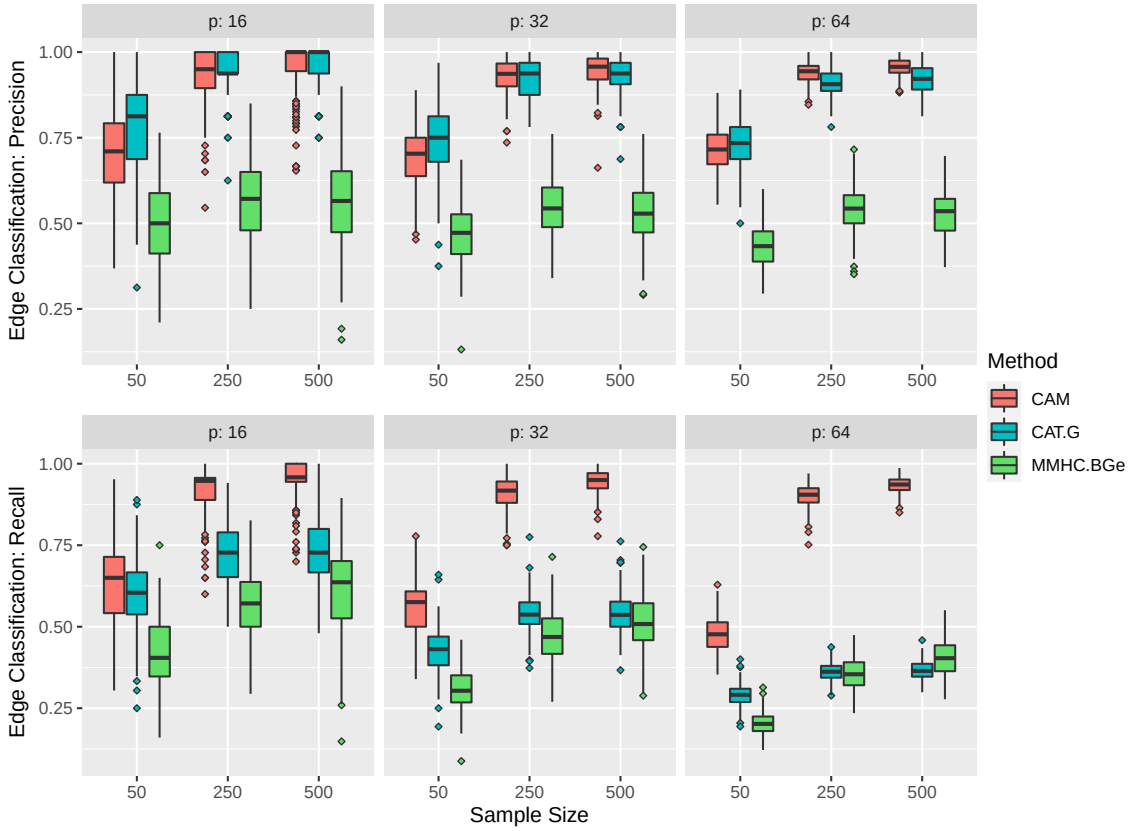

Figure 14: Boxplot of edge relations for the experiment in Section 6.2.

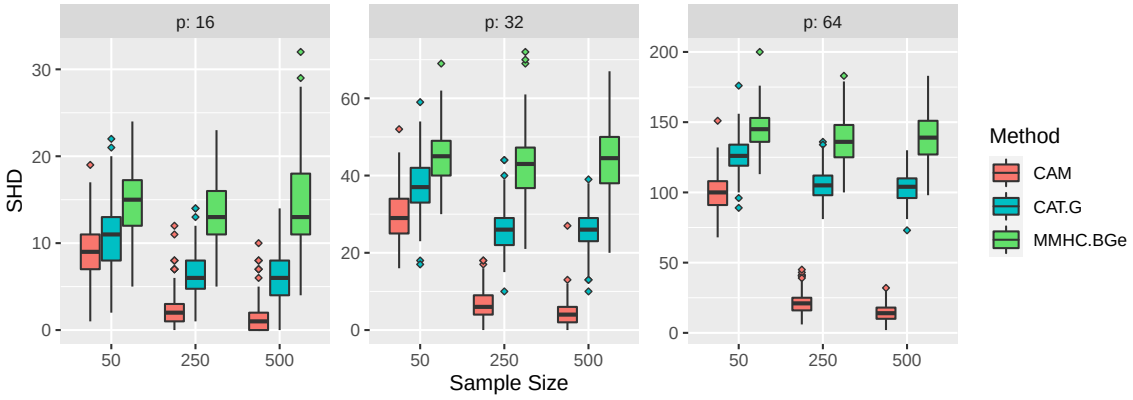

Figure 15: Boxplot of SHD for the experiment in Section 6.2.

Appendix D. Proofs

This section contains the proofs of all results presented in the main text.

D.1 Proofs of Section 2

Proof of Lemma 3. Let $\theta = (\mathcal{G}, (f_i), P_N) \in \mathcal{T}_p \times \mathcal{D}_3^p \times \mathcal{P}_{\mathrm{G}}^p$. Furthermore, let all causal functions (f_i) be nowhere constant and nonlinear. The additive noise is Gaussian, so the log density of N_i for all $i \in \{1, \dots, p\}$ is given by

$$\nu_i(x) = -\frac{1}{2}\log(2\pi\sigma_i^2) - \frac{x^2}{2\sigma_i^2}, \quad \nu_i'(x) = -\frac{x}{\sigma_i^2}, \quad \nu_i''(x) = -\frac{1}{\sigma_i^2}, \quad \nu_i'''(x) = 0.$$

By assumption we have that condition (i) of Definition 2 is satisfied, hence assume for contradiction that condition (ii) of Definition 2 is not satisfied. That is, we assume that there exists an $i \in \{1, ..., p\} \setminus \{\mathrm{rt}(\mathcal{G})\}$ such that for all

$$\begin{aligned}(x, y) \in \mathcal{J} :=& \{(x, y) \in \mathbb{R}^2 : \nu_i''(y - f_i(x))f_i'(x) \neq 0\} \\ =& \{(x, y) \in \mathbb{R}^2 : f_i'(x) \neq 0\},\end{aligned}$$

it holds that

$$\xi'''(x) - \xi''(x)\frac{f_i''(x)}{f_i'(x)} - \frac{2f_i''(x)f_i'(x)}{\sigma^2} = -\frac{y - f_i(x)}{\sigma^2}\left(f_i'''(x) - \frac{(f_i''(x))^2}{f_i'(x)}\right). \tag{18}$$

Henceforth, suppress the subscript i of f_i and σ_i. First note that $\{x \in \mathbb{R} : f'(x) = 0\}$ is closed by continuity of f'. The complement is open, hence there exists a countable collection of mutually disjoint open intervals $(O_k)_{k\in\mathbb{Z}}$ such that $\{x \in \mathbb{R} : f'(x) \neq 0\} = \cup_{k\in\mathbb{Z}} O_k$. Since f is nowhere constant we know that $\{x \in \mathbb{R} : f'(x) = 0\}$ has empty interior which implies that $\overline{\cup_{k\in\mathbb{Z}} O_k} = \mathbb{R}$. Now let $(O_k)_{k\in\mathbb{Z}}$ be indexed by $\mathbb{Z}$ such that for any $k, j \in \mathbb{Z}$ with $k < j$ and $x \in O_k, y \in O_j$ it holds that $x < y$. As the left-hand side of Equation (18) is constant in y it must hold that

$$0 = f'''(x) - \frac{(f''(x))^2}{f'(x)} = \frac{\frac{\partial f''(x)}{\partial x}f'(x) - f''(x)\frac{\partial f'(x)}{\partial x}}{(f'(x))^2} = \frac{\partial}{\partial x}\left(\frac{f''(x)}{f'(x)}\right),$$

i.e., $f''(x)/f'(x)$ is constant, for all $x \in \cup_{k\in\mathbb{Z}} O_k$.

On each O_k we have that $\partial/\partial x \log(\mathrm{sign}(f'(x))f'(x)) = c_{k,1} \iff \log(\mathrm{sign}(f'(x))f'(x)) = c_{k,1}x + c_{k,2} \iff \mathrm{sign}(f'(x))f'(x) = \exp(c_{k,1}x + c_{k,2}) \iff f'(x) = \pm\exp(c_{k,1}x + c_{k,2})$. Recall that we have assumed continuous differentiability of f'. That is, for any $k \in \mathbb{Z}$ and $t_k := \sup(O_k) = \inf(O_{k+1})$ we have $\lim_{x\uparrow t_k} f'(x) = \lim_{x\downarrow t_k} f'(x)$ and $\lim_{x\uparrow t_k} f''(x) = \lim_{x\downarrow t_k} f''(x)$. Assume without loss of generality that $f'(x) = \exp(c_{k,1}x + c_{k,2})$ for all $x \in O_k$ and $k \in \mathbb{Z}$. These conditions impose the restrictions $(c_{k,1} - c_{k+1,1})t_k = c_{k+1,2} - c_{k,2}$ and $\log(c_{k,1}/c_{k+1,1}) + (c_{k,1} - c_{k+1,1})t_k = c_{k+1,2} - c_{k,2}$ which entails that $c_{k,1} = c_{k+1,1}$ and $c_{k,2} = c_{k+1,2}$. This proves that there exists $c_1, c_2 \in \mathbb{R}$ such that $f'(x) = \exp(c_1x + c_2)$ for all $x \in \mathbb{R}$. Thus, the differential equation holds for all $x \in \mathbb{R}$,

$$0 = \xi'''(x) - \xi''(x)\frac{f''(x)}{f'(x)} - \frac{2f''(x)f'(x)}{\sigma^2} = \frac{\partial}{\partial x}\left(\frac{\xi''(x)}{f'(x)}\right) - 2\frac{f''(x)}{\sigma^2},$$

by division with $f'(x)$. By integration this implies that $0 = \xi''(x)/f'(x) - 2f'(x)/\sigma^2 + c_3$ such that $\xi''(x) = 2\exp(2c_1x+2c_2)/\sigma^2 - c_3\exp(c_1x+c_2)$ and $\xi'(x) = \exp(2c_1x+2c_2)/c_1\sigma^2 - c_3\exp(c_1x+c_2)/c_1 + c_4$ and

$$\xi(x) = \frac{\exp(2c_1x+2c_2)}{2c_1^2\sigma^2} - \frac{c_3\exp(c_1x+c_2)}{c_1^2} + c_4x + c_5.$$

We see that $\xi(x) \to \infty \iff p_{X_{\mathrm{pa}_{\mathcal{G}(i)}}}(x) \to \infty$ as $x \to \mathrm{sign}(c_1)\cdot\infty$, in contradiction with the assumption that $p_{X_{\mathrm{pa}_{\mathcal{G}(i)}}}(x)$ is a probability density function if $c_1 \neq 0$. Thus, it must hold that $f''(x)/f'(x) = 0$ for all $x \in \mathbb{R}$, or equivalently, that f is a linear function, yielding a contradiction.

This proves that whenever $f_i \in \mathcal{D}_3$ is a nowhere constant and nonlinear function and the additive noise is Gaussian then condition (ii) of Definition 2 is satisfied, so $\theta \in \Theta_R$. ■

Proof of Proposition 4. First, we consider the bivariate setting. Let (X, Y) be generated by an additive noise SCM $\theta \in \Theta_R \subset \mathcal{T}_2 \times \mathcal{D}_3^2 \times \mathcal{P}_{\mathcal{C}_3}^2$ given by $X := N_X$ and $Y := f(X) + N_Y$ with $P_X = p_X \cdot \lambda$ and $P_{N_Y} = p_{N_Y} \cdot \lambda$ having three times differentiable strictly positive densities and f is a three times differentiable nowhere constant function such that condition (ii) of Definition 2 holds.

Assume for contradiction that we do not have observational identifiability of the causal structure $\mathcal{G} = (V = \{X, Y\}, \mathcal{E} = \{(X \to Y)\})$. That is, there exists $\tilde{\theta} \in \mathcal{T}_2 \times \mathcal{D}_1^p \times \mathcal{P}_{\mathcal{C}_0}^p$ with causal graph $\tilde{\mathcal{G}} \neq \mathcal{G}$ or, equivalently, a differentiable function g and noise distributions $P_{\tilde{N}_X} = p_{\tilde{N}_X} \cdot \lambda$ and $P_{\tilde{N}_Y} = p_{\tilde{N}_Y} \cdot \lambda$ with continuous densities such that the structural assignments $\tilde{Y} := \tilde{N}_Y$ and $\tilde{X} := g(\tilde{Y}) + \tilde{N}_X$ induce the same distribution, i.e.,

$$P_{X,Y} = P_{\tilde{X},\tilde{Y}}. \tag{19}$$

By the additive noise structural assignments we know that both $P_{X,Y}$ and $P_{\tilde{X},\tilde{Y}}$ have densities with respect to λ^2 given by

$$p_{X,Y}(x,y) = p_X(x)p_{N_Y}(y - f(x)),$$
$$p_{\tilde{X},\tilde{Y}}(x,y) = p_{\tilde{N}_X}(x - g(y))p_{\tilde{Y}}(y),$$

for all $(x, y) \in \mathbb{R}^2$. By the equality of distributions in Equation (19) and strict positivity of p_X and p_{N_Y} we especially have that for λ^2-almost all $(x, y) \in \mathbb{R}^2$

$$0 < p_{X,Y}(x,y) = p_{\tilde{X},\tilde{Y}}(x,y). \tag{20}$$

However, as both $p_{X,Y}$ and $p_{\tilde{X},\tilde{Y}}$ are continuous we realize that the inequality in Equation (20) holds for all $(x, y) \in \mathbb{R}$ (if they were not everywhere equal there would exists a non-empty open ball in $\mathbb{R}^2$ on which they differ in contradiction with λ^2-almost everywhere equality). Furthermore, by the assumption that f is three times differentiable and p_X, p_{N_Y} are three times continuously differentiable we have that $\partial^3\pi/\partial x^3$ and $\partial^3\pi/\partial x^2\partial y$ are well-defined partial-derivatives of

$$\pi(x,y) := \log p_{X,Y}(x,y) = \log p_X(x) + \log p_{N_Y}(y - f(x)) =: \xi(x) + \nu(y - f(x)),$$

With $\tilde{\pi}(x,y) := \log p_{\tilde{X},\tilde{Y}}$ we have that

$$\tilde{\pi}(x,y) = \log p_{\tilde{N}_X}(x - g(y)) + \log p_{\tilde{Y}}(y) =: \tilde{\xi}(x - g(y)) + \tilde{\nu}(y).$$

Since it holds that $\pi = \tilde{\pi}$ by Equation (20) the partial-derivatives $\partial^3\tilde{\pi}/\partial x^3$ and $\partial^3\tilde{\pi}/\partial x^2\partial y$ are also well-defined. Now note that for any $x, y \in \mathbb{R}$

$$0 = \lim_{h\to 0} |\tilde{\pi}(x+h,y) - \tilde{\pi}(x,y)|/h = \lim_{h\to 0} |\tilde{\xi}(x - g(y) + h) - \tilde{\xi}(x - g(y))|/h,$$

implying that $\tilde{\xi}$ is differentiable in $x - g(y)$ for any $x, y \in \mathbb{R}$ or, equivalently, $\tilde{\xi}$ is everywhere differentiable. Similar arguments yield that $\tilde{\xi}$ is at least three times differentiable. We conclude that $\partial^2\tilde{\pi}(x,y)/\partial x^2 = \tilde{\xi}''(x - g(y))$ and $\partial^2\tilde{\pi}(x,y)/\partial x\partial y = -\tilde{\xi}''(x - g(y))g'(y)$ and for any $(x,y) \in \mathbb{R}^2$ such that $\partial^2\tilde{\pi}(x,y)/\partial x\partial y \neq 0$ or, equivalently,

$$\forall (x,y) \in \mathcal{J} := \left\{(x,y) : \frac{\partial^2\pi(x,y)}{\partial x\partial y} = -\nu''(y - f(x))f'(x) \neq 0\right\},$$

it holds that

$$\frac{\partial}{\partial x}\left(\frac{\frac{\partial^2}{\partial x^2}\tilde{\pi}(x,y)}{\frac{\partial^2}{\partial x\partial y}\tilde{\pi}(x,y)}\right) = \frac{\partial}{\partial x}\left(\frac{-1}{g'(y)}\right) = 0.$$

It is worth noting that $\mathcal{J} \neq \emptyset$ to ensure that the following derivations are not void of meaning. (This can be seen by noting that f is nowhere constant, i.e., $f'(x) \neq 0$ for λ-almost all $x \in \mathbb{R}$. Hence, $\mathcal{J} = \emptyset$ if and only if p_{N_Y} is a density such that $\{(x,y) \in \mathbb{R}^2 : f'(x) \neq 0\} \ni (x,y) \mapsto \nu''(y - f(x))$ is constantly zero or, equivalently, $\mathbb{R} \ni y \mapsto \nu''(y)$ is constantly zero. This holds if and only if p_{N_Y} is either exponentially decreasing or exponentially increasing everywhere, which is a contradiction as no continuously differentiable function integrating to one has this property.) For any $(x,y) \in \mathcal{J}$ we also have that

$$\begin{aligned} 0 = \frac{\partial}{\partial x}\left(\frac{\frac{\partial^2}{\partial x^2}\pi(x,y)}{\frac{\partial^2}{\partial x\partial y}\pi(x,y)}\right) &= \frac{\partial}{\partial x}\left(\frac{\xi''(x) + \nu''(y - f(x))f'(x)^2 - \nu'(y - f(x))f''(x)}{-\nu''(y - f(x))f'(x)}\right) \\ &= -2f'' + \frac{\nu' f'''}{\nu'' f'} - \frac{\xi'''}{\nu'' f'} + \frac{\nu'''\nu' f''}{(\nu'')^2} \\ &\quad - \frac{\nu'''\xi''}{(\nu'')^2} - \frac{(f'')^2\nu'}{\nu''(f')^2} + \frac{f''\xi''}{\nu''(f')^2}, \end{aligned}$$

which implies that

$$\xi''' = \xi''\left(\frac{f''}{f'} - \frac{f'\nu'''}{\nu''}\right) - 2\nu'' f'' f' + \nu' f''' + \frac{\nu'''\nu' f'' f'}{\nu''} - \frac{\nu'(f'')^2}{f'},$$

in contradiction with the assumption that condition (ii) of Definition 2 holds. We conclude that $P_{X,Y} \neq P_{\tilde{X},\tilde{Y}}$.

Now consider a multivariate restricted causal model $\theta \in \Theta_R$ over $X = (X_1, \ldots, X_p)$ with causal directed tree graph $\mathcal{G} = (V, \mathcal{E})$. Assume for contradiction that there exists an

alternative SCM $\tilde{\theta} = (\tilde{\mathcal{G}}, (\tilde{f}_i), P_{\tilde{N}}) \in \mathcal{T}_p \times \mathcal{D}_1^p \times \mathcal{P}_{\mathcal{C}_0}^p$ inducing $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_p)$ with causal graph $\tilde{\mathcal{G}} = (V, \tilde{\mathcal{E}}) \neq \mathcal{G}$, such that $P_X = P_{\tilde{X}}$.

Any SCM induced distribution is Markov with respect to the underlying causal graph. As such, we have that P_X is Markov with respect to both $\mathcal{G}$ and $\tilde{\mathcal{G}}$. Furthermore, since (in θ) the causal functions are non-constant and the noise innovations have strictly positive density, we have, by Proposition 17 of Peters et al. (2014), that P_X satisfies causal minimality with respect to causal graph $\mathcal{G}$ of θ, i.e., it is globally Markov with respect to $\mathcal{G}$ but not any proper subgraph of $\mathcal{G}$. If P_X also satisfies causal minimality with respect to $\tilde{\mathcal{G}}$, then, by Proposition 29 of Peters et al. (2014), there exist $i, j \in V$ such that $(j \to i) \in \mathcal{E}$ and $(i \to j) \in \tilde{\mathcal{E}}$.

> Assume for contradiction that P_X does not satisfy causal minimality with respect to $\tilde{\mathcal{G}}$. By Proposition 4 of Peters et al. (2014), we have that there exists $(j' \to i') \in \tilde{\mathcal{E}}$ such that $X_{j'} \perp\!\!\!\perp X_{i'}$. Define $A := \mathrm{nd}^{\tilde{\mathcal{G}}}(j') \cup \{j'\}$ and $B := \mathrm{de}^{\tilde{\mathcal{G}}}(i') \cup \{i'\}$. It holds that $A \perp\!\!\!\perp_{\tilde{\mathcal{G}}} (B \setminus \{i'\}) \,|\, i'$, i.e., A and $B \setminus \{i'\}$ are d-separated by i' in the directed tree $\tilde{\mathcal{G}}$. Since P_X is Markov with respect $\tilde{\mathcal{G}}$ it holds that $X_A \perp\!\!\!\perp X_{B\setminus\{i'\}} \,|\, X_{i'}$, hence $X_A \perp\!\!\!\perp X_B \,|\, X_{i'}$. Similarly, it holds that $X_A \perp\!\!\!\perp X_B \,|\, X_{j'}$ which implies that $X_A \perp\!\!\!\perp X_{i'} \,|\, X_{j'}$. By applying the contraction property of conditional independence, we get that
>
> $$\begin{aligned} &X_A \perp\!\!\!\perp X_{i'} \,|\, X_{j'} \quad \text{and} \quad X_{i'} \perp\!\!\!\perp X_{j'} \implies X_A \perp\!\!\!\perp X_{i'}, \text{ and} \\ &X_A \perp\!\!\!\perp X_B \,|\, X_{i'} \quad \text{and} \quad X_A \perp\!\!\!\perp X_{i'} \implies X_A \perp\!\!\!\perp X_B. \end{aligned}$$
>
> Since $A \cup B = V, A \cap B = \emptyset$ and $\mathcal{G}$ is a directed tree (that spans V) there exist either an edge $(j'' \to i'') \in \mathcal{E}$ with $j'' \in A$ and $i'' \in B$ or $j'' \in B$ and $i'' \in A$. In either case, we have that $X_{i''} \perp\!\!\!\perp X_{j''}$, which contradicts P_X satisfying causal minimality with respect to $\mathcal{G}$. We conclude that P_X also satisfies causal minimality with respect to the alternative graph $\tilde{\mathcal{G}}$.

Hence, the following structural equations hold for (X_i, X_j) and $(\tilde{X}_i, \tilde{X}_j)$

$$\begin{aligned} X_i &= f_i(X_j) + N_i, \quad \text{with} \quad X_j \perp\!\!\!\perp N_i, \\ \tilde{X}_j &= \tilde{f}_j(\tilde{X}_i) + \tilde{N}_j, \quad \text{with} \quad \tilde{X}_i \perp\!\!\!\perp \tilde{N}_j, \end{aligned}$$

with $P_{X_j, X_i} = P_{\tilde{X}_j, \tilde{X}_i}$. We can apply the same arguments as in the bivariate setup if we can argue that a density of X_j is three times differentiable and that a density of $\tilde{X}_i$ is a continuous density.

To this end, note that the density p_{X_j} is given by the convolution of two densities

$$p_{X_j}(y) = \int_{-\infty}^{\infty} p_{f_j(X_{\mathrm{pa}^{\mathcal{G}}(j)})}(t) p_{N_j}(y - t)\, dt, \tag{21}$$

as $X_j := f_j(X_{\mathrm{pa}^{\mathcal{G}}(j)}) + N_j$ with $X_{\mathrm{pa}^{\mathcal{G}}(j)} \perp\!\!\!\perp N_j$. Here we used that $f_j(X_{\mathrm{pa}^{\mathcal{G}}(j)})$ has density with respect to the Lebesgue measure.

To realize this note that $f_j \in \mathcal{C}_3$ and it is nowhere constant. By arguments similar to those in the proof of Lemma 3, this implies that $f'(x) = 0$ at only countably many points (d_k). Now let (O_k) be the countable collection of mutually disjoint open intervals that cover $\mathbb{R}$ except for the points (d_k). By continuity of f' we know that $f'(x)$ is either strictly positive or strictly negative on each O_k. That is, f is continuously differentiable and strictly monotone on each O_k. Thus, f has a continuously differentiable inverse on each O_k by, e.g., the inverse function theorem. This ensures that $f_j(X_{\mathrm{pa}_{\mathcal{G}}(j)})$ has a density with respect to the Lebesgue measure whenever $X_{\mathrm{pa}_{\mathcal{G}}(j)}$ does. By starting at the root node $X_{\mathrm{rt}(\mathcal{G})} = N_{\mathrm{rt}(\mathcal{G})}$, which by assumption has a density, we can iteratively apply the above argumentation down the directed path from $\mathrm{rt}(\mathcal{G})$ to j in order to conclude that any X_j for $j \in \{1, \dots, p\}$ has a density with respect to the Lebesgue measure.

Since p_{N_j} is assumed strictly positive three times continuous differentiable, the representation in Equation (21) furthermore yields that p_{X_j} is three times differentiable; see, e.g., Theorem 11.4 and 11.5 of Schilling (2017).

Now we argue that $\tilde{X}_i$ has a continuous density. First note that P_{X_i} at least has a continuous density p_{X_i} by arguments similar to those applied for Equation (21). By the assumption that $P_X = P_{\tilde{X}}$ we especially have that $P_{X_i} = P_{\tilde{X}_i}$ which implies that also $\tilde{X}_i$ has a continuous density. By virtue of the arguments for the bivariate setup we arrive at a contradiction, so it must hold that $P_X \neq P_{\tilde{X}}$. ■

Proof of Lemma 6. Consider an SCM $\tilde{\theta} = (\tilde{\mathcal{G}}, (\tilde{f}_i), P_{\tilde{N}}) \in \{\tilde{\mathcal{G}}\} \times \mathcal{D}_1^p \times \mathcal{P}_{\mathrm{G}}^p$ with $\tilde{\mathcal{G}} \neq \mathcal{G}$ and let $Q_{\tilde{\theta}}$ be the induced distribution. As $Q_{\tilde{\theta}}$ is Markov with respect to $\tilde{\mathcal{G}}$ and generated by an additive noise model the density $q_{\tilde{\theta}}$ factorizes as

$$q_{\tilde{\theta}}(x) = \prod_{i=1}^{p} q_{\tilde{\theta}}(x_i | x_{\mathrm{pa}_{\tilde{\mathcal{G}}}(i)}) = \prod_{i=1}^{p} q_{\tilde{N}_i}(x_i - \tilde{f}_i(x_{\mathrm{pa}_{\tilde{\mathcal{G}}}(i)})).$$

The cross entropy between P_X and $Q_{\tilde{\theta}}$ is then given by

$$\begin{aligned} h(P_X, Q_{\tilde{\theta}}) &:= \mathbb{E}\left[-\log\left(q_{\tilde{\theta}}(X)\right)\right] \\ &= \sum_{i=1}^{p} \mathbb{E}\left[-\log\left(q_{\tilde{N}_i}\left(X_i - \tilde{f}_i(X_{\mathrm{pa}_{\tilde{\mathcal{G}}}(i)})\right)\right)\right] \\ &= \sum_{i=1}^{p} h\left(X_i - \tilde{f}_i(X_{\mathrm{pa}_{\tilde{\mathcal{G}}}(i)}), \tilde{N}_i\right), \end{aligned}$$

where the latter is a sum of the cross entropies between the distribution of $X_i - \tilde{f}_i(X_{\mathrm{pa}_{\tilde{\mathcal{G}}}(i)})$ and the distribution of $\tilde{N}_i$. As $Q_{\tilde{\theta}}$ is generated by a causal additive tree model with Gaussian

noise, we have for all $1 \leq i \leq p$ that $\tilde{N}_i \sim \mathcal{N}(0, \tilde{\sigma}_i^2)$ for some $\tilde{\sigma}_i^2 > 0$. Hence for all $1 \leq i \leq p$,

$$\begin{aligned} h\left(X_i - \tilde{f}_i(X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}), \tilde{N}_i\right) &= \mathbb{E}\left[-\log\left(\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{\left(X_i - \tilde{f}_i(X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)})\right)^2}{2\tilde{\sigma}_i^2}\right)\right)\right] \\ &= \log(\sqrt{2\pi}\tilde{\sigma}_i) + \frac{\mathbb{E}\left[\left(X_i - \tilde{f}_i(X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)})\right)^2\right]}{2\tilde{\sigma}_i^2}. \end{aligned}$$

Thus, for given set of causal functions $(\tilde{f}_i)$ and a fixed i, the noise variance that minimizes the cross entropy is given by

$$\tilde{\sigma}_i = \sqrt{\mathbb{E}\left[\left(X_i - \tilde{f}_i(X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)})\right)^2\right]}.$$

We thus have

$$\begin{aligned} &\inf_{\tilde{\sigma}_i > 0}\left\{\log(\sqrt{2\pi}\tilde{\sigma}_i) + \frac{\mathbb{E}\left[\left(X_i - \tilde{f}_i(X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)})\right)^2\right]}{2\tilde{\sigma}_i^2}\right\} \\ &= \log\left(\sqrt{2\pi}\right) + \frac{1}{2}\log\left(\mathbb{E}\left[\left(X_i - \tilde{f}_i(X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)})\right)^2\right]\right) + \frac{1}{2}. \end{aligned}$$

We conclude that

$$\begin{aligned} &\inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{D}_1^p \times \mathcal{P}_{\mathrm{G}}^p} h(P_X, Q) \\ &= p\log(\sqrt{2\pi}) + \frac{p}{2} + \sum_{i=1}^{p}\frac{1}{2}\log\left(\inf_{\tilde{f}_i \in \mathcal{D}_1}\mathbb{E}\left[\left(X_i - \tilde{f}_i(X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)})\right)^2\right]\right). \end{aligned}$$

Finally, as $\mathcal{D}_1$ is dense in $\mathcal{L}^2(P_{X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}})$, we have that

$$\begin{aligned} \inf_{\tilde{f}_i \in \mathcal{D}_1}\mathbb{E}\left[\left(X_i - \tilde{f}_i(X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)})\right)^2\right] &= \mathbb{E}\left[\left(X_i - \mathbb{E}[X_i | X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}]\right)^2\right] \\ &\quad + \inf_{\tilde{f}_i \in \mathcal{D}_1}\mathbb{E}\left[\left(\mathbb{E}[X_i | X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}] - \tilde{f}_i(X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)})\right)^2\right] \\ &= \mathbb{E}\left[\left(X_i - \mathbb{E}[X_i | X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}]\right)^2\right]. \end{aligned}$$

Here we used that $X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}$ has density with respect to the Lebesgue measure, $P_{X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}} \ll \lambda$, and that the density is differentiable (see proof of Proposition 4). This concludes the first part of the proof.

For the second statement, we note that for any $Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p$ there exists some noise innovation distribution $P_{\tilde{N}} \in \mathcal{P}$ such that Q is the distribution of $\tilde{X}$ generated by structural assignments

$$\tilde{X}_i := \tilde{f}_i(X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}) + \tilde{N}_i = \mathbb{E}[X_i | X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}] + \tilde{N}_i,$$

for all $1 \leq j \leq p$ and mutually independent noise innovations $\tilde{N} = (\tilde{N}_1, \ldots, \tilde{N}_p) \sim P_{\tilde{N}} \in \mathcal{P}^p$. Let q denote the density of Q with respect to the Lebesgue measure and let $q_{\tilde{N}_i}$ denote the density of $\tilde{N}_i$ for all $1 \leq i \leq p$. As Q is Markov with respect to $\tilde{\mathcal{G}}$ and generated by an additive noise model the density factorizes as

$$q(x) = \prod_{i=1}^{p} q(x_i | x_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}) = \prod_{i=1}^{p} q_{\tilde{N}_i}(x_i - \mathbb{E}[X_i | X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)} = x_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}]).$$

The cross entropy between P_X and Q is given by

$$\begin{aligned}
h(P_X, Q) &= \mathbb{E}\left[-\log\left(q(X)\right)\right] \\
&= \sum_{i=1}^{p} \mathbb{E}\left[-\log\left(q(X_i | X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)})\right)\right] \\
&= \sum_{i=1}^{p} \mathbb{E}\left[-\log\left(q_{\tilde{N}_i}\left(X_i - \mathbb{E}\left[X_i | X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}\right]\right)\right)\right] \\
&= \sum_{i=1}^{p} h\left(X_i - \mathbb{E}\left[X_i | X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}\right], \tilde{N}_i\right).
\end{aligned}$$

Note that $h(P, Q) = h(P) + D_{\mathrm{KL}}(P \| Q) \geq h(P)$ with equality if and only if $Q = P$. Thus, the infimum is attained at noise innovations that are equal in distribution to $X_i - \mathbb{E}[X_i | X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}]$ (which has a density by assumption). That is,

$$\begin{aligned}
\inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p} h(P_X, Q) &= \sum_{i=1}^{p} \inf_{\tilde{N}_j \sim P_{\tilde{N}_j} \in \mathcal{P}} h\left(X_i - \mathbb{E}\left[X_i | X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}\right], \tilde{N}_i\right) \\
&= \sum_{i=1}^{p} h\left(X_i - \mathbb{E}\left[X_i | X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}\right]\right) \\
&= \ell_{\mathrm{E}}(\tilde{\mathcal{G}}).
\end{aligned}$$

■

Proof of Lemma 7. Let $\theta \in \Theta_R \subset \mathcal{T}_p \times \mathcal{D}_3^p \times \mathcal{P}_{\mathrm{G}}^p$ and assume that condition (a) is satisfied, i.e., that for all $i \neq j$ it holds that $x \mapsto \mathbb{E}[X_i | X_j = x]$ has a differentiable version. Note that

$$\ell_{\mathrm{G}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{G}}(\mathcal{G}) = \inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{D}_1^p \times \mathcal{P}_{\mathrm{G}}^p} h(P_X, Q) - h(P_X). \tag{22}$$

Furthermore, by the considerations in the proof of Lemma 6 the infimum in Equation (22) is attained for Q^*, where the functions are given by the conditional expectation functionals. When condition (a) is satisfied we therefore know that $Q^* \in \{\tilde{\mathcal{G}}\} \times \mathcal{D}_1^p \times \mathcal{P}_{\mathrm{G}}^p$. Finally,

$$\ell_{\mathrm{G}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{G}}(\mathcal{G}) = h(P_X, Q^*) - h(P_X) = D_{\mathrm{KL}}(P_X \,\|\, Q^*) > 0,$$

where the last strict inequality follows from Proposition 4.

Now let $\theta \in \Theta_R \subset \mathcal{T}_p \times \mathcal{D}_3^p \times \mathcal{P}_{\mathrm{G}}^p$. Assume that condition (b) is satisfied, i.e., for all $1 \leq i \leq p$ it holds that the causal function f_i is contained within a function class $\mathcal{F}_i \subseteq \mathcal{D}_1$, which for all $j \neq i$ satisfies

$$\operatorname*{arg\,min}_{\tilde{f}_i \in \mathcal{F}_i} \mathbb{E}\left[\left(X_i - \tilde{f}_i(X_j)\right)^2\right] \in \mathcal{F}_i. \tag{23}$$

Define the modified Gaussian score function

$$\ell_{\mathrm{G.mod}}(\tilde{\mathcal{G}}) := \sum_{i=1}^{p} \frac{1}{2} \log\left(\mathrm{Var}\left(X_i - f_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)i}(X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)})\right)\right),$$

where $f_{ji} : \mathbb{R} \to \mathbb{R}$ is given by

$$f_{ji} := \operatorname*{arg\,min}_{\tilde{f} \in \mathcal{F}_i} \mathbb{E}\left[\left(X_i - \tilde{f}(X_j)\right)^2\right],$$

for all $i \neq j$. Now, for any $\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}$, it holds that

$$\begin{aligned}
\ell_{\mathrm{G.mod}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{G.mod}}(\mathcal{G}) &= \inf_{Q \in \{\tilde{\mathcal{G}}\} \times (\mathcal{F}_i)_{1 \leq i \leq p} \times \mathcal{P}_{\mathrm{G}}^p} h(P_X, Q) - h(P_X) \\
&= h(P_X, Q^*) - h(P_X) \\
&= D_{\mathrm{KL}}(P_X \,\|\, Q^*) > 0.
\end{aligned}$$

Here we used the closedness in Equation (23) to argue that the infimum is attained for $Q^* \in \{\tilde{\mathcal{G}}\} \times (\mathcal{F}_i)_{1 \leq i \leq p} \times \mathcal{P}_{\mathrm{G}}^p$. Finally, since $(\mathcal{F}_i)_{1 \leq i \leq p} \subset \mathcal{D}_1^p$, Proposition 4 guarantees the strict inequality.

Now let $\theta \in \Theta_R \subset \mathcal{T}_p \times \mathcal{D}_3^p \times \mathcal{P}_{+\mathcal{C}_3}^p$. Assume that for all $i \neq j$ it holds that $x \mapsto \mathbb{E}[X_i | X_j = x]$ has a differentiable version, and assume that for all $i \neq j$ it holds that $X_i - \mathbb{E}[X_i|X_j]$ has a continuous density. With these assumptions we note that for any $\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}$ it holds, by the arguments in the proof of Lemma 6, that

$$\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) = \inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p} h(P_X, Q) - h(P_X) = h(P_X, Q^*) - h(P_X),$$

where Q^* is generated by an additive noise model $\tilde{\mathcal{G}} \times (f_i) \times (P_{\tilde{N}_i})_{1 \leq i \leq p}$ with causal graph $\tilde{\mathcal{G}} \in \mathcal{T}_p$, with causal functions $f_i \equiv x \mapsto \mathbb{E}[X_i | X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)} = x] \in \mathcal{D}_1$ and noise innovations given by $\tilde{N}_i \overset{\mathcal{D}}{=} X_i - \mathbb{E}[X_i|X_j] \sim P_{N_i} \in \mathcal{P}_{\mathcal{C}_0}$, i.e., noise innovations with continuous densities. Proposition 4 now yields that

$$\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) = D_{\mathrm{KL}}(P_X \,\|\, Q^*) > 0,$$

since P_X is induced by a restricted causal additive tree model and Q^* is induced by a causal additive tree model $\{\tilde{\mathcal{G}}\} \times (f_i) \times (P_{\tilde{N}_i})_{1 \leq i \leq p} \subset \mathcal{T}_p \times \mathcal{D}_1^p \times \mathcal{P}_{\mathcal{C}_0}^p$.

■

D.2 Proofs of Section 3

Proof of Theorem 8. Assume that $\theta = (\mathcal{G}, (f_i), P_N) \in \Theta_R$ with $P_N \in \mathcal{P}_{\mathrm{G}}^p$ and $\mathcal{G} = (V, \mathcal{E})$. For simplicity of the proof, we assume that $\mathbb{E}[X] = 0$ such that the Gaussian edge weight estimators simplify to

$$\hat{w}_{ji} := \hat{w}_{ji}^{\mathrm{G}} = \frac{1}{2} \log \left(\frac{\frac{1}{n} \sum_{k=1}^n (X_{k,i} - \hat{\varphi}_{ji}(X_{k,j}))^2}{\frac{1}{n} \sum_{k=1}^n X_{k,i}^2} \right),$$

for all $j \neq i$. Furthermore, define the Gaussian population (for $i \neq j$) and auxiliary (for $(j \to i) \notin \mathcal{E}$) edge weights by

$$w_{ji} := \frac{1}{2} \log \left(\frac{\mathbb{E}[(X_i - \varphi_{ji}(X_j))^2]}{\mathbb{E}[X_i^2]} \right), \quad w_{ji}^* := \frac{1}{2} \log \left(\frac{\mathbb{E}[(X_i - \tilde{\varphi}_{ji}(X_j))^2]}{\mathbb{E}[X_i^2]} \right),$$

respectively, where $\tilde{\varphi}_{ji} : \mathbb{R} \to \mathbb{R}$ is a fixed function satisfying $\mathbb{E}[(\hat{\varphi}_{ji}(X_j) - \tilde{\varphi}_{ji}(X_j))^2 | \tilde{\mathbf{X}}_n] \xrightarrow{P}_n 0$. Furthermore, for any $\tilde{\mathcal{G}} = (V, \tilde{\mathcal{E}}) \in \mathcal{T}_p$ denote

$$\hat{w}(\tilde{\mathcal{G}}) := \sum_{(j \to i) \in \tilde{\mathcal{E}}} \hat{w}_{ji}, \quad w(\tilde{\mathcal{G}}) := \sum_{(j \to i) \in \tilde{\mathcal{E}}} w_{ji}, \quad w^*(\tilde{\mathcal{G}}) := \sum_{(j \to i) \in \tilde{\mathcal{E}} \setminus \mathcal{E}} w_{ji}^* + \sum_{(j \to i) \in \tilde{\mathcal{E}} \cap \mathcal{E}} w_{ji},$$

as the total estimated, population and auxiliary edge weights for $\tilde{\mathcal{G}}$. As the conditional expectation minimizes the MSPE among measurable functions, i.e., $\varphi_{ji} = \arg\min_{f:\mathbb{R} \to \mathbb{R}} \mathbb{E}[(X_i - f(X_j))^2]$, we especially have, for any $i \neq j$, that

$$\mathbb{E}[(X_i - \tilde{\varphi}_{ji}(X_j))^2] \geq \mathbb{E}[(X_i - \varphi_{ji}(X_j))^2].$$

This construction entails, for any $\tilde{\mathcal{G}} \in \mathcal{T}_p$, that

$$w^*(\tilde{\mathcal{G}}) \geq w(\tilde{\mathcal{G}}), \quad \text{and} \quad w^*(\mathcal{G}) = w(\mathcal{G}). \tag{24}$$

Assumption 1 implies that there exists an $m > 0$ such that

$$\min_{\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}} \ell_{\mathrm{G}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{G}}(\mathcal{G}) = m > 0. \tag{25}$$

Thus, for any $\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}$ it holds that

$$\ell_{\mathrm{G}}(\mathcal{G}) + \frac{m}{2} \leq \ell_{\mathrm{G}}(\tilde{\mathcal{G}}) - \frac{m}{2}, \tag{26}$$

by the identifiability assumption of Equation (25). Now note that $\ell_{\mathrm{G}}(\tilde{\mathcal{G}}) = w(\tilde{\mathcal{G}}) + C$ with $C = \sum_{i=1}^p \log(\mathbb{E}[X_i^2])/2$ for all $\tilde{\mathcal{G}} \in \mathcal{T}_p$. Hence, we have, for all $\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}$, that

$$w^*(\mathcal{G}) - \frac{m}{2} = w(\mathcal{G}) + \frac{m}{2} \leq w(\tilde{\mathcal{G}}) - \frac{m}{2} \leq w^*(\tilde{\mathcal{G}}) - \frac{m}{2},$$

by the equality and inequalities in (26) and (24). Thus, we have that

$$P(\hat{\mathcal{G}} = \mathcal{G}) = P\left(\operatorname*{arg\,min}_{\tilde{\mathcal{G}}=(V,\tilde{\mathcal{E}})\in\mathcal{T}_p} \sum_{(j\to i)\in\tilde{\mathcal{E}}} \hat{w}_{ji} = \mathcal{G}\right)$$
$$\geq P\left(\bigcap_{\tilde{\mathcal{G}}\in\mathcal{T}_p} \left(|\hat{w}(\tilde{\mathcal{G}}) - w^*(\tilde{\mathcal{G}})| < \frac{m}{2}\right)\right).$$

We conclude that it suffices to show that

$$\sup_{\tilde{\mathcal{G}}\in\mathcal{T}_p} |\hat{w}(\tilde{\mathcal{G}}) - w^*(\tilde{\mathcal{G}})| \xrightarrow{P}_n 0.$$

To this end, let $\mathcal{E}^* := \{(j \to i) : i, j \in V, i \neq j\} \setminus \mathcal{E}$ and note that

$$\begin{aligned}
&\sup_{\tilde{\mathcal{G}}\in\mathcal{T}_p} |\hat{w}(\tilde{\mathcal{G}}) - w^*(\tilde{\mathcal{G}})| \\
&\leq \sup_{\tilde{\mathcal{G}}\in\mathcal{T}_p} \Bigg(\sum_{(j\to i)\in\tilde{\mathcal{E}}\setminus\mathcal{E}} \left|\hat{w}_{ji} - \frac{1}{2}\log\left(\frac{\mathbb{E}[(X_i - \tilde{\varphi}_{ji}(X_j))^2]}{\mathbb{E}[X_i^2]}\right)\right| \\
&\qquad + \sum_{(j\to i)\in\tilde{\mathcal{E}}\cap\mathcal{E}} \left|\hat{w}_{ji} - \frac{1}{2}\log\left(\frac{\mathbb{E}[(X_i - \varphi_{ji}(X_j))^2]}{\mathbb{E}[X_i^2]}\right)\right| \Bigg) \\
&\leq \sum_{(j\to i)\in\mathcal{E}^*} \left|\hat{w}_{ji} - \frac{1}{2}\log\left(\frac{\mathbb{E}[(X_i - \tilde{\varphi}_{ji}(X_j))^2]}{\mathbb{E}[X_i^2]}\right)\right| \\
&\qquad + \sum_{(j\to i)\in\mathcal{E}} \left|\hat{w}_{ji} - \frac{1}{2}\log\left(\frac{\mathbb{E}[(X_i - \varphi_{ji}(X_j))^2]}{\mathbb{E}[X_i^2]}\right)\right|. \qquad (27)
\end{aligned}$$

Now consider a fixed term $(j \to i) \in \mathcal{E}$ in the second sum of (27). We can upper bound the absolute difference by

$$\begin{aligned}
&\left|\hat{w}_{ji} - \frac{1}{2}\log\left(\frac{\mathbb{E}[(X_i - \varphi_{ji}(X_j))^2]}{\mathbb{E}[X_i^2]}\right)\right| \\
&\leq \frac{1}{2}\left|\log\left(\frac{1}{n}\sum_{k=1}^n (X_{k,i} - \hat{\varphi}_{ji}(X_{k,j}))^2\right) - \log\left(\mathbb{E}[(X_i - \varphi_{ji}(X_j))^2]\right)\right| \\
&\quad + \frac{1}{2}\left|\log(\mathbb{E}[X_i^2]) - \log\left(\frac{1}{n}\sum_{k=1}^n X_{k,i}^2\right)\right|. \qquad (28)
\end{aligned}$$

In the upper bound of (28), the last absolute difference vanishes in probability due to the law of large numbers and the continuous mapping theorem. The first absolute difference

also vanishes by the following arguments. Note that,

$$\begin{aligned}
0 &\leq \frac{1}{n}\sum_{k=1}^{n}\left(X_{k,i}-\hat{\varphi}_{ji}(X_{k,j})\right)^2 \\
&= \frac{1}{n}\sum_{k=1}^{n}\left(X_{k,i}-\varphi_{ji}(X_{k,j})\right)^2 + \frac{1}{n}\sum_{k=1}^{n}\left(\varphi_{ji}(X_{k,j})-\hat{\varphi}_{ji}(X_{k,j})\right)^2 \\
&\quad + \frac{2}{n}\sum_{k=1}^{n}\left(X_{k,i}-\varphi_{ji}(X_{k,j})\right)\left(\varphi_{ji}(X_{k,j})-\hat{\varphi}_{ji}(X_{k,j})\right).
\end{aligned}$$

Hence, it holds that

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{k=1}^{n}\left(X_{k,i}-\hat{\varphi}_{ji}(X_{k,j})\right)^2 - \frac{1}{n}\sum_{k=1}^{n}\left(X_{k,j}-\varphi_{ji}(X_{k,j})\right)^2\right| \\
=&\left|\frac{1}{n}\sum_{k=1}^{n}\left(\varphi_{ji}(X_{k,j})-\hat{\varphi}_{ji}(X_{k,j})\right)^2\right. \\
&\quad \left. + \frac{2}{n}\sum_{k=1}^{n}\left(X_{k,j}-\varphi_{ji}(X_{k,j})\right)\left(\varphi_{ji}(X_{k,j})-\hat{\varphi}_{ji}(X_{k,j})\right)\right| \\
\leq& \frac{1}{n}\sum_{k=1}^{n}\left(\varphi_{ji}(X_{k,j})-\hat{\varphi}_{ji}(X_{k,j})\right)^2 \\
&\quad + 2\sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(X_{k,j}-\varphi_{ji}(X_{k,j})\right)^2}\sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(\varphi_{ji}(X_{k,j})-\hat{\varphi}_{ji}(X_{k,j})\right)^2}, \qquad (29)
\end{aligned}$$

by Cauchy-Schwarz inequality. By the law of large numbers, we have that the first factor of the second term of (29) converges in probability to a constant,

$$\frac{1}{n}\sum_{k=1}^{n}\left(X_{k,j}-\varphi_{ji}(X_{k,j})\right)^2 \xrightarrow{P}_n \mathbb{E}[X_{1,i}-\varphi_{ji}(X_{1,j}))^2].$$

The first term and latter factor of the second term of Equation (29) vanish in probability by assumption. That is, for any $\varepsilon > 0$ we have that

$$\begin{aligned}
P\left(\left|\frac{1}{n}\sum_{k=1}\left(\varphi_{ji}(X_{k,j})-\hat{\varphi}_{ji}(X_{k,j})\right)^2\right| > \varepsilon\right) &= P\left(\left|\frac{1}{n}\sum_{k=1}\left(\varphi_{ji}(X_{k,j})-\hat{\varphi}_{ji}(X_{k,j})\right)^2\right| \wedge \varepsilon > \varepsilon\right) \\
&\leq \frac{\mathbb{E}\left[\left(\frac{1}{n}\sum_{k=1}^{n}\left(\varphi_{ji}(X_{k,j})-\hat{\varphi}_{ji}(X_{k,j})\right)^2\right)\wedge\varepsilon\right]}{\varepsilon} \\
&\leq \frac{\mathbb{E}\left[\mathbb{E}\left[\left(\varphi_{ji}(X_{1,j})-\hat{\varphi}_{ji}(X_{1,j})\right)^2 \middle| \tilde{\mathbf{X}}_n\right]\wedge\varepsilon\right]}{\varepsilon} \\
&\to_n 0,
\end{aligned}$$

using conditional Jensen's inequality ($x \mapsto \min(x, \varepsilon) = x \wedge \varepsilon$ is concave) and the dominated convergence theorem. This proves that

$$\frac{1}{n}\sum_{k=1}^{n}\left(X_{k,j} - \hat{\varphi}_{ji}(X_{k,j})\right)^2 \xrightarrow{P}_n \mathbb{E}[X_{1,i} - \varphi_{ji}(X_{1,j}))^2].$$

Thus, we have shown that the second term of (27) converges to zero in probability. Finally, the above arguments apply similarly to the first term of Equation (27) by exchanging every φ_{ji} with $\tilde{\varphi}_{ji}$. We have shown that $\sup_{\tilde{\mathcal{G}}\in\mathcal{T}_p} |\hat{w}(\tilde{\mathcal{G}}) - w^*(\tilde{\mathcal{G}})| \xrightarrow{P}_n 0$, which concludes the proof.

■

Proof of Theorem 9. Assume that for each sample size $n \in \mathbb{N}$ that $\theta_n = (\mathcal{G}, ...) \in \Theta_R$ with $\mathcal{G} = (V, \mathcal{E})$, additive Gaussian noise, and identifiability gap

$$\min_{\tilde{\mathcal{G}}\in\mathcal{T}_p\setminus\{\mathcal{G}\}} \ell_{\mathrm{G}}(\mathcal{G}) - \ell_{\mathrm{G}}(\tilde{\mathcal{G}}) = q_n > 0,$$

with $q_n^{-1} = o(\sqrt{n})$. For simplicity of the proof, we assume that $\mathbb{E}_{\theta_n}[X] = 0$ such that the edge weight estimators simplify to

$$\hat{w}_{ji} := \hat{w}^{\mathrm{G}}_{ji} = \hat{w}^{\mathrm{G}}_{ji}(\mathbf{X}_n, \tilde{\mathbf{X}}_n) = \frac{1}{2}\log\left(\frac{\frac{1}{n}\sum_{k=1}^{n}\left(X_{k,i} - \hat{\varphi}_{ji}(X_{k,j})\right)^2}{\frac{1}{n}\sum_{k=1}^{n} X_{k,i}^2}\right).$$

Furthermore, we continue with the notation and population quantities introduced in the proof of Theorem 8, i.e., $w_{ji} = \log(\mathbb{E}_{\theta_n}[(X_i - \mathbb{E}[X_i|X_j])^2])/\mathbb{E}_{\theta_n}[X_i^2])/2$, where we notionally have suppressed the dependence on n. We know that for each SCM θ_n it holds that

$$\ell_{\mathrm{G}}(\mathcal{G}) + q_n \leq \ell_{\mathrm{G}}(\tilde{\mathcal{G}}), \quad \text{hence} \quad w(\mathcal{G}) + q_n \leq w(\tilde{\mathcal{G}}),$$

for all $\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}$. Thus,

$$\begin{aligned}
&P_{\theta_n}\left(\operatorname*{arg\,min}_{\tilde{\mathcal{G}}=(V,\tilde{\mathcal{E}})\in\mathcal{T}_p} \sum_{(j\to i)\in\tilde{\mathcal{E}}} \hat{w}_{ji} = \mathcal{G}\right) \\
&\geq P_{\theta_n}\left(\left(|\hat{w}(\mathcal{G}) - w(\mathcal{G})| < \frac{q_n}{2}\right) \cap \bigcap_{\tilde{\mathcal{G}}\in\mathcal{T}_p\setminus\{\mathcal{G}\}} \left(\hat{w}(\tilde{\mathcal{G}}) - w(\tilde{\mathcal{G}}) \geq -\frac{q_n}{2}\right)\right).
\end{aligned}$$

For any $\tilde{\mathcal{G}} = (V, \tilde{\mathcal{E}}) \in \mathcal{T}_p$ we have that

$$\hat{w}(\tilde{\mathcal{G}}) - w(\tilde{\mathcal{G}}) = \sum_{(j\to i)\in\tilde{\mathcal{E}}\cap\mathcal{E}} \hat{w}_{ji} - w_{ji} + \sum_{(j\to i)\in\tilde{\mathcal{E}}\setminus\mathcal{E}} \hat{w}_{ji} - w_{ji},$$

where $\hat{w}_{ji}$ and w_{ji} denote the estimated and population Gaussian weights for the edge $(j \to i)$, respectively. Hence, it suffices to show that

$$\begin{aligned}
&\forall (j \to i) \in \mathcal{E}, \forall \varepsilon > 0 : P_{\theta_n}(|\hat{w}_{ji} - w_{ji}| < q_n\varepsilon) \to_n 1, \\
&\forall (j \to i) \notin \mathcal{E}, \forall \varepsilon > 0 : P_{\theta_n}\left(\hat{w}_{ji} - w_{ji} \geq -q_n\varepsilon\right) \to_n 1.
\end{aligned}$$

To see this, note that if the above statements hold, then

$$\begin{aligned} P_{\theta_n}\left(|\hat{w}(\mathcal{G}) - w(\mathcal{G})| < \frac{q_n}{2}\right) &\geq P_{\theta_n}\left(\sum_{(j\to i)\in\mathcal{E}} |\hat{w}_{ji} - w_{ji}| < \frac{q_n}{2}\right) \\ &\geq P_{\theta_n}\left(\bigcap_{(j\to i)\in\mathcal{E}} \left(|\hat{w}_{ji} - w_{ji}| < \frac{q_n}{2(p-1)}\right)\right) \\ &\to_n 1, \end{aligned}$$

and for any $\tilde{\mathcal{G}} = (V, \tilde{\mathcal{E}}) \in \mathcal{T}_p$

$$\begin{aligned} P_{\theta_n}\left(\hat{w}(\tilde{\mathcal{G}}) - w(\tilde{\mathcal{G}}) \geq -\frac{q_n}{2}\right) &= P_{\theta_n}\left(\sum_{(j\to i)\in\tilde{\mathcal{E}}\cap\mathcal{E}} \hat{w}_{ji} - w_{ji} + \sum_{(j\to i)\in\tilde{\mathcal{E}}\setminus\mathcal{E}} \hat{w}_{ji} - w_{ji} \geq -\frac{q_n}{2}\right) \\ &\geq P_{\theta_n}\left(\bigcap_{(j\to i)\in\tilde{\mathcal{E}}\cap\mathcal{E}} \left(|\hat{w}_{ji} - w_{ji}| \leq \frac{q_n}{2(p-1)}\right)\right. \\ &\qquad \left. \cap \bigcap_{(j\to i)\in\tilde{\mathcal{E}}\setminus\mathcal{E}} \left(\hat{w}_{ji} - w_{ji} \geq -\frac{q_n}{2(p-1)}\right)\right) \\ &\to_n 1, \end{aligned}$$

hence the probability of the intersections also converges to one.

The causal edges: Now fix $(j \to i) \in \mathcal{E}$. We want to show that for all $\varepsilon > 0$ it holds that

$$P_{\theta_n}(|\hat{w}_{ji} - w_{ji}| < q_n\varepsilon) \to_n 1.$$

First note that

$$\begin{aligned} |\hat{w}_{ji} - w_{ji}| \leq \frac{1}{2}\left|\log\left(\frac{1}{n}\sum_{k=1}^{n}\left(X_{k,i} - \hat{\varphi}_{ji}(X_{k,j})\right)^2\right) - \log\left(\mathbb{E}_{\theta_n}[(X_i - \varphi_{ji}(X_j))^2]\right)\right| \\ + \frac{1}{2}\left|\log(\mathbb{E}_{\theta_n}[X_i^2]) - \log\left(\frac{1}{n}\sum_{k=1}^{n} X_{k,i}^2\right)\right|, \end{aligned}$$

where $\hat{\varphi}_{ji}$ for each n is the estimated conditional expectation $x \mapsto \mathbb{E}_{\theta_n}[X_i|X_j = x]$ based on samples from the auxiliary data set. It suffices to show the desired convergence in probability for each of the above terms. Furthermore, for all sequences of positive random variables (Z_n) and positive constants $c > 0$ and for all $\varepsilon > 0$ there exists $\delta > 0$ such that

$$(q_n^{-1}|\log(Z_n) - \log(c)| \geq \varepsilon) \subseteq (q_n^{-1}|Z_n - c| \geq \delta),$$

for sufficiently large n. To see this, note that if $q_n^{-1}(\log(Z_n) - \log(c)) \geq \varepsilon$, then $Z_n > \exp(\log(c) + q_n\varepsilon) = c\exp(q_n\varepsilon) \geq c(1 + q_n\varepsilon)$, so $q_n^{-1}(Z_n - c) \geq c\varepsilon$. On the other hand, if $q_n^{-1}(\log(Z_n) - \log(c)) \leq -\varepsilon$, then $Z_n \leq c\exp(-\varepsilon q_n) \leq c(1 - \varepsilon q_n + \varepsilon^2 q_n^2)$, so $q_n^{-1}(Z_n - c) \leq$

$-c\varepsilon + c\varepsilon^2 q_n$. In summary, if $q_n^{-1}|\log(Z_n) - \log(c)| \geq \varepsilon$, then $q_n^{-1}|Z_n - c| \geq c\varepsilon - c\varepsilon^2 q_n > c\varepsilon(1-M) =: \delta$ where $1 > M > \varepsilon q_n$ for sufficiently large n. We conclude that it suffices to show that for all $\varepsilon > 0$ it holds that

$$P_{\theta_n}\left(\left|\frac{1}{n}\sum_{k=1}^{n}(X_{k,i} - \hat{\varphi}_{ji}(X_{k,j}))^2 - \mathbb{E}_{\theta_n}[(X_i - \varphi_{ji}(X_j))^2]\right| \geq q_n\varepsilon\right) \to_n 0 \tag{30}$$

and that

$$P_{\theta_n}\left(\left|\frac{1}{n}\sum_{k=1}^{n}X_{k,i}^2 - \mathbb{E}_{\theta_n}[X_i^2]\right| \geq q_n\varepsilon\right) \to_n 0, \tag{31}$$

Equation (31) is satisfied as the summands are mean zero i.i.d. Therefore, with

$$W_n := \frac{1}{n}\sum_{k=1}^{n}X_{k,i}^2 - \mathbb{E}_{\theta_n}[X_i^2],$$

where $\mathbb{E}_{\theta_n}[q_n^{-1}W_n] = 0$, we have that $\mathbb{E}_{\theta_n}[q_n^{-2}W_n^2] = \frac{q_n^{-2}}{n}\mathbb{E}_{\theta_n}[(X_i^2 - \mathbb{E}_{\theta_n}[X_i^2])^2]$, hence

$$\begin{aligned}
P_{\theta_n}(q_n^{-1}W_n \geq \varepsilon) &\leq q_n^{-2}\frac{\mathbb{E}_{\theta_n}[W_n^2]}{\varepsilon^2} \\
&\leq \frac{q_n^{-2}}{n}\frac{\sup_{n\in\mathbb{N}}\mathbb{E}_{\theta_n}[(X_i^2 - \mathbb{E}_{\theta_n}[X_i^2])^2]}{\varepsilon^2} \\
&\to_n 0,
\end{aligned}$$

for any $\varepsilon > 0$ as $\sup_{n\in\mathbb{N}}\mathbb{E}_{\theta_n}\|X\|_2^4 < \infty$ and $q_n^{-1} = o(\sqrt{n})$.

Now we show Equation (30). First, we simplify the notation by letting $Z_k := X_{k,i}$, $Y_k := X_{k,j}$ $f := \varphi_{ji}$ and $\hat{f} := \hat{\varphi}_{ji}$ for all $k \in \mathbb{N}$. Note that we have suppressed the dependence of $f = \varphi_{ji}$ on θ_n. We have that

$$\begin{aligned}
\frac{1}{n}\sum_{k=1}^{n}\left(Z_k - \hat{f}(Y_k)\right)^2 &= \frac{1}{n}\sum_{k=1}^{n}(Z_k - f(Y_k))^2 + \frac{1}{n}\sum_{k=1}^{n}(f(Y_k) - \hat{f}(Y_k))^2 \\
&\quad + \frac{2}{n}\sum_{k=1}^{n}(Z_k - f(Y_k))(f(Y_k) - \hat{f}(Y_k)) \\
&=: T_{1,n} + T_{2,n} + T_{3,n}.
\end{aligned}$$

It suffices to show that for all $\varepsilon > 0$ it holds that

(a) $P_{\theta_n}\left(|T_{1,n} - \mathbb{E}_{\theta_n}[(Z_1 - f(Y_1))^2]| \geq q_n\varepsilon\right) \to_n 0$,

(b) $P_{\theta_n}\left(|T_{2,n}| \geq q_n\varepsilon\right) \to_n 0$, and

(c) $P_{\theta_n}\left(|T_{3,n}| \geq q_n\varepsilon\right) \to_n 0$.

First we show (a). Each term in the sum of $T_{1,n} - \mathbb{E}_{\theta_n}[(Z_1 - f(Y_1))^2]$ is mean zero and i.i.d., i.e.,

$$q_n^{-1}\mathbb{E}_{\theta_n}[(Z_k - f(Y_k))^2 - \mathbb{E}_{\theta_n}[(Z_1 - f(Y_1))^2]] = 0.$$

Furthermore,

$$\begin{aligned}
&\mathrm{Var}_{\theta_n}(q_n^{-1}(T_{1,n} - \mathbb{E}_{\theta_n}[(Z_1 - f(Y_1))^2])) \\
=&\mathrm{Var}_{\theta_n}\left(\frac{q_n^{-1}}{n}\sum_{k=1}^{n}(Z_k - f(Y_k))^2 - \mathbb{E}_{\theta_n}[(Z_1 - f(Y_1))^2]\right) \\
=&\frac{q_n^{-2}}{n^2}\sum_{k=1}^{n}\mathrm{Var}_{\theta_n}\left((Z_k - f(Y_k))^2 - \mathbb{E}_{\theta_n}[(Z_1 - f(Y_1))^2]\right) \\
\leq&\frac{q_n^{-2}}{n}\sup_{n\in\mathbb{N}}\mathrm{Var}_{\theta_n}\left((Z_1 - f(Y_1))^2\right) \\
\to_n&0,
\end{aligned}$$

since $q_n^{-1} = o(\sqrt{n})$ and $\sup_{n\in\mathbb{N}}\mathbb{E}_{\theta_n}\|X\|_2^4 < \infty$. Hence,

$$\begin{aligned}
P_{\theta_n}\left(|q_n^{-1}(T_{1,n} - \mathbb{E}[(Z_1 - f(Y_1))^2])| \geq \varepsilon\right) &\leq \frac{\mathrm{Var}_{\theta_n}(q_n^{-1}(T_{,n} - \mathbb{E}[(Z_1 - f(Y_1))^2]))}{\varepsilon^2} \\
&\to_n 0.
\end{aligned}$$

by Chebyshev's inequality, proving (a).

Now we show (b). To that end, note that the terms of $T_{2,n}$ is i.i.d. conditional on $\tilde{\mathbf{X}}_n$. For a fixed $1 > \varepsilon > 0$ we have

$$\begin{aligned}
P_{\theta_n}\left(|q_n^{-1}T_{2,n}| \geq \varepsilon\right) &= \mathbb{E}_{\theta_n}\left[P_{\theta_n}\left(q_n^{-1}T_{2,n} \geq \varepsilon|\tilde{\mathbf{X}}_n\right) \wedge 1\right] \\
&\leq \frac{\mathbb{E}_{\theta_n}\left[\mathbb{E}_{\theta_n}\left[q_n^{-1}T_{2,n}|\tilde{\mathbf{X}}_n\right] \wedge 1\right]}{\varepsilon} \\
&= \frac{\mathbb{E}_{\theta_n}\left[q_n^{-1}\mathbb{E}_{\theta_n}\left[(f(Y_1) - \hat{f}(Y_1))^2|\tilde{\mathbf{X}}_n\right] \wedge 1\right]}{\varepsilon},
\end{aligned}$$

where we used the conditional Markov's inequality. Now fix $1 > \delta > 0$ and define $A_{n,\delta} := (q_n^{-1}\mathbb{E}_{\theta_n}\left[(f(Y_1) - \hat{f}(Y_1))^2|\tilde{\mathbf{X}}_n\right] > \delta)$ and note that by assumption there exists an $N_\delta \in \mathbb{N}$ such that $\forall n \geq N_\delta : P_{\theta_n}(A_{n,\delta}) < \delta$. Hence, for $n \geq N_\delta$ we have that

$$\begin{aligned}
\mathbb{E}_{\theta_n}\left[q_n^{-1}\mathbb{E}_{\theta_n}\left[(f(Y_1) - \hat{f}(Y_1))^2|\tilde{\mathbf{X}}_n\right] \wedge 1\right] &= \mathbb{E}_{\theta_n}\left[1_{A_{n,\delta}}q_n^{-1}\mathbb{E}_{\theta_n}\left[(f(Y_1) - \hat{f}(Y_1))^2|\tilde{\mathbf{X}}_n\right] \wedge 1\right] \\
&\quad + \mathbb{E}_{\theta_n}\left[1_{A_{n,\delta}^c}q_n^{-1}\mathbb{E}_{\theta_n}\left[(f(Y_1) - \hat{f}(Y_1))^2|\tilde{\mathbf{X}}_n\right] \wedge 1\right] \\
&\leq \mathbb{E}_{\theta_n}\left[1_{A_{n,\delta}}q_n^{-1}\mathbb{E}_{\theta_n}\left[(f(Y_1) - \hat{f}(Y_1))^2|\tilde{\mathbf{X}}_n\right] \wedge 1\right] \\
&\quad + \mathbb{E}_{\theta_n}\left[1_{A_{n,\delta}^c}\delta\right] \\
&\leq \mathbb{E}_{\theta_n}\left[1_{A_{n,\delta}}\right] + \delta \\
&= P_{\theta_n}(A_{n,\delta}) + \delta < 2\delta, \qquad (32)
\end{aligned}$$

hence $\limsup_{n\to\infty} P_{\theta_n}\left(|q_n^{-1}T_{2,n}| \geq \varepsilon\right) < 2\delta/\varepsilon$, i.e., $P_{\theta_n}\left(|q_n^{-1}T_{2,n}| \geq \varepsilon\right) \to 0$ as $\delta > 0$ was chosen arbitrarily, proving (b).

Now we prove (c). To this end, recall that

$$T_{3,n} := \frac{2}{n}\sum_{k=1}^{n}(Z_k - f(Y_k))(f(Y_k) - \hat{f}(Y_k)),$$

is, conditional on $\tilde{\mathbf{X}}$, an i.i.d. sum with conditional mean zero

$$\begin{aligned}\mathbb{E}_{\theta_n}[T_{3,n}|\tilde{\mathbf{X}}_n] &= 2\mathbb{E}_{\theta_n}[(Z_k - f(Y_k))(f(Y_k) - \hat{f}(Y_k))|\tilde{\mathbf{X}}_n] \\ &= 2\mathbb{E}_{\theta_n}[(\mathbb{E}_{\theta_n}[Z_k|Y_k, \tilde{\mathbf{X}}_n] - f(Y_k))(f(Y_k) - \hat{f}(Y_k))|\tilde{\mathbf{X}}_n] \\ &= 2\mathbb{E}_{\theta_n}[(f(Y_k) - f(Y_k))(f(Y_k) - \hat{f}(Y_k))|\tilde{\mathbf{X}}_n] = 0,\end{aligned}$$

and conditional second moment given by

$$\begin{aligned}\mathbb{E}_{\theta_n}[T_{3,n}^2|\tilde{\mathbf{X}}_n] &= \frac{4}{n^2}\sum_{k=1}^{n}\mathbb{E}_{\theta_n}[(Z_k - f(Y_k))^2(f(Y_k) - \hat{f}(Y_k))^2|\tilde{\mathbf{X}}_n] \\ &= \frac{4}{n}\mathbb{E}_{\theta_n}\left[(Z_k - f(Y_k))^2(f(Y_k) - \hat{f}(Y_k))^2|\tilde{\mathbf{X}}_n\right] \\ &= \frac{4}{n}\mathbb{E}_{\theta_n}\left[\mathbb{E}_{\theta_n}\left[(Z_k - f(Y_k))^2|\tilde{\mathbf{X}}_n, Y_k\right](f(Y_k) - \hat{f}(Y_k))^2|\tilde{\mathbf{X}}_n\right] \\ &= \frac{4}{n}\mathbb{E}_{\theta_n}\left[\mathrm{Var}_{\theta_n}(Z_k|Y_k)(f(Y_k) - \hat{f}(Y_k))^2|\tilde{\mathbf{X}}_n\right] \\ &\leq \frac{C}{n}\mathbb{E}_{\theta_n}\left[(f(Y_k) - \hat{f}(Y_k))^2|\tilde{\mathbf{X}}_n\right],\end{aligned}$$

P_{θ_n}-almost surely. Hence, w.l.o.g. assume that $0 < \varepsilon < 1$ and note that the conditional Markov's inequality yields

$$\begin{aligned}P_{\theta_n}(|q_n^{-1}T_{3,n}| \geq \varepsilon) &= \mathbb{E}_{\theta_n}[P_{\theta_n}(|q_n^{-1}T_{3,n}| \geq \varepsilon|\tilde{\mathbf{X}}_n) \wedge 1] \\ &\leq \frac{1}{\varepsilon^2}\mathbb{E}_{\theta_n}\left[\mathbb{E}_{\theta_n}\left[q_n^{-2}T_{3,n}^2|\tilde{\mathbf{X}}_n\right] \wedge 1\right] \\ &\leq \frac{C}{\varepsilon^2}\mathbb{E}_{\theta_n}\left[\frac{q_n^{-2}}{n}\mathbb{E}_{\theta_n}\left[(f(Y_k) - \hat{f}(Y_k))^2|\tilde{\mathbf{X}}_n\right] \wedge 1\right].\end{aligned} \tag{33}$$

By conditional Jensen's inequality, we have that

$$\begin{aligned}\mathbb{E}_{\theta_n}\left[(f(Y_k) - \hat{f}(Y_k))^2|\tilde{\mathbf{X}}_n\right] &\leq 1 + \mathbb{E}_{\theta_n}\left[(f(Y_k) - \hat{f}(Y_k))^2|\tilde{\mathbf{X}}_n\right]^2 \\ &\leq 1 + \mathbb{E}_{\theta_n}\left[(f(Y_k) - \hat{f}(Y_k))^4|\tilde{\mathbf{X}}_n\right].\end{aligned}$$

Fix $\delta > 0$ and let $A_{n,\delta} := \left(\frac{q_n^{-2}}{n}\mathbb{E}_{\theta_n}\left[(f(Y_k) - \hat{f}(Y_k))^4|\tilde{\mathbf{X}}_n\right] > \delta\right)$ and note that $P_{\theta_n}(A_{n,\delta}) \to_n 0$, hence there exists an $N_\delta \in \mathbb{N}$ such that $\forall n \geq N_\delta : P_{\theta_n}(A_{n,\delta}) < \delta$. Furthermore, as $q_n^{-1} = o(\sqrt{n})$ there exists an $N \in \mathbb{N}$ such that $q_n^{-2}/n < \delta$ for all $n \geq N$. Similar to the

arguments in Equation (32) we then have that

$$\begin{aligned}\frac{\varepsilon^2}{C}P_{\theta_n}(|q_n^{-1}T_{3,n}| \geq \varepsilon) &\leq \mathbb{E}_{\theta_n}\left[\frac{q_n^{-2}}{n}\left(1 + \mathbb{E}_{\theta_n}\left[(f(Y_k) - \hat{f}(Y_k))^4|\tilde{\mathbf{X}}_n\right]\right) \wedge 1\right] \\ &\leq \frac{q_n^{-2}}{n} + E_{\theta_n}\left[\frac{q_n^{-2}}{n}\mathbb{E}_{\theta_n}\left[(f(Y_k) - \hat{f}(Y_k))^4|\tilde{\mathbf{X}}_n\right] \wedge 1\right] \\ &\leq \frac{q_n^{-2}}{n} + \mathbb{E}_{\theta_n}[1_{A_{n,\delta}}] + \mathbb{E}_{\theta_n}[1_{A_{n,\delta}^c}\delta] \\ &< \delta + P_{\theta_n}(A_{n,\delta}) + \delta < 3\delta,\end{aligned}$$

for any $n \geq N_\delta \vee N$, so $P_{\theta_n}(q_n^{-1}T_{3,n} \geq \varepsilon) \to_n 0$, proving (c).

The non-causal edges: Now fix $(j \to i) \notin \mathcal{E}$, we want to show, for any $\varepsilon > 0$ that

$$P_{\theta_n}(\hat{w}_{ji} - w_{ji} \geq -q_n\varepsilon) \to_n 1,$$

where

$$\begin{aligned}\hat{w}_{ji} - w_{ji} = \frac{1}{2}\Bigg(&\left[\log\left(\frac{1}{n}\sum_{k=1}^n (X_{k,i} - \hat{\varphi}_{ji}(X_{k,j}))^2\right) - \log\left(\mathbb{E}[(X_i - \varphi_{ji}(X_j))^2]\right)\right] \\ &+ \left[\log(\mathbb{E}[X_i^2]) - \log\left(\frac{1}{n}\sum_{k=1}^n X_{k,i}^2\right)\right]\Bigg) =: \frac{1}{2}(D_{1,n} + D_{2,n}).\end{aligned}$$

We have that $P_{\theta_n}(\hat{w}_{ji} - w_{ji} \geq -q_n\varepsilon) \geq P_{\theta_n}\left((D_{1,n} \geq -q_n\varepsilon) \cap (|D_{2,n}| < q_n\varepsilon)\right)$, where the second event has already been shown to have probability converging to one in Equation (31). Thus, it suffices to show that

$$P_{\theta_n}\left(D_{1,n} \geq -q_n\varepsilon\right) \to_n 1.$$

By similar arguments as above we have for any sequence of positive random variables $(K_n)_{n\geq 1}$ and a positive constant K that for all $\varepsilon > 0$ there exists an $\delta > 0$ such that $P_{\theta_n}\left(\log(K_n) - \log(K) < -q_n\varepsilon\right) \leq P_{\theta_n}(K_n - K < -q_n\delta)$, for sufficiently large $n \in \mathbb{N}$. To see this, note that if $\log(K_n) - \log(K) < -q_n\varepsilon$, then $K_n < K\exp(-\varepsilon q_n) \leq K(1 - \varepsilon q_n + \varepsilon^2 q_n^2)$, so $q_n^{-1}(K_n - K) < -K\varepsilon + K\varepsilon^2 q_n < -K\varepsilon(1 - M) =: -\delta$ where $1 > M > \varepsilon q_n$ for sufficiently large n, since $q_n \downarrow 0$. Thus, it suffices to show that for any $\varepsilon > 0$ it holds that

$$P_{\theta_n}\left(\frac{1}{n}\sum_{k=1}^n (X_{k,i} - \hat{\varphi}_{ji}(X_{k,j}))^2 - \mathbb{E}_{\theta_n}[(X_i - \varphi_{ji}(X_j))^2] \geq -q_n\varepsilon\right) \to_n 1.$$

Again, we simplify the notation $Z_k := X_{k,i}$, $Y_k := X_{k,j}$, $f = \varphi_{ji}$ and $\hat{f} := \hat{\varphi}_{ji}$ for all $k \in \mathbb{N}$. Now define the following terms

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n}\left(Z_k - \hat{f}(Y_k)\right)^2 &= \frac{1}{n}\sum_{k=1}^{n}(Z_k - f(Y_k))^2 \\ &\quad + \frac{1}{n}\sum_{k=1}^{n}\{(f(Y_k) - \hat{f}(Y_k))^2 - \delta^2_{n,\theta_n}\} \\ &\quad + \frac{2}{n}\sum_{k=1}^{n}\{(Z_k - f(Y_k))(f(Y_k) - \hat{f}(Y_k)) + \delta^2_{n,\theta_n}/2\} \\ &=: T_{1,n} + \tilde{T}_{2,n} + \tilde{T}_{3,n},\end{aligned}$$

where $\delta^2_{n,\theta_n} := \mathbb{E}_{\theta_n}[(f(Y_1) - \hat{f}(Y_1))^2|\tilde{\mathbf{X}}_n] = \mathbb{E}_{\theta_n}[(\varphi_{ji}(X_j) - \hat{\varphi}_{ji}(X_j))^2|\tilde{\mathbf{X}}_n]$. It suffices to show that for all $\varepsilon > 0$ it holds that

(d) $P_{\theta_n}\left(|T_{1,n} - \mathbb{E}_{\theta_n}[(Z_1 - f(Y_1))^2]| \geq q_n\varepsilon\right) \to_n 0$,

(e) $P_{\theta_n}\left(|\tilde{T}_{2,n}| \geq q_n\varepsilon\right) \to_n 0$, and

(f) $P_{\theta_n}\left(\tilde{T}_{3,n} \geq -q_n\varepsilon\right) \to_n 1$.

Condition (d) holds by arguments similar to (a) for the causal edges.

Now we prove (e). The expansion, conditional on $\tilde{\mathbf{X}}_n$, is a sum of mean zero i.i.d. terms, hence

$$\begin{aligned}\mathbb{E}_{\theta_n}\left(q_n^{-2}\tilde{T}^2_{2,n}\middle|\,\tilde{\mathbf{X}}_n\right) &= \frac{q_n^{-2}}{n}\mathbb{E}_{\theta_n}\left[\{(f(Y_k) - \hat{f}(Y_k))^2 - \delta^2_{n,\theta_n}\}^2|\tilde{\mathbf{X}}_n\right] \\ &= \frac{q_n^{-2}}{n}\mathbb{E}_{\theta_n}\left[(f(Y_k) - \hat{f}(Y_k))^4 + (\delta^2_{n,\theta_n})^2 - 2(f(Y_k) - \hat{f}(Y_k))^2\delta^2_{n,\theta_n}|\tilde{\mathbf{X}}_n\right] \\ &= \frac{q_n^{-2}}{n}\left(\mathbb{E}_{\theta_n}\left[(f(Y_k) - \hat{f}(Y_k))^4|\tilde{\mathbf{X}}_n\right] - (\delta^2_{n,\theta_n})^2\right) \\ &\leq \frac{q_n^{-2}}{n}\mathbb{E}_{\theta_n}\left[(f(Y_k) - \hat{f}(Y_k))^4|\tilde{\mathbf{X}}_n\right],\end{aligned}$$

using that $(\delta^2_{n,\theta_n})^2 \geq 0$. Fix $1 > \delta > 0$ and let $A_{n,\delta} := \left(\frac{q_n^{-2}}{n}\mathbb{E}_{\theta_n}\left[(f(Y_k) - \hat{f}(Y_k))^4|\tilde{\mathbf{X}}_n\right] > \delta\right)$ and note that there exists an $N_\delta \in \mathbb{N}$ such that $\forall n \geq N_\delta : P_{\theta_n}(A_{n,\delta}) < \delta$. Similar to the previous arguments we have for any $1 > \varepsilon > 0$ and $n \geq N_\delta$ that

$$\begin{aligned}P_{\theta_n}\left(\left|\tilde{T}_{2,n}\right| \geq q_n\varepsilon\right) &= \mathbb{E}_{\theta_n}\left[P_{\theta_n}\left(\left|q_n^{-1}\tilde{T}_{2,n}\right| \geq \varepsilon\middle|\,\tilde{\mathbf{X}}_n\right) \wedge 1\right] \\ &\leq \frac{1}{\varepsilon^2}\mathbb{E}_{\theta_n}\left[\mathbb{E}_{\theta_n}\left[q_n^{-2}\tilde{T}^2_{2,n}|\tilde{\mathbf{X}}_n\right] \wedge 1\right] \\ &\leq \frac{1}{\varepsilon^2}\mathbb{E}_{\theta_n}\left[\frac{q_n^{-2}}{n}\mathbb{E}_{\theta_n}\left[(f(Y_k) - \hat{f}(Y_k))^4|\tilde{\mathbf{X}}_n\right] \wedge 1\right] \\ &\leq \frac{1}{\varepsilon^2}\left(\mathbb{E}_{\theta_n}[1_{A_{n,\delta}}] + \mathbb{E}_{\theta_n}[1_{A^c_{n,\delta}}\delta]\right) < \frac{2\delta}{\varepsilon^2},\end{aligned}$$

by the conditional Markov's inequality. Since $\delta > 0$ was chosen arbitrarily, we conclude that (e) holds.

Finally we show (f). Recall that in the analysis of the causal edges, we defined

$$T_{3,n} := \frac{2}{n}\sum_{k=1}^{n}(Z_k - f(Y_k))(f(Y_k) - \hat{f}(Y_k)).$$

Hence, we have that $\tilde{T}_{3,n} = T_{3,n} + \delta^2_{n,\theta_n}$. We realize that for any $0 < \varepsilon < 1$

$$\begin{aligned}
P_{\theta_n}(\tilde{T}_{3,n} < -q_n\varepsilon) &\leq P_{\theta_n}(T_{3,n} + \delta^2_{n,\theta_n} \leq -q_n\varepsilon) \\
&= P_{\theta_n}\left(T_{3,n} \leq -\left(q_n\varepsilon + \delta^2_{n,\theta_n}\right)\right) \\
&\leq P_{\theta_n}\left(T^2_{3,n} \geq \left(q_n\varepsilon + \delta^2_{n,\theta_n}\right)^2\right) \\
&\leq P_{\theta_n}\left(T^2_{3,n} \geq (q_n\varepsilon)^2\right) \\
&= P_{\theta_n}\left(q_n^{-2}T^2_{3,n} \geq \varepsilon^2\right) \\
&= \mathbb{E}_{\theta_n}\left[P_{\theta_n}\left(q_n^{-2}T^2_{3,n} \geq \varepsilon^2 | \tilde{\mathbf{X}}_n\right) \wedge 1\right] \\
&\leq \frac{1}{\varepsilon^2}\mathbb{E}_{\theta_n}\left[\mathbb{E}_{\theta_n}\left[q_n^{-2}T^2_{3,n}\middle|\, \tilde{\mathbf{X}}_n\right] \wedge 1\right] \\
&\to_n 0,
\end{aligned}$$

where we used the convergence shown in the proof of (c); see Equation (33). To see that the former arguments apply to non-causal edges, simply note that they did not use any conditions restricted to causal edges. This concludes the proof.

■

D.3 Proofs of Section 4

Lemma D.1 *Consider an i.i.d. sequence $(X_m)_{m\geq 1}$ of random variables with $X_m \in \mathbb{R}^d$ independent from a random infinite sequence $\tilde{\mathbf{X}} \in \prod_{i=1}^{\infty}\mathbb{R}^d$. Let $(\psi_n)_{n\geq 1}$ be a sequence of measurable functions s.t. for all $n \geq 1$, $\psi_n : \mathbb{R}^d \times (\prod_{i=1}^{\infty}\mathbb{R}^d) \to \mathbb{R}^q$ satisfies the following conditions:*

(a) $\mathbb{E}[\psi_n(X_m, \tilde{\mathbf{X}})|\tilde{\mathbf{X}}] = 0$ *almost surely,*

(b) $\exists\, \Sigma \in \mathbb{R}^{q\times q} : \sum_{m=1}^{n} \mathrm{Var}(\psi_n(X_m, \tilde{\mathbf{X}})|\tilde{\mathbf{X}}) \xrightarrow{P}_n \Sigma$, *and*

(c) $\exists\, \varepsilon > 0 : \sum_{m=1}^{n} \mathbb{E}[\|\psi_n(X_m, \tilde{\mathbf{X}})\|_2^{2+\varepsilon}|\tilde{\mathbf{X}}] \xrightarrow{P}_n 0$.

It holds that

$$\sum_{m=1}^{n}\psi_n(X_m, \tilde{\mathbf{X}}) \xrightarrow{\mathcal{D}}_n \mathcal{N}(0, \Sigma),$$

Proof of Lemma D.1. Let the random sequences be defined on a common probability space $(\Omega, \mathbb{F}, P)$ and define

$$\begin{aligned} A_{nm} &:= \mathbb{E}[\psi_n(X_m, \tilde{\mathbf{X}})|\tilde{\mathbf{X}}], \\ B_n &:= \Sigma - \sum_{m=1}^{n} \mathrm{Var}(\psi_n(X_m, \tilde{\mathbf{X}})|\tilde{\mathbf{X}}), \\ C_n &:= \sum_{m=1}^{n} \mathbb{E}[\|\psi_n(X_m, \tilde{\mathbf{X}})\|_2^{2+\varepsilon}|\tilde{\mathbf{X}}]. \end{aligned}$$

By assumption we have that $P(\cap_{n,m}(A_{nm} = 0)) = 1, B_n \xrightarrow{P} 0$ and $C_n \xrightarrow{P} 0$ as $n \to \infty$. First, note that for any subsequence $(n_k)_{k\geq 1}$ of the positive integers, there exists a subsequence $(n_{k_l})_{l\in\mathbb{N}}$ such that

$$P(\lim_{l\to\infty} B_{n_{k_l}} = 0) = 1 \quad \text{for} \quad (\lim_{l\to\infty} B_{n_{k_l}} = 0) := \{\omega \in \Omega : \lim_{l\to\infty} B_{n_{k_l}}(\omega) = 0\},$$

and

$$P(\lim_{l\to\infty} C_{n_{k_l}} = 0) = 1 \quad \text{for} \quad (\lim_{l\to\infty} C_{n_{k_l}} = 0) := \{\omega \in \Omega : \lim_{l\to\infty} C_{n_{k_l}}(\omega) = 0\}.$$

Thus, define

$$G := (\cap_{n,m}(A_{nm} = 0) \cap (\lim_{l\to\infty} B_{n_{k_l}} = 0) \cap (\lim_{l\to\infty} C_{n_{k_l}} = 0)) \subseteq \Omega, \quad \text{with} \quad P(G) = 1.$$

Now fix $\tilde{x} \in \tilde{\mathbf{X}}(G) := \{\tilde{\mathbf{X}}(\omega) \in \prod_{j=1}^{\infty} \mathbb{R}^d : \omega \in G\}$ and note that

$$\begin{aligned} &\forall l \geq 1, \forall 1 \leq m \leq n_{k_l} : \mathbb{E}[\psi_{n_{k_l}}(X_m, \tilde{x})] = 0, \\ &\sum_{m=1}^{n_{k_l}} \mathrm{Var}(\psi_{n_{k_l}}(X_m, \tilde{x})) \to_l \Sigma, \text{ and} \\ &\sum_{m=1}^{n_{k_l}} \mathbb{E}[\|\psi_{n_{k_l}}(X_m, \tilde{x})\|_2^{2+\varepsilon}] \to_l 0. \end{aligned}$$

Furthermore, for any $l \geq 1$

$$\psi_{n_{k_l}}(X_1, \tilde{x}), ..., \psi_{n_{k_l}}(X_{n_{k_l}}, \tilde{x}), \quad \text{are jointly independent,}$$

hence by Lyapunov's central limit theorem for triangular arrays (see, e.g., Van der Vaart, 2000, Proposition 2.27, and recall that Lyapunov's condition implies the Lindeberg–Feller condition) that

$$\sum_{m=1}^{n_{k_l}} \psi_{n_{k_l}}(X_m, \tilde{x}) \xrightarrow{\mathcal{D}}_l Z \sim \mathcal{N}(0, \Sigma).$$

The above convergence in distribution is equivalent to the following statement: for any continuous bounded function $g : \mathbb{R}^q \to \mathbb{R}$ it holds that

$$\lim_{l\to\infty} \mathbb{E}\left[g\left(\sum_{m=1}^{n_{k_l}} \psi_{n_{k_l}}(X_m, \tilde{x})\right)\right] = \mathbb{E}\left[g(Z)\right].$$

Fix a continuous and bounded g and note that the above convergence holds for all $\tilde{x} \in \tilde{\mathbf{X}}(G)$ with $P(G) = 1$. Thus, it must hold that

$$\mathbb{E}\left[g\left(\sum_{m=1}^{n_{k_l}} \psi_{n_{k_l}}(X_m, \tilde{\mathbf{X}})\right) \big| \tilde{\mathbf{X}}\right] \xrightarrow{a.s.}_l \mathbb{E}\left[g(Z)\right].$$

Finally, as $(n_{k_l})_{l\geq 1}$ is a subsequence of an arbitrary subsequence of positive integers, we have that

$$\mathbb{E}\left[g\left(\sum_{m=1}^{n} \psi_n(X_m, \tilde{x})\right) \big| \tilde{\mathbf{X}}\right] \xrightarrow{P}_n \mathbb{E}\left[g(Z)\right],$$

and since g is bounded the dominated convergence theorem yields that

$$\begin{aligned}
&\mathbb{E}\left[g\left(\sum_{m=1}^{n} \psi_n(X_m, \tilde{\mathbf{X}})\right)\right] \\
=&\mathbb{E}\left[\mathbb{E}\left[g\left(\sum_{m=1}^{n} \psi_n(X_m, \tilde{\mathbf{X}})\right) \big| \tilde{\mathbf{X}}\right]\right] \to_n \mathbb{E}\left[g(Z)\right].
\end{aligned}$$

As g was chosen arbitrarily, the above convergence holds for any continuous bounded g. We conclude that

$$\sum_{m=1}^{n} \psi_n(X_m, \tilde{\mathbf{X}}) \xrightarrow{\mathcal{D}}_n \mathcal{N}(0, \Sigma),$$

proving the theorem. ■

Lemma D.2 (Shah and Peters, 2020, Lemma 19) *Let $\mathcal{P}$ be a family of distributions for a random variable $\zeta \in \mathbb{R}$ and suppose $\zeta_1, \zeta_2, \ldots$ are i.i.d. copies of ζ. For each $n \in \mathbb{N}$ let $S_n = n^{-1}\sum_{i=1}^{n} \zeta_i$. Suppose that for all $P \in \mathcal{P}$ we have $\mathbb{E}_P(\zeta) = 0$ and $\mathbb{E}_P\left(|\zeta|^{1+\eta}\right) < c$ for some $\eta, c > 0$. We have that for all $\varepsilon > 0$,*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} P\left(|S_n| > \varepsilon\right) = 0.$$

Lemma D.3 *Let U be a random element and let $(Z_n)_{n\geq 1}$ be an i.i.d. sequence of random variables such that $U \perp\!\!\!\perp (Z_n)_{n\geq 1}$ and let $((W_{nm})_{m\leq n})_{n\geq 1}$ be a triangular array of random variables and $(g_n)_{n\geq 1}$ be measurable mappings with the following properties:*

(a) $\forall n \geq 1, \forall m \leq n : W_{nm} = g_n(Z_m, U)$,

(b) $\exists \eta > 0 : \mathbb{E}\left(|W_{n1}|^{1+\eta} \mid U\right) = O_p(1)$, *as* $n \to \infty$.

Then, writing $\bar{W}_n := \sum_{m=1}^{n} W_{nm}/n$, *we have*

$$\left|\bar{W}_n - \mathbb{E}\left(W_{n1} \mid U\right)\right| \xrightarrow{P}_n 0.$$

Proof of Lemma D.3. Denote

$$j_n(Z_m, U) := g_n(Z_m, U) - \mathbb{E}[g_n(Z_1, U)|U],$$

for any $n \geq 1$ and $m \leq n$. Let $\delta > 0$ be given. Pick $M > 0$ and $N \in \mathbb{N}$ such that the events

$$\Omega_n := \left\{\mathbb{E}\left[|g_n(Z_1, U)|^{1+\eta} \mid U\right] \leq M\right\},$$

satisfy $\mathbb{P}\left(\Omega_n^c\right) < \delta$ for $n \geq N$. Notice that

$$U(\Omega_n) = \left\{\tilde{u}_n : \mathbb{E}\left[|g_n(Z_1, \tilde{u}_n)|^{1+\eta}\right] \leq M\right\},$$

since $U \perp\!\!\!\perp (Z_n)_{n\geq 1}$. Fix $\varepsilon > 0$. Then, for all $n \geq N$

$$\begin{aligned} P\left(\left|\bar{W}_n - \mathbb{E}\left(W_n \mid U\right)\right| > \varepsilon\right) &= P\left(\left|\frac{1}{n}\sum_{m=1}^{n} j_n(Z_m, U)\right| > \varepsilon\right) \\ &< \mathbb{E}\left[P\left(\left|\frac{1}{n}\sum_{m=1}^{n} j_n(Z_m, U)\right| > \varepsilon \mid U\right) 1_{\Omega_n}\right] + \delta. \end{aligned}$$

By the dominated convergence theorem, the first term on the RHS converges to 0 if

$$\begin{aligned} &\sup_{\omega \in \Omega_n} P\left(\left|\frac{1}{n}\sum_{m=1}^{n} j_n(Z_m, U)\right| > \varepsilon \mid U\right)(\omega) \\ &= \sup_{\tilde{u}_n \in U(\Omega_n)} P\left(\left|\frac{1}{n}\sum_{m=1}^{n} j_n(Z_m, \tilde{u}_n)\right| > \varepsilon\right) \to_n 0, \end{aligned}$$

which implies the desired statement as $\delta > 0$ was chosen arbitrarily. Now note that for any $n \in \mathbb{N}, \tilde{u}_n \in U(\Omega_n)$ and all $m \in \mathbb{N}$ it holds that

$$\begin{aligned} \mathbb{E}[|j_n(Z_m, \tilde{u}_n)|^{1+\eta}] &= \mathbb{E}[|g_n(Z_m, \tilde{u}_n) - \mathbb{E}[g_n(Z_1, \tilde{u}_n)]|^{1+\eta}] \\ &\leq 2^{\eta}\left(\mathbb{E}[|g_n(Z_m, \tilde{u}_n)|^{1+\eta}] + |\mathbb{E}[g_n(Z_1, \tilde{u}_n)]|^{1+\eta}\right) \\ &\leq 2^{\eta}\left(\mathbb{E}[|g_n(Z_m, \tilde{u}_n)|^{1+\eta}] + \mathbb{E}[|g_n(Z_1, \tilde{u}_n)|^{1+\eta}]\right) \\ &< 2^{\eta+1} M =: c \end{aligned}$$

by the cr and Jensen's inequalities, and

$$\mathbb{E}[j_n(Z_m, \tilde{u}_n)] = 0.$$

For any $n \in \mathbb{N}$, define the following set of pushforward measures

$$\mathcal{P}_n := \{P' = (j_n(Z_1, \tilde{u}_n))(P) : \tilde{u}_n \in U(\Omega_n)\}.$$

For any $P' \in \mathcal{P}_n$, let $(Y_m)_{m\geq 1}$ be a sequence of i.i.d. random variables such that $Y_1 \overset{\mathcal{D}}{=} j_n(Z_1, \tilde{u}_n)$ for some $\tilde{u}_n \in U(\Omega_n)$. Notice that for all $n \in \mathbb{N}$ and $P' \in \mathcal{P}_n$ it holds that $\mathbb{E}_{P'}|Y_1|^{1+\eta} < c$ and $\mathbb{E}_{P'}[Y_1] = 0$. Thus,

$$\begin{aligned}\sup_{\tilde{u}_n \in U(\Omega_n)} P\left(\left|\frac{1}{n}\sum_{m=1}^{n} j_n(Z_m, \tilde{u}_n)\right| > \varepsilon\right) &= \sup_{P' \in \mathcal{P}_n} P'\left(\left|\frac{1}{n}\sum_{m=1}^{n} Y_m\right| > \varepsilon\right) \\ &\leq \sup_{P' \in \cup_k \mathcal{P}_k} P'\left(\left|\frac{1}{n}\sum_{m=1}^{n} Y_m\right| > \varepsilon\right) \\ &\to_n 0,\end{aligned}$$

by the weak uniform law of large numbers, Lemma D.2. ■

Lemma D.4 (Asymptotic normality of edge weight components) *Let for each sample size $n \in \mathbb{N}$, $\hat{\varphi}^n_{ji}$ denote the estimated conditional mean function φ_{ji} based on the auxiliary sample $\tilde{\mathbf{X}}_n$. For any $j \neq i$ and $m \leq n$, define*

$$\begin{aligned}&\hat{R}_{nm,ji} := \{X_{m,i} - \hat{\varphi}^n_{ji}(X_{m,j})\}, && \hat{\mu}_{n,ji} := \frac{1}{n}\sum_{m=1}^{n} \hat{R}^2_{nm,ji}, \\ &R_{m,ji} := \{X_{m,i} - \varphi_{ji}(X_{m,j})\}, && \mu_{ji} := \mathbb{E}[R^2_{1,ji}], \\ &\hat{V}_{m,i} := \left(X_{m,i} - \frac{1}{n}\sum_{k=1}^{n} X_{k,i}\right)^2, && \hat{\nu}_{n,i} := \frac{1}{n}\sum_{m=1}^{n} \hat{V}_{m,i}, \\ &\nu_i := \mathrm{Var}(X_{1,i}), && \delta^2_{n,ji} := \mathbb{E}[(\hat{\varphi}^n_{ji}(X_{1,j}) - \varphi_{ji}(X_{1,j}))^2 | \tilde{\mathbf{X}}_n].\end{aligned}$$

Let

$$\widehat{\Sigma}_n := \begin{bmatrix} \widehat{\Sigma}_{n,R} & \widehat{\Sigma}_{n,RV} \\ \widehat{\Sigma}^\intercal_{n,RV} & \widehat{\Sigma}_{n,V} \end{bmatrix} := \frac{1}{n}\sum_{m=1}^{n} \begin{bmatrix} \hat{R}^2_{nm}(\hat{R}^2_{nm})^\intercal - \hat{\mu}_n\hat{\mu}_n^\intercal & \hat{R}^2_{nm}\hat{V}_m^\intercal - \hat{\mu}_n\hat{\nu}_n^\intercal \\ \hat{V}_m(\hat{R}^2_{nm})^\intercal - \hat{\nu}_n\hat{\mu}_n^\intercal & \hat{V}_m\hat{V}_m^\intercal - \hat{\nu}_n\hat{\nu}_n^\intercal \end{bmatrix},$$

denote the $p^2 \times p^2$ matrix empirical covariance matrix, where the squaring of vectors means that each entry is squared. Suppose there exists $\xi > 0$ such that for all $j \neq i$, the following three conditions hold:

(i) $\mathbb{E}\|X\|^{4+\xi} < \infty$.

(ii) $\mathbb{E}[|\hat{\varphi}^n_{ji}(X_j) - \varphi_{ji}(X_j)|^{4+\xi} | \tilde{\mathbf{X}}_n] = O_p(1)$, *as* $n \to \infty$.

(iii) $\exists \Sigma \in \mathbb{R}^{p^2 \times p^2} : \mathrm{Var}\left(\begin{bmatrix} \hat{R}^2_{n1} - \delta^2_n - \mu \\ \hat{V}_1 - \nu \end{bmatrix} \middle| \tilde{\mathbf{X}}_n\right) \xrightarrow{P}_n \Sigma$, *where* Σ *is constant.*

Then we have that $\widehat{\Sigma}_n \xrightarrow{P} \Sigma \in \mathbb{R}^{p^2 \times p^2}$ *and*

$$\frac{1}{\sqrt{n}} \sum_{m=1}^{n} \begin{bmatrix} \hat{R}^2_{nm} - \delta^2_n - \mu \\ \hat{V}_m - \nu \end{bmatrix} = \sqrt{n} \begin{bmatrix} \hat{\mu}_n - \delta^2_n - \mu \\ \hat{\nu}_n - \nu \end{bmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma). \tag{34}$$

Proof of Lemma D.4. We prove the lemma under the assumption that $\mathbb{E}[X] = 0$ under which the variance estimator simplify to $\hat{V}_{m,i} := X^2_{m,i}$ and $\hat{\nu}_{n,i} := \frac{1}{n}\sum_{m=1}^{n} \hat{V}_{m,i}$ for all $1 \leq i \leq p$. The proof only gets more notionally cumbersome without this assumption. It should follow in all generality by applying expansion techniques and Slutsky's theorem similar to the standard arguments showing asymptotic normality of the regular sample variance.

Let $\tilde{\mathbf{X}}$ denote the auxilliary i.i.d. process such that $\tilde{\mathbf{X}}_n$ is the first n-coordinates of said process. Note that conditioning $\hat{\varphi}^n_{ji}$ on $\tilde{\mathbf{X}}$ it is equivalent to conditioning on $\tilde{\mathbf{X}}_n$ by the i.i.d. structure of $\tilde{\mathbf{X}}$ and that $\hat{\varphi}^n_{ji}$ only depends on $\tilde{\mathbf{X}}_n$. First, we define for all $j \neq i$, $n \in \mathbb{N}$ and $m \leq n$ the following conditional expectation regression error $\hat{\delta}_{nm,ji} := \{\varphi_{ji}(X_{m,j}) - \hat{\varphi}^n_{ji}(X_{m,j})\}$. Furthermore, for each $n \in \mathbb{N}$ and $m \leq n$ define

$$\Psi_n(X_m, \tilde{\mathbf{X}}) := \begin{bmatrix} \hat{R}^2_{nm} - \delta^2_n - \mu \\ \hat{V}_m - \nu \end{bmatrix} \in \mathbb{R}^{p^2},$$

where only $\tilde{\mathbf{X}}_n$ (containing the first n coordinates of $\tilde{\mathbf{X}}$) is used, and

$$\psi_n(X_m, \tilde{\mathbf{X}}) := \frac{1}{\sqrt{n}} \Psi_n(X_m, \tilde{\mathbf{X}}).$$

Note that the desired conclusion of Equation (34) follows by verifying condition (a), (b) and (c) of Lemma D.1. First, we show (a), the conditional mean zero condition. To that end, note that for any $i \in \{1, \ldots, p\}$ and $j \in \{1, \ldots, p\} \setminus \{i\}$ it holds that

$$\begin{aligned} \hat{R}^2_{nm,ji} &= (X_{m,i} - \varphi_{ji}(X_{m,j}) + \varphi_{ji}(X_{m,j}) - \hat{\varphi}^n_{ji}(X_{m,j}))^2 \\ &= (R_{m,ji} + \hat{\delta}_{nm,ji})^2 \\ &= R^2_{m,ji} + \hat{\delta}^2_{nm,ji} + 2R_{m,ji}\hat{\delta}_{nm,ji}. \end{aligned}$$

Hence, we have that

$$\hat{R}^2_{nm,ji} - \mu_{ji} - \delta^2_{n,ji} = (R^2_{m,ji} - \mu_{ji}) + (\hat{\delta}^2_{nm,ji} - \delta^2_{n,ji}) + 2R_{m,ji}\hat{\delta}_{nm,ji}. \tag{35}$$

The terms of Equation (35) are mean zero conditionally on $\tilde{\mathbf{X}}$, since $\mathbb{E}[R^2_{m,ji}|\tilde{\mathbf{X}}] = \mathbb{E}[R^2_{m,ji}] = \mu_{ji}$, $\mathbb{E}[\hat{\delta}^2_{nm,ji}|\tilde{\mathbf{X}}] = \delta^2_{n,ji}$ and

$$\begin{aligned} \mathbb{E}[R_{m,ji}\hat{\delta}_{nm,ji}|\tilde{\mathbf{X}}] &= \mathbb{E}[\mathbb{E}[R_{m,ji}\hat{\delta}_{nm,ji}|\tilde{\mathbf{X}}, X_{m,j}]|\tilde{\mathbf{X}}] \\ &= \mathbb{E}[\mathbb{E}[X_{m,i} - \varphi_{ji}(X_{m,j})|\tilde{\mathbf{X}}, X_{m,j}]\hat{\delta}_{nm,ji}|\tilde{\mathbf{X}}] \\ &= \mathbb{E}[(\mathbb{E}[X_{m,i}|X_{m,j}] - \varphi_{ji}(X_{m,j}))\hat{\delta}_{nm,ji}|\tilde{\mathbf{X}}] \\ &= 0, \end{aligned}$$

as $\varphi_{ji}(X_{m,j}) = \mathbb{E}[X_{m,i}|X_{m,j}]$ almost surely. Furthermore,

$$\mathbb{E}[X_{m,i}^2 - \mathrm{Var}(X_i)|\tilde{\mathbf{X}}] = \mathbb{E}[X_{m,i}^2] - \mathrm{Var}(X_i) = 0.$$

We conclude that

$$\mathbb{E}[\psi_n(X_m, \tilde{\mathbf{X}})|\tilde{\mathbf{X}}] = \frac{1}{\sqrt{n}}\mathbb{E}\left[\begin{bmatrix}\hat{R}_{nm}^2 - \delta_n^2 - \mu \\ \hat{V}_m - \nu\end{bmatrix}\middle|\tilde{\mathbf{X}}\right] = 0,$$

almost surely. With respect to (b), convergence of the sum of variances, we have, by assumption, that

$$\Sigma_n := \begin{bmatrix}\Sigma_{n,R} & \Sigma_{n,RV} \\ \Sigma_{n,RV}^\intercal & \Sigma_{n,V}\end{bmatrix} := \mathrm{Var}\left(\Psi_n(X_1, \tilde{\mathbf{X}})|\tilde{\mathbf{X}}\right) \xrightarrow{P}_n \Sigma,$$

where Σ is a positive semi-definite matrix. Furthermore, we have that $(X_m)_{m\geq 1}$ is an i.i.d. sequence independent of $\tilde{\mathbf{X}}$. Therefore,

$$\begin{aligned}\sum_{m=1}^n \mathrm{Var}(\psi_n(X_m, \tilde{\mathbf{X}})|\tilde{\mathbf{X}}) &= \sum_{m=1}^n \frac{1}{n}\mathrm{Var}(\Psi_n(X_m, \tilde{\mathbf{X}})|\tilde{\mathbf{X}}) \\ &= \sum_{m=1}^n \frac{1}{n}\Sigma_n \\ &= \Sigma_n \\ &\xrightarrow{P}_n \Sigma.\end{aligned}$$

Finally, we show that condition (c), a conditional Lindeberg-Feller condition, is fulfilled. To this end, note that with $\varepsilon := \xi/2 > 0$ we have that

$$\begin{aligned}&\mathbb{E}\left[\|\psi_n(X_m, \tilde{\mathbf{X}})\|_2^{2+\varepsilon}\middle|\tilde{\mathbf{X}}\right] \\ &= \mathbb{E}\left[\left\|\frac{1}{\sqrt{n}}\begin{bmatrix}\hat{R}_{nm}^2 - \delta_n^2 - \mu \\ \hat{V}_m - \nu\end{bmatrix}\right\|_2^{2+\varepsilon}\middle|\tilde{\mathbf{X}}\right] \\ &= \frac{1}{n^{\frac{2+\varepsilon}{2}}}\mathbb{E}\left[\left\|\begin{bmatrix}\hat{R}_{nm}^2 - \delta_n^2 - \mu \\ \hat{V}_m - \nu\end{bmatrix}\right\|_2^{2+\varepsilon}\middle|\tilde{\mathbf{X}}\right] \\ &\leq \frac{1}{n^{\frac{2+\varepsilon}{2}}}2^{(\frac{2+\varepsilon}{2}-1)}\Bigg(\sum_{i\neq j}\mathbb{E}\left[|\hat{R}_{nm,ji}^2 - \mu_{ji} - \delta_{n,ji}^2|^{2+\varepsilon}|\tilde{\mathbf{X}}\right] \\ &\qquad + \sum_{i=1}^p \mathbb{E}|X_{m,i}^2 - \mathrm{Var}(X_i)|^{2+\varepsilon}\Bigg),\end{aligned} \tag{36}$$

by the cr inequality. We now realize that the second factor of Equation (36) is stochastically bounded. To see this, note that for any $j \neq i$ it holds that

$$\mathbb{E}\left[|\hat{R}_{nm,ji}^2 - \mu_{ji} - \delta_{n,ji}^2|^{2+\varepsilon}|\tilde{\mathbf{X}}\right] \leq 2^{1+\varepsilon}(\mathbb{E}[|\hat{R}_{nm,ji}|^{4+2\varepsilon}|\tilde{\mathbf{X}}] + \mu_{ji}^{2+\varepsilon} + \mathbb{E}[|\delta_{n,ji}^2(\tilde{\mathbf{X}})|^{2+\varepsilon}|\tilde{\mathbf{X}}]). \tag{37}$$

The first term of the upper bound in Equation (37) is $O_p(1)$,

$$\begin{aligned}\mathbb{E}[|\hat{R}_{nm,ji}|^{4+2\varepsilon}|\tilde{\mathbf{X}}] &= \mathbb{E}[|X_{m,i} - \hat{\varphi}^n_{ji}(X_{m,j})|^{4+2\varepsilon}|\tilde{\mathbf{X}}] \\ &\leq 2^{3+2\varepsilon}(\mathbb{E}|X_{m,i} - \varphi_{ji}(X_{m,j})|^{4+2\varepsilon} + \mathbb{E}[|\varphi_{ji}(X_{m,i}) - \hat{\varphi}^n_{ji}(X_{m,j})|^{4+2\varepsilon}|\tilde{\mathbf{X}}]) \\ &= 2^{3+2\varepsilon}(\mathbb{E}[|R_{m,ji}|^{4+\xi}] + \mathbb{E}[|\hat{\delta}_{nm,ji}|^{4+\xi}|\tilde{\mathbf{X}}]) = O_p(1),\end{aligned}$$

as $\mathbb{E}\|X\|_2^{4+\xi} < \infty$ and $\mathbb{E}[|\hat{\delta}_{nm,ji}|^{4+\xi}|\tilde{\mathbf{X}}] = O_p(1)$. This holds because $R_{m,ji} = \{X_{m,i} - \mathbb{E}[X_{m,i}|X_{m,j}]\}$ and both terms are in $\mathcal{L}^{4+\xi}(P)$ if $X_{m,i} \in \mathcal{L}^{4+\xi}(P)$ which is guaranteed as $\mathbb{E}\|X\|_2^{4+\xi} < \infty$. For the third term in the upper bound of Equation (37), we note that by the conditional Jensen's inequality, we have that

$$\mathbb{E}[|\delta^2_{n,ji}|^{2+\varepsilon}|\tilde{\mathbf{X}}] \leq \mathbb{E}[|\varphi_{ji}(X_{m,i}) - \hat{\varphi}^n_{ji}(X_{m,j})|^{4+2\varepsilon}|\tilde{\mathbf{X}}] = \mathbb{E}[|\hat{\delta}_{nm,ji}|^{4+\xi}|\tilde{\mathbf{X}}] = O_p(1),$$

by assumption. Therefore, we have that

$$\sum_{m=1}^{n} \mathbb{E}\left[\|\Psi_n(X_m, \tilde{\mathbf{X}})\|_2^{2+\varepsilon}\middle|\tilde{\mathbf{X}}\right] \leq \frac{n}{n^{\frac{2+\varepsilon}{2}}} O_p(1) = n^{-\varepsilon/2} O_p(1) \xrightarrow{P}_n 0,$$

proving the conditional Lindeberg-Feller condition. By Lemma D.1 it holds that

$$\frac{1}{\sqrt{n}} \sum_{m=1}^{n} \psi_n(X_m, \tilde{\mathbf{X}}) \xrightarrow{\mathcal{D}}_n \mathcal{N}(0, \Sigma).$$

Now it only remains to prove that

$$\|\widehat{\Sigma}_n - \Sigma_n\| \xrightarrow{P} 0,$$

or, equivalently, that each entry converges to zero in probability. For example, for the entries of the first block matrix with $j \neq i$ and $l \neq r$ we prove that

$$|\widehat{\Sigma}_{n,R,ji,lr} - \Sigma_{n,R,ji,lr}| \xrightarrow{P} 0.$$

Now note that the observable estimated covariance matrix entry is given by

$$\widehat{\Sigma}_{n,R,ji,lr} = \frac{1}{n} \sum_{m=1}^{n} \hat{R}^2_{nm,ji} \hat{R}^2_{nm,lr} - \hat{\mu}_{n,ji} \hat{\mu}_{n,lr},$$

while the unobservable conditional covariance matrix is given by

$$\begin{aligned}\Sigma_{n,R,ji,lr} &= \mathbb{E}[(\hat{R}^2_{nm,ji} - \mu_{ji} - \delta^2_{n,ji})(\hat{R}^2_{nm,lr} - \mu_{lr} - \delta^2_{n,lr})|\tilde{\mathbf{X}}] \\ &= \mathbb{E}[\hat{R}^2_{nm,ji} \hat{R}^2_{nm,lr}|\tilde{\mathbf{X}}] - (\mu_{ji} + \delta^2_{n,ji})(\mu_{lr} + \delta^2_{n,lr}) \\ &= \mathbb{E}[\hat{R}^2_{nm,ji} \hat{R}^2_{nm,lr}|\tilde{\mathbf{X}}] - \mathbb{E}[\hat{R}^2_{nm,ji}|\tilde{\mathbf{X}}]\mathbb{E}[\hat{R}^2_{nm,lr}|\tilde{\mathbf{X}}],\end{aligned}$$

where we have used that $\mathbb{E}[\hat{R}^2_{nm,ji}|\tilde{\mathbf{X}}] = \mu_{ji} + \delta^2_{n,ji}$; see Equation (35) and its discussion. Note that the second term of the covariance matrix estimator expands to

$$\begin{aligned}
\hat{\mu}_{n,ji}\hat{\mu}_{n,lr} &= \left(\frac{1}{n}\sum_{m=1}^{n}\hat{R}^2_{nm,ji}\right)\left(\frac{1}{n}\sum_{m=1}^{n}\hat{R}^2_{nm,lr}\right) \\
&= \left(\frac{1}{n}\sum_{m=1}^{n}\hat{R}^2_{nm,ji} - \mathbb{E}[\hat{R}^2_{nm,ji}]\right)\left(\frac{1}{n}\sum_{m=1}^{n}\hat{R}^2_{nm,lr} - \mathbb{E}[\hat{R}^2_{nm,lr}]\right) \\
&\quad - \mathbb{E}[\hat{R}^2_{nm,ji}]\mathbb{E}[\hat{R}^2_{nm,lr}] \\
&\quad + \frac{1}{n}\sum_{m=1}^{n}\hat{R}^2_{nm,ji}\mathbb{E}[\hat{R}^2_{nm,lr}] \\
&\quad + \frac{1}{n}\sum_{m=1}^{n}\hat{R}^2_{nm,lr}\mathbb{E}[\hat{R}^2_{nm,ji}],
\end{aligned}$$

Thus

$$\begin{aligned}
&|\widehat{\Sigma}_{n,R,ji,lr} - \Sigma_{n,R,ji,lr}| \\
&= \Bigg|\frac{1}{n}\sum_{m=1}^{n}(\hat{R}^2_{nm,ji}\hat{R}^2_{nm,lr} - \mathbb{E}[\hat{R}^2_{nm,ji}\hat{R}^2_{nm,lr}|\tilde{\mathbf{X}}]) \\
&\quad - \left(\frac{1}{n}\sum_{m=1}^{n}\hat{R}^2_{nm,ji} - \mathbb{E}[\hat{R}^2_{nm,ji}|\tilde{\mathbf{X}}]\right)\left(\frac{1}{n}\sum_{m=1}^{n}\hat{R}^2_{nm,lr} - \mathbb{E}[\hat{R}^2_{nm,lr}|\tilde{\mathbf{X}}]\right) \\
&\quad - \frac{1}{n}\sum_{m=1}^{n}(\hat{R}^2_{nm,ji}\mathbb{E}[\hat{R}^2_{nm,lr}|\tilde{\mathbf{X}}] - \mathbb{E}[\hat{R}^2_{nm,ji}|\tilde{\mathbf{X}}]\mathbb{E}[\hat{R}^2_{nm,lr}|\tilde{\mathbf{X}}]) \\
&\quad - \frac{1}{n}\sum_{m=1}^{n}(\hat{R}^2_{nm,lr}\mathbb{E}[\hat{R}^2_{nm,ji}|\tilde{\mathbf{X}}] - \mathbb{E}[\hat{R}^2_{nm,ji}|\tilde{\mathbf{X}}]\mathbb{E}[\hat{R}^2_{nm,lr}|\tilde{\mathbf{X}}])\Bigg|. \qquad (38)
\end{aligned}$$

Each of these terms tends to zero in probability by Lemma D.3. For example, for the first term of Equation (38) it suffices to show that

$$\mathbb{E}\left[|\hat{R}^2_{nm,ji}\hat{R}^2_{nm,lr}|^{1+\varepsilon}|\tilde{\mathbf{X}}\right] = O_p(1),$$

for some $\varepsilon > 0$. Fix $\varepsilon = \xi/4$ and note, by the cr-inequality, that

$$\begin{aligned}
\hat{R}^2_{nm,ji}\hat{R}^2_{nm,lr} &= (X_{m,i} - \hat{\varphi}^n_{ji}(X_{m,j}))^2(X_{m,r} - \hat{\varphi}^n_{lr}(X_{m,l}))^2 \\
&\leq 4(R^2_{m,ji} + \hat{\delta}^2_{nm,ji})(R^2_{m,lr} + \hat{\delta}^2_{nm,lr}).
\end{aligned}$$

Thus, by the cr-inequality and the conditional Cauchy-Schwarz inequality we have, with $c = 4^{1+\varepsilon}2^{2\varepsilon}$, that

$$\begin{aligned}
&c^{-1}\mathbb{E}[|\hat{R}^2_{nm,ji}\hat{R}^2_{nm,lr}|^{1+\varepsilon}|\tilde{\mathbf{X}}] \\
&\leq c^{-1}4^{1+\varepsilon}\mathbb{E}[|R^2_{m,ji} + \hat{\delta}^2_{nm,ji}|^{1+\varepsilon}|R^2_{m,lr} + \hat{\delta}^2_{nm,lr}|^{1+\varepsilon}|\tilde{\mathbf{X}}] \\
&\leq \mathbb{E}[(|R_{m,ji}|^{2+2\varepsilon} + |\hat{\delta}_{nm,ji}|^{2+2\varepsilon})(|R_{m,lr}|^{2+2\varepsilon} + |\hat{\delta}_{nm,lr}|^{2+2\varepsilon})|\tilde{\mathbf{X}}] \\
&\leq \mathbb{E}[|R_{m,ji}|^{2+2\varepsilon}|R_{m,lr}|^{2+2\varepsilon}|\tilde{\mathbf{X}}] + \mathbb{E}[|R_{m,ji}|^{2+2\varepsilon}|\hat{\delta}_{nm,lr}|^{2+2\varepsilon}|\tilde{\mathbf{X}}] \\
&\quad + \mathbb{E}[|\hat{\delta}_{nm,ji}|^{2+2\varepsilon}|R_{m,lr}|^{2+2\varepsilon}|\tilde{\mathbf{X}}] + \mathbb{E}[|\hat{\delta}_{nm,ji}|^{2+2\varepsilon}|\hat{\delta}_{nm,lr}|^{2+2\varepsilon}|\tilde{\mathbf{X}}] \\
&\leq \mathbb{E}[|R_{m,ji}|^{4+\xi}]\mathbb{E}[|R_{m,lr}|^{4+\xi}] + \mathbb{E}[|R_{m,ji}|^{4+\xi}]\mathbb{E}[|\hat{\delta}_{nm,lr}|^{4+\xi}|\tilde{\mathbf{X}}] \\
&\quad + \mathbb{E}[|\hat{\delta}_{nm,ji}|^{4+\xi}|\tilde{\mathbf{X}}]\mathbb{E}[|R_{m,lr}|^{4+\xi}] + \mathbb{E}[|\hat{\delta}_{nm,ji}|^{4+\xi}|\tilde{\mathbf{X}}]\mathbb{E}[|\hat{\delta}_{nm,lr}|^{4+\xi}|\tilde{\mathbf{X}}] \\
&= O_p(1),
\end{aligned}$$

as $\mathbb{E}[|\hat{\delta}_{nm,ji}|^{4+\xi}|\tilde{\mathbf{X}}] = O_p(1)$ for all $j \neq i$ by assumption and $\mathbb{E}[|R_{m,ji}|^{4+\xi}] < \infty$ since $\mathbb{E}\|X\|_2^{4+\xi} < \infty$.

Similar arguments show convergence in probability of the entries in the other block submatrices of $\widehat{\Sigma}_n$ less Σ_n, yielding the desired conclusion.

■

Proof of Theorem 10. We prove the theorem under the simplifying assumption that $\mathbb{E}[X] = 0$ for which we can simplify the variance estimator by $\hat{V}_{m,i} := X^2_{m,i}$ and $\hat{\nu}_{n,i} := \frac{1}{n}\sum_{m=1}^n \hat{V}_{m,i}$ for all $1 \leq i \leq p$.

First, note (using the notation introduced in Lemma D.4) that $\hat{M}_1 = \{\hat{R}^2_{n1,ji}\}_{j\neq i}$, $\hat{\mu} = \hat{\mu}_n$, $\hat{\nu} = \hat{\nu}_n$ and $\widehat{\Sigma} = \widehat{\Sigma}_n$. The conditional mean of $\hat{M}_1$ given $\tilde{\mathbf{X}}_n$ is given by

$$\mathbb{E}[\hat{M}_1|\tilde{\mathbf{X}}_n] = \mathbb{E}[\{\hat{R}^2_{n1,ji}\}_{j\neq i}|\tilde{\mathbf{X}}_n] = \mu + \delta^2_n,$$

see Equation (35). Similarly we have that $\mathbb{E}[\hat{V}_1|\tilde{\mathbf{X}}_n] = \mathbb{E}[\hat{V}_1] = \nu$. Subtracting a constant (conditional on $\tilde{\mathbf{X}}_n$) does not change the conditional variance, hence

$$\mathrm{Var}\left(\begin{bmatrix}\hat{R}^2_{n1} - \delta^2_n - \mu \\ \hat{V}_1 - \nu\end{bmatrix}\middle|\tilde{\mathbf{X}}_n\right) = \mathrm{Var}\left((\hat{M}_1^\intercal, \hat{V}_1^\intercal)^\intercal\middle|\tilde{\mathbf{X}}_n\right) \xrightarrow{P}_n \Sigma.$$

Σ is constant and positive semi-definite with strictly positive diagonal. As such, the conditions of Lemma D.4 is satisfied, which yields that

$$\frac{1}{\sqrt{n}}\sum_{m=1}^n \begin{bmatrix}\hat{R}^2_{nm} - \delta^2_n - \mu \\ \hat{V}_m - \nu\end{bmatrix} = \sqrt{n}\begin{bmatrix}\hat{\mu} - \delta^2_n - \mu \\ \hat{\nu} - \nu\end{bmatrix} \xrightarrow{\mathcal{D}}_n \mathcal{N}(0, \Sigma), \tag{39}$$

and that

$$\widehat{\Sigma} = \begin{bmatrix}\widehat{\Sigma}_M & \widehat{\Sigma}_{MV} \\ \widehat{\Sigma}_{MV}^\intercal & \widehat{\Sigma}_V\end{bmatrix} \xrightarrow{P} \Sigma =: \begin{bmatrix}\Sigma_M & \Sigma_{MV} \\ \Sigma_{MV}^\intercal & \Sigma_V\end{bmatrix} \in \mathbb{R}^{p^2\times p^2}.$$

For any $j \neq i$ we denote

$$\hat{w}_{ji} := \frac{1}{2}\log\left(\frac{\hat{\mu}_{ji}}{\hat{\nu}_i}\right), \quad \tilde{w}_{ji} := \frac{1}{2}\log\left(\frac{\hat{\mu}_{ji} - \delta^2_{n,ji}}{\hat{\nu}_i}\right), \quad w_{ji} := \frac{1}{2}\log\left(\frac{\mu_{ji}}{\nu_i}\right),$$

where the latter is a shorthand notation for the Gaussian edge weight w^{G}_{ji}. Fix $\alpha \in (0,1)$. First, consider $(j \to i) \in \mathcal{E}$ and note that

$$\sqrt{n}\left(\begin{bmatrix}\hat{\mu}_{ji} - \mu_{ji} \\ \hat{\nu}_i - \nu_i\end{bmatrix} - \begin{bmatrix}\hat{\mu}_{ji} - \delta^2_{n,ji} - \mu_{ji} \\ \hat{\nu}_i - \nu_i\end{bmatrix}\right) = \sqrt{n}\begin{bmatrix}\delta^2_{n,ji} \\ 0\end{bmatrix} = \sqrt{n}\begin{bmatrix}\mathbb{E}[\hat{\delta}^2_{nm,ji}|\tilde{\mathbf{X}}_n] \\ 0\end{bmatrix} \xrightarrow{P}_n 0, \quad (40)$$

by assumption (iv). Hence, Equation (39), Equation (40) and the delta method yields that

$$\begin{aligned}\sqrt{n}\,(\hat{w}_{ji} - w_{ji}) &= \sqrt{n}\left(\log\left(\frac{\hat{\mu}_{ji}}{\hat{\nu}_i}\right) - \log\left(\frac{\mu_{ji}}{\nu_i}\right)\right) \\ &= \sqrt{n}(\log(\hat{\mu}_{ji}) - \log(\mu_{ji}) - \log(\hat{\nu}_i) + \log(\nu_i)) \\ &\xrightarrow{\mathcal{D}}_n \mathcal{N}(0, \sigma^2_{ji}),\end{aligned}$$

where

$$\hat{\sigma}^2_{ji} := \frac{\widehat{\Sigma}_{M,ji}}{\hat{\mu}^2_{ji}} + \frac{\widehat{\Sigma}_{V,i}}{\hat{\nu}^2_i} - 2\frac{\widehat{\Sigma}_{MV,ji,i}}{\hat{\mu}_{ji}\hat{\nu}_i} \xrightarrow{P} \sigma^2_{ji} := \frac{\Sigma_{M,ji}}{\mu^2_{ji}} + \frac{\Sigma_{V,i}}{\nu^2_i} - 2\frac{\Sigma_{MV,ji,i}}{\mu_{ji}\nu_i} \geq 0.$$

Here $\widehat{\Sigma}_{M,ji}$ and $\widehat{\Sigma}_{V,i}$ and their limits use a shorthand notation that denote the corresponding diagonal element, e.g., $\widehat{\Sigma}_{M,ji} := \widehat{\Sigma}_{M,ji,ji}$.

An asymptotically valid marginal confidence interval for w_{ji} with level α is, by virtue of the above convergence in distribution, given by

$$\hat{w}_{ji} \pm \hat{\sigma}_{ji}\frac{q(1-\frac{\alpha}{2})}{2\sqrt{n}},$$

where $q(1-\frac{\alpha}{2})$ is the $1-\alpha/2$ quantile of the standard normal distribution. That is,

$$P\left(\hat{w}_{ji} - \hat{\sigma}_{ji}\frac{q(1-\frac{\alpha}{2})}{2\sqrt{n}} \leq w_{ji} \leq \hat{w}_{ji} + \hat{\sigma}_{ji}\frac{q(1-\frac{\alpha}{2})}{2\sqrt{n}}\right) \longrightarrow_n 1-\alpha.$$

On the other hand, for any $(j \to i) \notin \mathcal{E}$ we have, by similar arguments, except that no assumption guarantees that $\sqrt{n}\delta^2_{n,ji}$ vanishes, that

$$P\left(\tilde{w}_{ji} - \tilde{\sigma}_{ji}\frac{q(1-\frac{\alpha}{2})}{2\sqrt{n}} \leq w_{ji} \leq \tilde{w}_{ji} + \tilde{\sigma}_{ji}\frac{q(1-\frac{\alpha}{2})}{2\sqrt{n}}\right) \longrightarrow_n 1-\alpha,$$

where

$$\begin{aligned}\tilde{\sigma}^2_{ji} &:= \frac{\widehat{\Sigma}_{M,ji}}{(\hat{\mu}_{ji} - \delta^2_{n,ji})^2} + \frac{\widehat{\Sigma}_{V,i}}{\hat{\nu}^2_i} - 2\frac{\widehat{\Sigma}_{MV,ji,i}}{(\hat{\mu}_{ji} - \delta^2_{n,ji})\hat{\nu}_i} \\ &\xrightarrow{P} \sigma^2_{ji} := \frac{\Sigma_{M,ji}}{\mu^2_{ji}} + \frac{\Sigma_{V,i}}{\nu^2_i} - 2\frac{\Sigma_{MV,ji,i}}{\mu_{ji}\nu_i} \geq 0,\end{aligned}$$

by the convergence in Equation (39). Note that $\tilde{\sigma}_{ji}^2$ is not observable since $\delta_{n,ji}^2$ is not observable. Now define

$$\hat{u}_{\alpha,ji},\, \hat{l}_{\alpha,ji} := \hat{w}_{ji} \pm \hat{\sigma}_{ji} \frac{q\left(1 - \frac{\alpha}{2p(p-1)}\right)}{2\sqrt{n}},$$
$$\tilde{u}_{\alpha,ji},\, \tilde{l}_{\alpha,ji} := \tilde{w}_{ji} \pm \tilde{\sigma}_{ji} \frac{q\left(1 - \frac{\alpha}{2p(p-1)}\right)}{2\sqrt{n}},$$

for all $j \neq i$. Thus, we have the following Bonferroni corrected simultaneous confidence interval for the Gaussian edge weights

$$\liminf_{n\to\infty} P\left(\bigcap_{(j\to i)\in\mathcal{E}} \left(w_{ji} \in \left[\hat{l}_{\alpha,ji}, \hat{u}_{\alpha,ji}\right]\right) \bigcap_{j\to i\notin\mathcal{E}} \left(w_{ji} \in \left[\tilde{l}_{\alpha,ji}, \tilde{u}_{\alpha,ji}\right]\right)\right) \geq 1 - \alpha.$$

The above confidence region has the correct asymptotic level, but it is infeasible to compute in that $\tilde{w}_{ji}$, $\tilde{\sigma}_{ji}$ and $\mathcal{E}$ are not directly observable from data. Furthermore, define

$$C(\hat{l}_\alpha, \tilde{l}_\alpha, \hat{u}_\alpha, \tilde{u}_\alpha) := \Bigg\{ \underset{\tilde{\mathcal{G}}=(V,\tilde{\mathcal{E}})\in\mathcal{T}_p}{\arg\min} \sum_{(j\to i)\in\tilde{\mathcal{E}}} w'_{ji} : \forall (j \to i) \in \mathcal{E}, w'_{ji} \in [\hat{l}_{\alpha,ji}, \hat{u}_{\alpha,ji}],$$
$$\forall (j \to i) \notin \mathcal{E}, w'_{ji} \in [\tilde{l}_{\alpha,ji}, \tilde{u}_{\alpha,ji}] \Bigg\},$$

and note that this is an unobservable confidence region for the causal graph. That is,

$$\begin{aligned}
&\liminf_{n\to\infty} P(\mathcal{G} \in C(\hat{l}_\alpha, \tilde{l}_\alpha, \hat{u}_\alpha, \tilde{u}_\alpha)) \\
&\geq \liminf_{n\to\infty} P\left(\bigcap_{(j\to i)\in\mathcal{E}} (w_{ji} \in [\hat{l}_{\alpha,ji}, \hat{u}_{\alpha,ji}]) \bigcap_{(j\to i)\notin\mathcal{E}} (w_{ji} \in [\tilde{l}_{\alpha,ji}, \tilde{u}_{\alpha,ji}])\right) \\
&\geq 1 - \alpha.
\end{aligned}$$

Our proposed confidence region has the form

$$\hat{C} := C(\hat{l}_\alpha, \hat{u}_\alpha) := \Bigg\{ \underset{\tilde{\mathcal{G}}=(V,\tilde{\mathcal{E}})\in\mathcal{T}_p}{\arg\min} \sum_{(j\to i)\in\tilde{\mathcal{E}}} w'_{ji} : \forall j \neq i, w'_{ji} \in [\hat{l}_{\alpha,ji}, \hat{u}_{\alpha,ji}] \Bigg\},$$

which corresponds to the biased but computable confidence region

$$\prod_{j\neq i} [\hat{l}_{\alpha,ji}, \hat{u}_{\alpha,ji}] = \prod_{j\neq i} \left[\hat{w}_{ji} \pm \hat{\sigma}_{ji} \frac{q\left(1 - \frac{\alpha}{2p(p-1)}\right)}{2\sqrt{n}}\right].$$

for the Gaussian edge weights, where the product is over all combinations of possible edges $1 \leq j \neq i \leq p$. The biased confidence region $\prod_{j\neq i} [\hat{l}_{\alpha,ji}, \hat{u}_{\alpha,ji}]$ does not necessarily contain the population Gaussian edge weights with a probability of at least $1-\alpha$ in the large sample limit. However, it can be used to construct a conservative confidence region for the causal

graph. To see this, note that by further penalizing the wrong (non-causal) edge weights, the causal graph still yields the minimum edge weight directed spanning tree. Hence,

$$\begin{aligned}
&\liminf_{n\to\infty} P(\mathcal{G} \in C(\hat{l}_\alpha, \hat{u}_\alpha)) \\
\geq &\liminf_{n\to\infty} P\left(\bigcap_{(j\to i)\in\mathcal{E}} (w_{ji} \in [\hat{l}_{\alpha,ji}, \hat{u}_{\alpha,ji}]) \bigcap_{(j\to i)\notin\mathcal{E}} (w_{ji} \in [\tilde{l}_{\alpha,ji}, \tilde{u}_{\alpha,ji}]) \bigcap_{(j\to i)\notin\mathcal{E}} (\tilde{u}_{\alpha,ji} \leq \hat{u}_{\alpha,ji}) \right) \\
\geq &1-\alpha,
\end{aligned}$$

as $P\left(\tilde{u}_{\alpha,ji} \leq \hat{u}_{\alpha,ji}\right) \to_n 1$ for all $(j \to i) \notin \mathcal{E}$ by Lemma D.5 below. ■

Lemma D.5 *Suppose that the assumptions of Lemma D.4 hold. It holds that*

$$\forall (j \to i) \notin \mathcal{E}, \forall \alpha \in (0,1) : P\left(\tilde{u}_{\alpha,ji} \leq \hat{u}_{\alpha,ji}\right) \to_n 1.$$

Proof of Lemma D.5. Fix any $(j \to i) \notin \mathcal{E}$ and $\alpha \in (0,1)$ and note that we want to show that

$$\begin{aligned}
&\tilde{u}_{\alpha,ji} \leq \hat{u}_{\alpha,ji} \\
\iff &\tilde{w}_{ji} + c\frac{\tilde{\sigma}_{ji}}{\sqrt{n}} \leq \hat{w}_{ji} + c\frac{\hat{\sigma}_{ji}}{\sqrt{n}} \\
\iff &0 \leq \log\left(\hat{\mu}_{ji}\right) + c\frac{\hat{\sigma}_{ji}}{\sqrt{n}} - \log\left(\hat{\mu}_{ji} - \delta^2_{n,ji}\right) - c\frac{\tilde{\sigma}_{ji}}{\sqrt{n}}
\end{aligned}$$

holds with probability converging to one, where c is a strictly positive constant. It suffices to show that an even smaller quantity is non-negative with probability converging to one. That is, it suffices to show that

$$0 \leq \log\left(\hat{\mu}_{ji}\right) + c\frac{\hat{\sigma}_{ji}}{\sqrt{n}} - \log\left(\hat{\mu}_{ji} - \delta^2_{n,ji}\right) - c\frac{\tilde{\sigma}^*_{ji}}{\sqrt{n}},$$

with increasing probability, where

$$\tilde{\sigma}^*_{ji} := \sqrt{\frac{\widehat{\Sigma}_{M,ji}}{(\hat{\mu}_{ji} - \delta^2_{n,ji})^2} + \frac{\widehat{\Sigma}_{V,i}}{\hat{\nu}_i^2} + 2\frac{|\widehat{\Sigma}_{MV,ji,i}|}{(\hat{\mu}_{ji} - \delta^2_{n,ji})\hat{\nu}_i}} \geq \tilde{\sigma}_{ji},$$

with $P(\tilde{\sigma}^*_{ji} > 0) \to_n 1$. Let $d_n(t) : [0,\infty) \to \mathbb{R}$ denote the random function given by

$$\begin{aligned}
d_n(t) := &\log\left(\hat{\mu}_{ji}\right) + c\frac{\hat{\sigma}_{ji}}{\sqrt{n}} - \log\left(\hat{\mu}_{ji} - t\right) \\
&- \frac{c}{\sqrt{n}}\sqrt{\frac{\widehat{\Sigma}_{M,ji}}{(\hat{\mu}_{ji} - t)^2} + \frac{\widehat{\Sigma}_{V,i}}{\hat{\nu}_i^2} + 2\frac{|\widehat{\Sigma}_{MV,ji,i}|}{(\hat{\mu}_{ji} - t)\hat{\nu}_i}}.
\end{aligned}$$

It holds that $d_n(0) = 0$ surely, so by the mean value theorem, the desired conclusion holds if it with probability one (as n tends to infinity) holds, for all $t \in [0, \delta^2_{n,ji}]$, that $d'_n(t) \geq 0$.

Now fix $\eta > 0$ and choose $M_\eta, \varepsilon_1, \ldots, \varepsilon_5 > 0$ such that the constant lower bounds in the following inequalities are strictly positive

$$\begin{aligned}
\Omega_n(1) &:= (\hat{\mu}_{ji} \leq M_\eta),\\
\Omega_n(2) &:= (\Sigma_{M,ji} - \varepsilon_1 \leq \widehat{\Sigma}_{M,ji} \leq \Sigma_{M,ji} + \varepsilon_1),\\
\Omega_n(3) &:= (\Sigma_{V,i} - \varepsilon_2 \leq \widehat{\Sigma}_{V,i} \leq \Sigma_{V,i} + \varepsilon_2),\\
\Omega_n(4) &:= (0 \leq |\widehat{\Sigma}_{MV,ji,i}| \leq |\Sigma_{MV,ji,i}| + \varepsilon_3),\\
\Omega_n(5) &:= (\mu_{ji} - \varepsilon_4 \leq \hat{\mu}_{ji} - \delta^2_{n,ji} \leq \mu_{ji} + \varepsilon_4),\\
\Omega_n(6) &:= (\nu_i - \varepsilon_5 \leq \hat{\nu}_i \leq \nu_i + \varepsilon_5),
\end{aligned}$$

and $\liminf_{n\to\infty} P(\Omega_n(1)) > 1 - \eta$. This is possible as $\hat{\mu}_{ji} - \delta^2_{n,ji} \xrightarrow{P}_n \mu_{ji} > 0$ and

$$\begin{aligned}
\delta^2_{n,ji} = E[|\hat{\delta}_{nm,ji}|^2|\tilde{\mathbf{X}}] &= E[|\hat{\delta}_{nm,ji}|^{\frac{4+\xi}{2+\xi/2}}|\tilde{\mathbf{X}}]\\
&\leq E[|\hat{\delta}_{nm,ji}|^{4+\xi}|\tilde{\mathbf{X}}]^{\frac{1}{2+\xi/2}} = O_p(1),
\end{aligned}$$

by the conditional Jensen's inequality and concavity of $[0,\infty) \ni x \mapsto x^{\frac{1}{2+\xi/2}}$, which implies that $\hat{\mu}_{ji} = (\hat{\mu}_{ji} - \hat{\delta}^2_{n,ji} - \mu_{ji}) + (\hat{\delta}^2_{n,ji} + \mu_{ji}) = o_p(1) + O_p(1) = O_p(1)$. Furthermore, as

$$\begin{aligned}
&\widehat{\Sigma}_{M,ji} \xrightarrow{P}_n \Sigma_{M,ji} > 0, \quad \widehat{\Sigma}_{V,i} \xrightarrow{P}_n \Sigma_{V,i} > 0,\\
&|\widehat{\Sigma}_{MV,ji,i}| \xrightarrow{P}_n |\Sigma_{MV,ji,i}| \geq 0, \quad \hat{\nu}_i \xrightarrow{P}_n \nu_i > 0,
\end{aligned}$$

it holds that

$$\begin{aligned}
\limsup_{n\to\infty} P\left(\bigcup_{1\leq k\leq 6} \Omega_n(k)^c\right) &\leq \sum_{1\leq k\leq 6} \limsup_{n\to\infty} P(\Omega_n(k)^c)\\
&= \limsup_{n\to\infty} P(\Omega_n(1)^c) \leq \eta.
\end{aligned}$$

Here we used that the diagonal elements of the limit covariance matrix are assumed strictly positive. That $\mu_{ji}, \nu_i > 0$ follows from the fact that $X_i - \mathbb{E}[X_i|X_j]$ is assumed to have a density (w.r.t. Lebesgue measure) and that the variables are non-degenerate $\nu_i = \mathrm{Var}(X_i) > 0$. Thus, we have that

$$\liminf_{n\to\infty} P\left(\bigcap_{1\leq k\leq 6} \Omega_n(k)\right) > 1 - \eta.$$

Now consider a fixed $\omega \in \bigcap_{1\leq k\leq 6} \Omega_n(k)$ and note that with $g_n : [0, \delta^2_{n,ji}] \to \mathbb{R}$ given by $g_n(t) = \hat{\mu}_{ji} - t$ we have that g_n is decreasing and that

$$g_n([0, \delta^2_{n,ji}]) \subset [\mu_{ji} - \varepsilon_4, \hat{\mu}_{ji}] \subset (0, M_\eta]$$

We have for any $t \in [0, \delta_{n,ji}^2]$ that

$$d_n'(t) = \frac{1}{\hat{\mu}_{ji} - t} - \frac{c}{\sqrt{n}} \left(\frac{\widehat{\Sigma}_{M,ji}}{(\hat{\mu}_{ji} - t)^2} + \frac{\widehat{\Sigma}_{V,i}}{\hat{\nu}_i^2} + \frac{2|\widehat{\Sigma}_{MV,ji,i}|}{(\hat{\mu}_{ji} - t)\hat{\nu}_i} \right)^{-1/2} \times \left(\frac{\widehat{\Sigma}_{M,ji}}{(\hat{\mu}_{ji} - t)^3} + \frac{|\widehat{\Sigma}_{MV,ji,i}|}{(\hat{\mu}_{ji} - t)^2 \hat{\nu}_i} \right),$$

hence,

$$\begin{aligned}
d_n'(t) &= \frac{1}{\hat{\mu}_{ji} - t} - \frac{c}{\sqrt{n}} \left(\frac{\widehat{\Sigma}_{M,ji}}{g_n(t)^2} + \frac{\widehat{\Sigma}_{V,i}}{\hat{\nu}_i^2} + \frac{2|\widehat{\Sigma}_{MV,ji,i}|}{g_n(t)\hat{\nu}_i} \right)^{-1/2} \\
&\quad \times \left(\frac{\widehat{\Sigma}_{M,ji}}{g_n(t)^3} + \frac{|\widehat{\Sigma}_{MV,ji,i}|}{g_n(t)^2 \hat{\nu}_i} \right) \\
&\geq \frac{1}{\hat{\mu}_{ji}} - \frac{c}{\sqrt{n}} \left(\frac{\widehat{\Sigma}_{M,ji}}{\hat{\mu}_{ji}^2} + \frac{\widehat{\Sigma}_{V,i}}{\hat{\nu}_i^2} \right)^{-1/2} \\
&\quad \times \left(\frac{\widehat{\Sigma}_{M,ji}}{(\hat{\mu}_{ji} - \delta_{n,ji}^2)^3} + \frac{|\widehat{\Sigma}_{MV,ji,i}|}{(\hat{\mu}_{ji} - \delta_{n,ji}^2)^2 \hat{\nu}_i} \right) \\
&\geq \frac{1}{M_\eta} - \frac{c}{\sqrt{n}} \left(\frac{\Sigma_{M,ji} - \varepsilon_1}{M_\eta^2} + \frac{\Sigma_{V,i} - \varepsilon_2}{(\nu_i + \varepsilon_5)^2} \right)^{-1/2} \\
&\quad \times \left(\frac{\Sigma_{M,ji} + \varepsilon_1}{(\mu_{ji} - \varepsilon_4)^3} + \frac{|\Sigma_{MV,ji,i}| + \varepsilon_3}{(\mu_{ji} - \varepsilon_4)^2 (\nu_i - \varepsilon_5)} \right) \\
&=: \frac{1}{M_\eta} - \frac{C_{M_\eta, \varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5}}{\sqrt{n}} \\
&\geq 0,
\end{aligned}$$

for $n \geq (C_{M_\eta, \varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5} M_\eta)^2$. We conclude that for $n \geq (C_{M_\eta, \varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5} M_\eta)^2$

$$\begin{aligned}
P\left(\tilde{u}_{\alpha,ji} \leq \hat{u}_{\alpha,ji}\right) &= P\left(0 \leq \log\left(\hat{\mu}_{ji}\right) + c\frac{\hat{\sigma}_{ji}}{\sqrt{n}} - \log\left(\hat{\mu}_{ji} - \delta_{n,ji}^2\right) - c\frac{\tilde{\sigma}_{ji}}{\sqrt{n}}\right) \\
&\geq P\left(\forall t \in [0, \delta_{n,ji}^2] : d_n'(t) \geq 0\right) \\
&\geq P\left(\bigcap_{1 \leq k \leq 6} \Omega_n(k)\right).
\end{aligned}$$

Hence,

$$\liminf_{n \to \infty} P\left(\tilde{u}_{\alpha,ji} \leq \hat{u}_{\alpha,ji}\right) \geq \liminf_{n \to \infty} P\left(\bigcap_{1 \leq k \leq 6} \Omega_n(k)\right) \geq 1 - \eta,$$

and as $\eta > 0$ was chosen arbitrarily, we have the desired conclusion

$$P\left(\tilde{u}_{\alpha,ji} \leq \hat{u}_{\alpha,ji}\right) \to_n 1.$$

■

Proof of Theorem 11. Consider a collection of arbitrary and possibly data-dependent substructures $\mathcal{R}_1, \mathcal{R}_2, ...$ and level $\alpha \in (0,1)$. First, we note that the score associated with two sets of edge weights w_1 and w_2 is weakly monotone, that is, $S_{\mathcal{T}_p}(w_1) \leq S_{\mathcal{T}_p}(w_2)$ if w_1 and w_2 satisfy the component-wise partial ordering $w_1 \leq w_2$. Furthermore, the restricted score function $w \mapsto S_{\mathcal{T}_p(\mathcal{R})}(w)$ is also weakly monotone for any set of restrictions $\mathcal{R}$.

Let $k \in \mathbb{N}$ and suppose that the null hypothesis

$$\mathcal{H}_0(\mathcal{R}_k) : \mathcal{E}_{\mathcal{R}_k} \setminus \mathcal{E} = \emptyset,\ \mathcal{E} \setminus \mathcal{E}^{\text{miss}}_{\mathcal{R}_k} = \emptyset,\ r_k = \text{rt}(\mathcal{G}),$$

corresponding to the restriction $\mathcal{R}_k = (\mathcal{E}_{\mathcal{R}_k}, \mathcal{E}^{\text{miss}}_{\mathcal{R}_k}, r_k)$ is true.

If there is a graph in $\hat{C}_{\text{Bon}} := \hat{C}(\hat{l}_\alpha, \hat{u}_\alpha)$ satisfying the restrictions imposed by the substructure $\mathcal{R}_k$, then there exist $\hat{l}_\alpha \leq w' \leq \hat{u}_\alpha$ such that $S_{\mathcal{T}_p}(w')$ attains its minimum value in a graph satisfying $\mathcal{R}_k$. Penalizing (or removing) edges that are not present in the minimum edge weight directed tree does not affect the score of the minimum edge weigh directed tree. Hence, it holds that

$$S_{\mathcal{T}_p(\mathcal{R}_k)}(w') = S_{\mathcal{T}_p}(w').$$

Monotonicity of $S_{\mathcal{T}_p(\mathcal{R}_k)}$ and $S_{\mathcal{T}_p}$ in the edge weights imply that

$$S_{\mathcal{T}_p(\mathcal{R}_k)}(\hat{l}_\alpha) \leq S_{\mathcal{T}_p(\mathcal{R}_k)}(w') = S_{\mathcal{T}_p}(w') \leq S_{\mathcal{T}_p}(\hat{u}_\alpha).$$

Hence, $S_{\mathcal{T}_p(\mathcal{R}_k)}(\hat{l}_\alpha) > S_{\mathcal{T}_p}(\hat{u}_\alpha)$ entails that no graph in $\hat{C}$ satisfies the restrictions of $\mathcal{R}_k$. (This is a slightly conservative criterion as $S_{\mathcal{T}_p(\mathcal{R}_k)}(\hat{l}_\alpha) \leq S_{\mathcal{T}_p}(\hat{u}_\alpha)$ does not necessarily guarantee that a graph in $\hat{C}_{\text{Bon}}$ satisfies the restrictions of $\mathcal{R}_k$.)

Therefore, if $\psi^{\text{CheckC}}_{\mathcal{R}_k} = 1$, then we know that there is no graph in $\hat{C}_{\text{Bon}}$ satisfying the restrictions of $\mathcal{R}_k$. As the causal graph $\mathcal{G}$ satisfies the restriction $\mathcal{R}_k$ we conclude that $\mathcal{G}$ is not contained in $\hat{C}_{\text{Bon}}$. Thus for any true $\mathcal{R}_k$ we have that

$$(\psi^{\text{CheckC}}_{\mathcal{R}_k} = 1) \subseteq (\mathcal{G} \notin \hat{C}_{\text{Bon}}).$$

Since this holds for any true $\mathcal{R}_k$, the conclusion follows by noting that

$$\limsup_{n\to\infty} P\left(\bigcup_{k:\mathcal{H}_0(\mathcal{R}_k) \text{ is true}} (\psi^{\text{CheckC}}_{\mathcal{R}_k} = 1) \right) \leq \limsup_{n\to\infty} P(\mathcal{G} \notin \hat{C}_{\text{Bon}}) \leq \alpha,$$

where we used Theorem 10.

For the claim about the level guarantee of the ConvB test, let $k \in \mathbb{N}$ and consider a true substructure restriction $\mathcal{R}_k = (\mathcal{E}_{\mathcal{R}_k}, \mathcal{E}^{\text{miss}}_{\mathcal{R}_k}, r_k)$. Suppose that $\mathcal{G} \in \hat{C}_{\text{Bon}}$. This implies that there exist $\hat{l}_\alpha \leq w' \leq \hat{u}_\alpha$ such that $S_{\mathcal{T}_p}(w')$ attains its minimum value in a graph satisfying $\mathcal{R}_k$. Now let $w'' = (w''_{ji})_{j \neq i}$ be given by

$$w''_{ji} = \begin{cases} \hat{u}_{ji} & \text{if } [\exists l \neq j : (l \to i) \in \mathcal{E}_{\mathcal{R}_k}] \vee [(i \to j) \in \mathcal{E}_{\mathcal{R}_k}] \vee [(j \to i) \in \mathcal{E}^{\text{miss}}_{\mathcal{R}_k}] \vee [i = r], \\ w'_{ji} & \text{otherwise,} \end{cases}$$

where we penalize edges that are in disagreement with the substructure restriction $\mathcal{R}_k$. It is clear that the MWDST using the edge weights w' and w'', i.e., $\mathcal{G}^*_{\mathcal{T}_p}(w')$ and $\mathcal{G}^*_{\mathcal{T}_p}(w'')$, both satisfy the substructure restriction $\mathcal{R}_k$. However, as w' is unknown, so is w''. We lower bound the unknown w' by $\hat{l}$ and define $\check{w} = (\check{w}_{ji})_{j \neq i}$ as

$$\check{w}_{ji} = \begin{cases} \hat{u}_{ji} & \text{if } [\exists l \neq j : (l \to i) \in \mathcal{E}_{\mathcal{R}_k}] \vee [(i \to j) \in \mathcal{E}_{\mathcal{R}_k}] \vee [(j \to i) \in \mathcal{E}^{\text{miss}}_{\mathcal{R}_k}] \vee [i = r], \\ \hat{l}_{ji} & \text{otherwise,} \end{cases}$$

Now, the MWDST $\mathcal{G}^*_{\mathcal{T}_p}(\tilde{w})$ may use edges that are in disagreement with $\mathcal{R}_k$ or not satisfy $\mathcal{R}_k$. (For example, consider a three node causal graph $V = \{1, 2, 3\}$ with edges $1 \to 2 \to 3$ and consider the substructure restriction $\mathcal{E}_{\mathcal{R}_k} = \{(1 \to 2)\}$. Now it may happen that $\hat{l}_{12} + \hat{l}_{23} > \hat{l}_{13} + \hat{u}_{32}$ or $\hat{l}_{12} + \hat{l}_{23} > \hat{l}_{23} + \hat{l}_{31}$, that is, the MWDST $\mathcal{G}^*_{\mathcal{T}_p}(\tilde{w})$ does not satisfy the substructure restriction.) We now argue that this happens with probability tending to zero.

By the assumed identifiability, i.e., that Assumption 1 holds, we have that

$$\Delta := \min_{\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \mathcal{G}} \ell_{\text{G}}(\tilde{\mathcal{G}}) - \ell_{\text{G}}(\mathcal{G}) > 0. \tag{41}$$

Now consider the events $(A_n)_{n \in \mathbb{N}}$ (that are independent of k) given by

$$A_n := \bigcap_{j \neq i} \left(|\hat{l}_{ji} - w^{\text{G}}_{ji}| < \frac{\Delta}{p-1} \right), \quad n \in \mathbb{N}.$$

We realize that on A_n it must hold that the MWDST $\mathcal{G}^*_{\mathcal{T}_p}(\tilde{w})$ satisfies $\mathcal{R}_k$. Thus, for a true substructure restriction $\mathcal{R}_k$ we have that

$$A_n \cap (\mathcal{G} \in \hat{C}_{\text{Bon}}) \subseteq (\psi^{\text{ConvB}}_{\mathcal{R}_k} = 0) \iff A_n^c \cup (\mathcal{G} \notin \hat{C}_{\text{Bon}}) \supseteq (\psi^{\text{ConvB}}_{\mathcal{R}_k} = 1),$$

for all $n \in \mathbb{N}$. Hence, we have that

$$\begin{aligned} \limsup_{n \to \infty} P\left(\bigcup_{k : \mathcal{H}_0(\mathcal{R}_k) \text{ is true}} (\psi^{\text{ConvB}}_{\mathcal{R}_k} = 1) \right) &= \limsup_{n \to \infty} P\left((\mathcal{G} \notin \hat{C}_{\text{Bon}}) \cup A_n^c \right) \\ &\leq \limsup_{n \to \infty} P(\mathcal{G} \notin \hat{C}_{\text{Bon}}) + \limsup_{n \to \infty} P(A_n^c) \\ &\leq \alpha, \end{aligned}$$

by Theorem 10, proving the claim. It only remains to argue that $\limsup_{n \to \infty} P(A_n^c) = 0$. To that end, note that by Theorem 10 it holds that $\hat{\Sigma} \xrightarrow{P}_n \Sigma$ and that

$$\sqrt{n} \left(\begin{pmatrix} \hat{\mu} - \delta^2_n \\ \hat{\nu} \end{pmatrix} - \begin{pmatrix} \mu \\ \nu \end{pmatrix} \right) \xrightarrow{\mathcal{D}}_n \mathcal{N}(0, \Sigma).$$

By the strengthened assumptions, i.e., the $\sqrt{n}$-convergence for the non-causal edges, we have that (see the arguments for the causal edges from Theorem 10)

$$\sqrt{n} \left(\begin{pmatrix} \hat{\mu} \\ \hat{\nu} \end{pmatrix} - \begin{pmatrix} \mu \\ \nu \end{pmatrix} \right) \xrightarrow{\mathcal{D}}_n \mathcal{N}(0, \Sigma).$$

Thus, for any $j \neq i$, that

$$\hat{u}_{ji}, \hat{l}_{ji} = \frac{1}{2}\log\left(\frac{\hat{\mu}_{ji}}{\hat{\nu}_i}\right) \pm z_\alpha \frac{\hat{\sigma}_{ji}}{2\sqrt{n}} \xrightarrow{P}_n \frac{1}{2}\log\left(\frac{\mu_{ji}}{\nu_{ji}}\right) = w^{\mathrm{G}}_{ji}, \tag{42}$$

since $\hat{\mu} \xrightarrow{P}_n \mu$, $\hat{\nu} \xrightarrow{P}_n \nu$, and $\hat{\sigma}_{ji} \xrightarrow{P}_n \sigma_{ji}$. The convergence statements in Equation (42) obviously implies that $P(A_n) \to_n 1$, since Δ is strictly positive, see Equation (41)). This concludes the proof.

■

D.4 Proofs of Section 5

D.4.1 Proofs of first results in Section 5

Proof of Lemma 12. As conditioning reduces entropy we always have that

$$\begin{aligned}
\ell_{\mathrm{CE}}(\tilde{\mathcal{G}}, i) = h(X_i | X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}) &= h(X_i - \mathbb{E}[X_i | X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}] | X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}) \\
&\leq h(X_i - \mathbb{E}[X_i | X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}]) \\
&= \ell_{\mathrm{E}}(\tilde{\mathcal{G}}, i).
\end{aligned}$$

Furthermore, note that when conditioning we 'throw out' dependence information captured by the mutual information $I(X_i - \mathbb{E}[X_i | X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}]; X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)})$, which is zero if and only if $X_i - \mathbb{E}[X_i | X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}] \perp\!\!\!\perp X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}$. This is especially the case for the true graph, i.e., $X_i - \mathbb{E}[X_i | X_{\mathrm{pa}^{\mathcal{G}}(i)}] \perp\!\!\!\perp X_{\mathrm{pa}^{\mathcal{G}}(i)}$, implying that $\ell_{\mathrm{CE}}(\mathcal{G}, i) = \ell_{\mathrm{E}}(\mathcal{G}, i)$. Consequently, we have that the local conditional entropy score gap lower bounds the local entropy score gap,

$$\ell_{\mathrm{CE}}(\tilde{\mathcal{G}}, i) - \ell_{\mathrm{CE}}(\mathcal{G}, i) \leq \ell_{\mathrm{E}}(\tilde{\mathcal{G}}, i) - \ell_{\mathrm{E}}(\mathcal{G}, i).$$

Furthermore, from the arguments in the proof of Lemma 6 we have that

$$\begin{aligned}
\ell_{\mathrm{E}}(\tilde{\mathcal{G}}, i) &= \inf_{\tilde{N}_i \sim P_{\tilde{N}_i} \in \mathcal{P}} h\left(X_i - \mathbb{E}\left[X_i | X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}\right], \tilde{N}_i\right) \\
&\leq \inf_{\tilde{N}_i \sim P_{\tilde{N}_i} \in \mathcal{P}_G} h\left(X_i - \mathbb{E}\left[X_i | X_{\mathrm{pa}^{\tilde{\mathcal{G}}}(i)}\right], \tilde{N}_i\right) \\
&= \ell_{\mathrm{G}}(\tilde{\mathcal{G}}, i) + \log(\sqrt{2\pi e}).
\end{aligned}$$

If X is generated by a causal additive tree model with Gaussian noise, i.e., with generating SCM $\theta = (\mathcal{G}, (f_i), P_N)$ with $P_N \in \mathcal{P}^p_{\mathrm{G}}$, then $\ell_{\mathrm{E}}(\mathcal{G}, i) = h(X_i - \mathbb{E}[X_i | X_{\mathrm{pa}^{\mathcal{G}}(i)}]) = h(N_i) = \log(\sqrt{2\pi e}\sigma_i) = \log(\sqrt{2\pi e}) + \frac{1}{2}\log(\mathbb{E}[N_i^2]) = \log(\sqrt{2\pi e}) + l_{\mathrm{G}}(\mathcal{G}, i)$, in which case the local entropy score gap lower bounds the local Gaussian score gap

$$\ell_{\mathrm{E}}(\tilde{\mathcal{G}}, i) - \ell_{\mathrm{E}}(\mathcal{G}, i) \leq \ell_{\mathrm{G}}(\tilde{\mathcal{G}}, i) - \ell_{\mathrm{G}}(\mathcal{G}, i).$$

■

Proof of Lemma 13. Note that $\mathbb{E}[Y|X] = \mathbb{E}[f(X) + N_Y|X] = f(X) + \mathbb{E}[N_Y|X] = f(X) + \mathbb{E}[N_Y]$, since $N_Y \perp\!\!\!\perp N_X = X$. Hence, the score difference can be written as

$$\begin{aligned}\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) &= \ell_{\mathrm{E}}(\tilde{\mathcal{G}}, X) - \ell_{\mathrm{E}}(\mathcal{G}, X) + \ell_{\mathrm{E}}(\tilde{\mathcal{G}}, Y) - \ell_{\mathrm{E}}(\mathcal{G}, Y) \\ &= h(X - \mathbb{E}(X|Y)) - h(X) + h(Y) - h(Y - E(Y|X)) \\ &= h(X - \mathbb{E}(X|Y)) - h(X) + h(Y) - h(N_Y + \mathbb{E}[N_Y]) \\ &= h(X - \mathbb{E}(X|Y)) - h(X) + h(Y) - h(N_Y),\end{aligned}$$

as the differential entropy is translation invariant. Now note that as $N_Y \perp\!\!\!\perp N_X$ it holds that $N_Y \perp\!\!\!\perp f(X)$, so conditioning on $f(X)$ yields that

$$\begin{aligned}h(Y) &= h(Y|f(X)) + I(Y; f(X)) \\ &= h(f(X) + N_Y|f(X)) + I(Y; f(X)) \\ &= h(N_Y) + I(Y; f(X)).\end{aligned}$$

Similarly, conditioning on X yields that

$$\begin{aligned}h(Y) &= h(Y|X) + I(Y; X) \\ &= h(N_Y) + I(Y; X),\end{aligned}$$

which proves that

$$I(Y; f(X)) = I(Y; X).$$

This equality is normally derived by restricting f to be bijective, but here it holds regardless by the structural assignment form, as Y only depends on X through $f(X)$. Furthermore, we have that

$$\begin{aligned}h(X - \mathbb{E}[X|Y]) &= I(X - \mathbb{E}[X|Y]; Y) + h(X - \mathbb{E}[X|Y]|Y) \\ &= I(X - \mathbb{E}[X|Y]; Y) + h(X|Y).\end{aligned}$$

Hence,

$$\begin{aligned}h(X - \mathbb{E}[X|Y]) - h(X) &= I(X - \mathbb{E}[X|Y]; Y) + h(X|Y) - h(X) \\ &= I(X - \mathbb{E}[X|Y]; Y) - I(Y; X).\end{aligned}$$

Thus

$$\begin{aligned}\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) &= h(X - \mathbb{E}[X|Y]) - h(X) + h(Y) - h(N_Y) \\ &= I(X - \mathbb{E}[X|Y]; Y) - I(Y; X) + h(N_Y) + I(Y; f(X)) - h(N_Y) \\ &= I(X - \mathbb{E}[X|Y]; Y) - I(Y; X) + I(Y; f(X)) \\ &= I(X - \mathbb{E}[X|Y]; Y),\end{aligned}$$

proving the claim. ■

Proof of Proposition 14. As the conditional mean $\mathbb{E}[X|Y]$ vanishes, we have that

$$\begin{aligned}\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) &= I(X - \mathbb{E}(X|Y); Y) \\ &= I(X;Y) \\ &= I(Y;X) \\ &= I(Y; f(X)),\end{aligned}$$

where the last equality was derived in the proof of Lemma 13. Now let $f(X)^{\mathrm{G}}$ and N_Y^{G} be independent normal distributed random variables with the same mean and variance as $f(X)$ and N_Y. That is, $f(X)^{\mathrm{G}} \sim \mathcal{N}(\mathbb{E}[f(X)], \mathrm{Var}(f(X)))$, $N_Y^{\mathrm{G}} \sim \mathcal{N}(\mathbb{E}[N_Y], \mathrm{Var}(N_Y))$ with $N_Y^{\mathrm{G}} \perp\!\!\!\perp f(X)^{\mathrm{G}}$ such that $f(X)^{\mathrm{G}} + N_Y^{\mathrm{G}} \sim \mathcal{N}(\mathbb{E}[f(X)] + \mathbb{E}[N_Y], \mathrm{Var}(f(X)) + \mathrm{Var}(N_Y))$.

(a) If $D_{\mathrm{KL}}(f(X) \| f(X)^{\mathrm{G}}) \leq D_{\mathrm{KL}}(N_Y \| N_Y^{\mathrm{G}})$ then by Lemma C.1 of Silva (2009) we have, since $X \perp\!\!\!\perp N_Y$, that

$$I(Y; f(X)) = I(f(X) + N_Y; f(X)) \geq I(f(X)^{\mathrm{G}} + N_Y^{\mathrm{G}}; f(X)^{\mathrm{G}}),$$

Note, we have equality if and only if $f(X)$ and N_Y are jointly Gaussian. Furthermore,

$$\begin{aligned}I(f(X)^{\mathrm{G}} + N_Y^{\mathrm{G}}; f(X)^{\mathrm{G}}) &= h(f(X)^{\mathrm{G}} + N_Y^{\mathrm{G}}) - h(f(X)^{\mathrm{G}} + N_Y^{\mathrm{G}} | f(X)^{\mathrm{G}}) \\ &= h(f(X)^{\mathrm{G}} + N_Y^{\mathrm{G}}) - h(N_Y^{\mathrm{G}}) \\ &= \log(\sqrt{2\pi(\mathrm{Var}(f(X)) + \mathrm{Var}(N_Y))}) - \log(\sqrt{2\pi \mathrm{Var}(N_Y)}) \\ &= \frac{1}{2} \log\left(\frac{\mathrm{Var}(f(X)) + \mathrm{Var}(N_Y)}{\mathrm{Var}(N_Y)}\right) \\ &= \frac{1}{2} \log\left(1 + \frac{\mathrm{Var}(f(X))}{\mathrm{Var}(N_Y)}\right).\end{aligned}$$

(b) If $f(X)+N_Y$ is log-concave distributed, then by Theorem 3 of Marsiglietti and Kostina (2018) we have that

$$h(f(X) + N_Y) \geq \frac{1}{2} \log\left(4\mathrm{Var}(f(X) + N_Y)\right) = \frac{1}{2} \log\left(4(\mathrm{Var}(f(X)) + \mathrm{Var}(N_Y)\right).$$

Furthermore, it is well known that for fixed variance, the normal distribution maximizes entropy, hence

$$h(N_Y) \leq h(N_Y^{\mathrm{G}}) = \frac{1}{2} \log\left(2\pi \mathrm{Var}(N_Y)\right).$$

Therefore, we get that

$$\begin{aligned}I(Y; f(X)) &= I(f(X) + N_Y; f(X)) \\ &= h(f(X) + N_Y) - h(f(X) + N_Y | f(X)) \\ &= h(f(X) + N_Y) - h(N_Y) \\ &\geq \frac{1}{2} \log\left(4(\mathrm{Var}(f(X)) + \mathrm{Var}(N_Y)\right) - \frac{1}{2} \log\left(2\pi e \mathrm{Var}(N_Y)\right) \\ &= \frac{1}{2} \log\left(\frac{2}{\pi e} + \frac{2}{\pi e} \frac{\mathrm{Var}(f(X))}{\mathrm{Var}(N_Y)}\right),\end{aligned}$$

which yields a strictly positive lower bound if and only if

$$\frac{2}{\pi e} + \frac{2}{\pi e}\frac{\mathrm{Var}(f(X))}{\mathrm{Var}(N_Y)} > 1 \iff \frac{\mathrm{Var}(f(X))}{\mathrm{Var}(N_Y)} > \frac{\pi e}{2} - 1 \approx 3.27.$$

■

Lemma D.6 *Two different but Markov equivalent trees $\tilde{\mathcal{G}}$ and $\hat{\mathcal{G}}$ share the exact same edges except for a single reversed directed path between the two root nodes of the graphs,*

$$\begin{array}{lccccccccc} \hat{\mathcal{G}}: & c_1 & \to & c_2 & \to & \cdots & \to & c_{r-1} & \to & c_r, \\ \tilde{\mathcal{G}}: & c_r & \to & c_{r-1} & \to & \cdots & \to & c_2 & \to & c_1, \end{array}$$

with $c_1 = \mathrm{rt}(\hat{\mathcal{G}})$ and $c_r = \mathrm{rt}(\tilde{\mathcal{G}})$.

Proof of Lemma D.6. First, note that there always exists a unique directed path in $\hat{\mathcal{G}}$ from $\mathrm{rt}(\hat{\mathcal{G}})$ to $\mathrm{rt}(\tilde{\mathcal{G}})$

$$\hat{\mathcal{G}} : \mathrm{rt}(\hat{\mathcal{G}}) = c_1 \to \cdots \to c_{r-1} \to c_r = \mathrm{rt}(\tilde{\mathcal{G}}).$$

Since $\tilde{\mathcal{G}}$ and $\hat{\mathcal{G}}$ are Markov equivalent, they share the same skeleton, so in $\tilde{\mathcal{G}}$ the above path must be reversed. That is, there exists a unique directed path in $\tilde{\mathcal{G}}$ from $\mathrm{rt}(\tilde{\mathcal{G}})$ to $\mathrm{rt}(\hat{\mathcal{G}})$ given by

$$\tilde{\mathcal{G}} : \mathrm{rt}(\tilde{\mathcal{G}}) = c_r \to c_{r-1} \to \cdots \to c_1 = \mathrm{rt}(\hat{\mathcal{G}}),$$

If $r = p$ we are done, so assume $r < p$. As $\hat{\mathcal{G}}$ is a directed tree there must exists a node z_2 which is not a part of the above path but is a child of a node in the path. That is, there exists a node $z_1 \in \{c_1, \ldots, c_r\}$ such that $\hat{\mathcal{G}}$ contains the edge

$$\hat{\mathcal{G}} : z_1 \to z_2.$$

Furthermore, by equality of skeleton, this edge must also be present in $\tilde{\mathcal{G}}$,

$$\tilde{\mathcal{G}} : z_1 - z_2.$$

Assume for contradiction that $z_2 \to z_1$ in $\tilde{\mathcal{G}}$. As such, it must hold that $z_1 = c_r = \mathrm{rt}(\tilde{\mathcal{G}})$ for otherwise if $z_1 \in \{c_1, \ldots, c_{r-1}\}$ then z_1 would have two parents in $\tilde{\mathcal{G}}$, a contradiction since $\tilde{\mathcal{G}}$ is a directed tree. However, if $z_1 = c_r = \mathrm{rt}(\tilde{\mathcal{G}})$ then there is an incoming edge into the root node, a contradiction. We conclude that the directed edge $z_1 \to z_2$ also is present in $\tilde{\mathcal{G}}$.

Any paths further out on this branch will coincide in both graphs for otherwise there exists nodes with two parents. These arguments show that any paths branching out from the main reversed path will coincide in both $\hat{\mathcal{G}}$ and $\tilde{\mathcal{G}}$. Thus, the two graphs coincide up to a directed path between root nodes that is reversed.

■

Proof of Proposition 15. By Lemma D.6 there exists a path reversal

$$\begin{array}{lccccccccc} \mathcal{G}: & \mathrm{rt}(\mathcal{G}) = c_1 & \to & c_2 & \to & \cdots & \to & c_{r-1} & \to & c_r = \mathrm{rt}(\tilde{\mathcal{G}}), \\ \tilde{\mathcal{G}}: & \mathrm{rt}(\tilde{\mathcal{G}}) = c_r & \to & c_{r-1} & \to & \cdots & \to & c_2 & \to & c_1 = \mathrm{rt}(\mathcal{G}), \end{array}$$

while all other edges in $\mathcal{G} = (V, \mathcal{E})$ and $\tilde{\mathcal{G}} = (V, \tilde{\mathcal{E}})$ coincide. Hence, the entropy score difference reduces to

$$\begin{aligned} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) &= h(X_{\mathrm{rt}(\tilde{\mathcal{G}})}) + \sum_{(j,i)\in\tilde{\mathcal{E}}} h(X_i - \mathbb{E}[X_i|X_j]) \\ &\quad - h(X_{\mathrm{rt}(\mathcal{G})}) - \sum_{(j,i)\in\mathcal{E}} h(X_i - \mathbb{E}[X_i|X_j]) \\ &= h(X_{c_r}) + \sum_{i=1}^{r-1} h(X_{c_i} - \mathbb{E}[X_{c_i}|X_{c_{i+1}}]) \\ &\quad - h(X_{c_1}) - \sum_{i=2}^{r} h(X_{c_i} - \mathbb{E}[X_{c_i}|X_{c_{i-1}}]). \end{aligned}$$

Note that

$$h(X_{c_r}) - h(X_{c_1}) = \sum_{i=2}^{r} h(X_{c_i}) - \sum_{i=1}^{r-1} h(X_{c_i}) = \sum_{i=1}^{r-1} h(X_{c_{i+1}}) - h(X_{c_i}).$$

Hence,

$$\begin{aligned} &\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) \\ &= \sum_{i=1}^{r-1} h(X_{c_i} - \mathbb{E}[X_{c_i}|X_{c_{i+1}}]) + h(X_{c_{i+1}}) - h(X_{c_{i+1}} - \mathbb{E}[X_{c_{i+1}}|X_{c_i}]) - h(X_{c_i}) \\ &= \sum_{i=1}^{r-1} \Delta\ell_{\mathrm{E}}(c_i \xleftarrow{\cdot\cdot} c_{i+1}) \\ &\geq \min_{1\leq i\leq r-1} \Delta\ell_{\mathrm{E}}(c_i \xleftarrow{\cdot\cdot} c_{i+1}), \end{aligned}$$

which concludes the proof. ∎

D.4.2 Proof of Theorem 16

We first describe the graphs that result from the reduction technique described in 5.3. To do so, define

$$\mathbb{L}(\mathcal{G}, \tilde{\mathcal{G}}) := \{L \in V_R : \mathrm{ch}^{\mathcal{G}_R}(L) = \emptyset \wedge (\mathrm{pa}^{\tilde{\mathcal{G}}_R}(L) \neq \mathrm{pa}^{\mathcal{G}_R}(L) \vee \mathrm{ch}^{\tilde{\mathcal{G}}_R}(L) \neq \emptyset)\},$$

containing the sink nodes in $\mathcal{G}_R$ that are either not sink nodes in $\tilde{\mathcal{G}}_R$ or sink nodes in $\tilde{\mathcal{G}}_R$ with different parents: $\mathrm{pa}^{\mathcal{G}_R}(L) \neq \mathrm{pa}^{\tilde{\mathcal{G}}_R}(L)$. Now fix any $L \in \mathbb{L}(\mathcal{G}, \tilde{\mathcal{G}}) \subset V_R$ and note that its only parent in $\mathcal{G}_R$, $\mathrm{pa}^{\mathcal{G}_R}(L)$, is either also a parent of L, a child of L or not adjacent to L, in $\tilde{\mathcal{G}}_R$. That is, one and only one of the following sets is non-empty

$$
\begin{aligned}
Z(L) &:= \mathrm{pa}^{\mathcal{G}_R}(L) \cap \mathrm{pa}^{\tilde{\mathcal{G}}_R}(L), && \text{('staying parents')} \\
Y(L) &:= \mathrm{pa}^{\mathcal{G}_R}(L) \cap \mathrm{ch}^{\tilde{\mathcal{G}}_R}(L), && \text{('parents to children')} \\
W(L) &:= \mathrm{pa}^{\mathcal{G}_R}(L) \cap (V \setminus \{L \cup \mathrm{ch}^{\tilde{\mathcal{G}}_R}(L) \cup \mathrm{pa}^{\tilde{\mathcal{G}}_R}(L)\}) && \text{('removing parents')}
\end{aligned}
$$

We define the $\tilde{\mathcal{G}}_R$ parent and children of L that are not adjacent to L in $\mathcal{G}_R$ as

$$
\begin{aligned}
D(L) &:= \mathrm{pa}^{\tilde{\mathcal{G}}_R}(L) \cap (V \setminus \{L \cup \mathrm{ch}^{\mathcal{G}_R}(L) \cup \mathrm{pa}^{\mathcal{G}_R}(L)\}), \text{ and} \\
O(L) &:= \mathrm{ch}^{\tilde{\mathcal{G}}_R}(L) \cap (V \setminus \{L \cup \mathrm{ch}^{\mathcal{G}_R}(L) \cup \mathrm{pa}^{\mathcal{G}_R}(L)\}),
\end{aligned}
$$

respectively. All such sets contain at most one node and by slight abuse of notation, we use the same letters to refer to the nodes. We will henceforth suppress the dependence on L if the choice is clear from the context. Figure 2 visualizes the above sets.

Now partition $\mathcal{T}_p \setminus \{\mathcal{G}\}$ into the three following disjoint partitions for which there exists a reduced graph sink node $L \in \mathbb{L}(\mathcal{G}, \tilde{\mathcal{G}})$ such that $W(L)$, $Y(L)$ and $Z(L)$ is non-empty, respectively. That is, we define

$$
\begin{aligned}
\mathcal{T}_p(\mathcal{G}, W) &:= \{\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\} : \exists L \in \mathbb{L}(\mathcal{G}, \tilde{\mathcal{G}}) \text{ s.t. } W(L) \neq \emptyset\}, \\
\mathcal{T}_p(\mathcal{G}, Y) &:= \{\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\} : \exists L \in \mathbb{L}(\mathcal{G}, \tilde{\mathcal{G}}) \text{ s.t. } Y(L) \neq \emptyset\} \setminus \mathcal{T}_p(\mathcal{G}, W), \\
\mathcal{T}_p(\mathcal{G}, Z) &:= \{\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\} : \exists L \in \mathbb{L}(\mathcal{G}, \tilde{\mathcal{G}}) \text{ s.t. } Z(L) \neq \emptyset\} \setminus (\mathcal{T}_p(\mathcal{G}, W) \cup \mathcal{T}_p(\mathcal{G}, Y)).
\end{aligned}
$$

Using that $\mathcal{T}_p(\mathcal{G}, W) \cup \mathcal{T}_p(\mathcal{G}, Y) \cup \mathcal{T}_p(\mathcal{G}, Z) = \mathcal{T}_p(\mathcal{G})$, we can now find a lower bound for the score gap that holds uniformly over all alternative directed tree graphs $\mathcal{T}_p \setminus \{\mathcal{G}\}$:

$$
\begin{aligned}
\min_{\tilde{\mathcal{G}} \in \mathcal{T}_p \setminus \{\mathcal{G}\}} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) = \min \Bigg\{ & \min_{\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, Z)} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}), \\
& \min_{\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, W)} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}), \min_{\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, Y)} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) \Bigg\}.
\end{aligned}
$$

We now turn to each of these three terms individually and first consider alternative graphs in the partitioning $\mathcal{T}_p(\mathcal{G}, Z)$. The following lower bound consists of possibly non-localized conditional dependence properties of the observable distribution P_X. (That is, the bound may involve nodes that are not close to each other in the graph $\mathcal{G}$.)

Lemma D.7 *Let $\Pi_Z(\mathcal{G})$ denote all tuples $(z, l, o) \in V^3$ of adjacent nodes $(z \to l) \in \mathcal{E}$ for which there exists a node $o \in \mathrm{nd}^{\mathcal{G}}(l) \setminus \{z, l\}$. It holds that*

$$
\min_{\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, Z)} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) \geq \min_{(z,l,o) \in \Pi_Z(\mathcal{G})} I(X_z; X_o | X_l).
$$

The next result proves a lower bound that holds uniformly over all alternative graphs in $\mathcal{T}_p(\mathcal{G}, W)$. The lower bound consists only of local conditional dependence properties. That

is, for any subgraph of the causal graph $\mathcal{G}$ of the form $X_o \to X_w \to X_l$ or $X_o \leftarrow X_w \to X_l$ we measure, by means of conditional mutual information, the conditional dependence of the two adjacent nodes X_w and X_l conditional on X_o, $I(X_w; X_l|X_o)$. The lower bound consists of the smallest of all such local conditional dependence measures.

Lemma D.8 *Let $\Pi_W(\mathcal{G})$ denote all tuples $(w, l, o) \in V^3$ of adjacent nodes $(w \to l) \in \mathcal{E}$ and $o \in (\mathrm{ch}^{\mathcal{G}}(w) \setminus \{l\}) \cup \mathrm{pa}^{\mathcal{G}}(w)$. It holds that that*

$$\min_{\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, W)} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) \geq \min_{(w,l,o) \in \Pi_W(\mathcal{G})} I(X_w; X_l|X_o).$$

A uniform lower bound of the score gap over all alternative graphs in the final partition $\mathcal{T}_p(\mathcal{G}, Y)$ is given by the smallest edge-reversal of any edge in the causal graph $\mathcal{G}$.

Lemma D.9 *It holds that*

$$\min_{\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, Y)} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) \geq \min_{(j \to i) \in \mathcal{E}} \Delta\ell_{\mathrm{E}}(j \xleftrightarrow{} i).$$

An immediate consequence of Lemmas D.7 to D.9 is that the entropy identifiability gap is given by the smallest of the lower bounds derived for each partition, see Theorem 16. Thus, it only remains to prove Lemmas D.7 to D.9.

Proof of Lemma D.7. Let $\tilde{\mathcal{G}} \in \Pi_Z(\mathcal{G})$ such that $Z \neq \emptyset$. This implies that $Y = W = \emptyset$ as L can only have one parent in $\mathcal{G}$. Furthermore, $D = \emptyset$ as L can only have one parent in $\tilde{\mathcal{G}}$ and $O \neq \emptyset$ for otherwise L would have been deleted by the deletion procedure in Section 5. Assume without loss of generality that $O = \{O_1, \ldots, O_k\}$ for some $k \in \mathbb{N}$. The two subgraphs are illustrated in Figure 16.

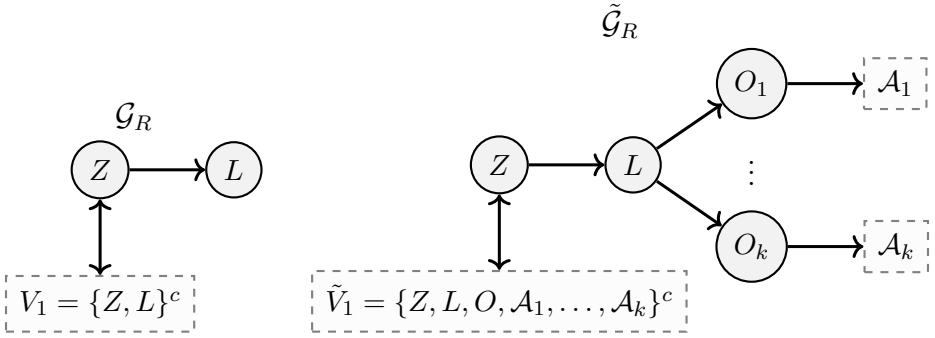

Figure 16: Illustration of the reduced form graphs $\mathcal{G}_R$ and $\tilde{\mathcal{G}}_R$ for the case $\tilde{\mathcal{G}} \in \Pi_Z(\mathcal{G})$. $\mathcal{A}_1, \ldots, \mathcal{A}_k$ are possibly empty sets of nodes, and dashed rectangle nodes denotes a possibly multi-node subgraph over the variables enclosed. The bi-directed edges means that the edge can be directed in both directions. An edge pointing into the multi-node subgraph, can possibly be multiple edges into distinct nodes of the subgraph.

For ease of notation, fix any $1 \leq i \leq k$ and denote $O := O_i$. We note that in $\tilde{\mathcal{G}}$ the following d-separation holds

$$Z \perp\!\!\!\perp_{\tilde{\mathcal{G}}} O \,|\, L.$$

Thus, we have for all probability measures $Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p$ over nodes V that $Z \perp\!\!\!\perp O \,|\, L$ (as the path between Z and O is blocked by L and all probability measures generated in accordance with an SCM are Markovian with respect to the generating graph $\tilde{\mathcal{G}}$). Recall that

$$\begin{aligned}\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) &= \inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p} D_{\mathrm{KL}}(P_X \| Q) \\ &= \inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p} h(P_X, Q) - h(P_X).\end{aligned}$$

Now fix $Q = q \cdot \lambda^p \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p$ and note that it factorizes as $Q = Q_{A|Z,O,L} Q_{Z|L} Q_{O|L} Q_L$, i.e., the density q factorizes as

$$\begin{aligned}q(x) &= q_{A|Z,O,L}(a|z,o,l) q_{Z,O,L}(z,o,l) \\ &= q_{A|Z,O,L}(a|z,o,l) q_{Z|L}(z|l) q_{O|L}(o|l) q_L(l),\end{aligned}$$

for λ^p-almost all $x = (a,z,o,l) \in \mathbb{R}^p$ where $A = V \setminus \{Z, O, L\}$. Hence, the cross entropy splits additively into

$$\begin{aligned}h(P_X, Q) \geq\ & \mathbb{E}[-\log(q_{A|Z,O,L}(A|Z,O,L))] \\ &+ \mathbb{E}[-\log(q_{Z|L}(Z|L))] \\ &+ \mathbb{E}[-\log(q_{O|L}(O|L))] \\ &+ \mathbb{E}[-\log(q_L(L))]. \end{aligned} \tag{43}$$

Now note, e.g., that for a conditional distribution (Markov kernel) $Q_{Z|L}$ it holds that

$$\begin{aligned}0 \leq D_{\mathrm{KL}}(P_{Z|L} P_L \| Q_{Z|L} P_L) &= \mathbb{E}\left[-\log\left(\frac{q_{Z|L}(Z|L) p_L(L)}{p_{Z|L}(Z|L) p_L(L)}\right)\right] \\ &= \mathbb{E}[-\log(q_{Z|L}(Z|L))] - \mathbb{E}[-\log(p_{Z|L}(Z|L))],\end{aligned}$$

proving that

$$\mathbb{E}[-\log(q_{Z|L}(Z|L))] \geq \mathbb{E}[-\log(p_{Z|L}(Z|L))].$$

By similar arguments, we get that the three other terms in the lower bound of Equation (43) are bounded below by

$$\begin{aligned}\mathbb{E}[-\log(q_{A|Z,O,L}(A|Z,O,L))] &\geq \mathbb{E}[-\log(p_{A|Z,O,L}(A|Z,O,L))], \\ \mathbb{E}[-\log(q_{O|L}(O|L))] &\geq \mathbb{E}[-\log(p_{O|L}(O|L))], \\ \mathbb{E}[-\log(q_L(L))] &\geq \mathbb{E}[-\log(p_L(L))].\end{aligned}$$

This implies that

$$\inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p} h(P_X, Q) \geq h(P_X, Q^*),$$

where $Q^* = P_{A|Z,O,L}P_{Z|L}P_{O|L}P_L$. On the other hand, we know that P_X factorizes as $P_X = P_{A|Z,O,L}P_{Z,O|L}P_L$. Thus we have the following entropy score gap lower bound

$$\begin{aligned}\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) &\geq h(P_X, Q^*) - h(P_X) \\ &= D_{\mathrm{KL}}(P_X \| Q^*) \\ &= D_{\mathrm{KL}}(P_{A|Z,O,L}P_{Z,O|L}P_L \| P_{A|Z,O,L}P_{Z|L}P_{O|L}P_L) \\ &= D_{\mathrm{KL}}(P_{Z,O|L}P_L \| P_{Z|L}P_{O|L}P_L) \\ &= D_{\mathrm{KL}}(P_{Z,O|L} \| P_{Z|L}P_{O|L} | P_L) \\ &= I(Z; O|L).\end{aligned}$$

$\Pi_Z(\mathcal{G})$ denotes all tuples $(z, l, o) \in V^3$ of adjacent nodes $(z \to l) \in \mathcal{E}$ for which there exists a node $o \in \mathrm{nd}^{\mathcal{G}}(l) \setminus \{z, l\}$. For any graph $\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, Z)$ we can, by the above considerations, find a tuple $(z, l, o) \in \Pi_Z(\mathcal{G})$ such that

$$\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) \geq I(X_o; X_z \,|\, X_l).$$

We conclude that

$$\min_{\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, Z)} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) \geq \min_{(z,l,o) \in \Pi_Z(\mathcal{G})} I(X_o; X_z | X_l).$$

■

Proof of Lemma D.8. Fix any $\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, W)$ and L with $W \neq \emptyset$ such that $Z = Y = \emptyset$. We have illustrated the subgraph $\mathcal{G}_R$ in Figure 17 and the possible subgraphs $\tilde{\mathcal{G}}_R$ in Figure 18.

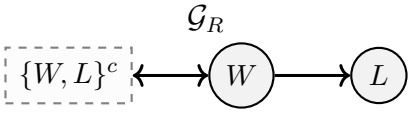

Figure 17: Illustrations of the $\mathcal{G}_R$ subgraph for for $\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, W)$.

Note that for any of the three possible local graph structures presented in Figure 18 there exists an $A \in \{O_1, \ldots, O_k, D\}$ such that $L \perp\!\!\!\perp_{\tilde{\mathcal{G}}_R} W \,|\, A$, i.e., A blocks the path between L and W. Thus, for all probability measures $Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p$ over nodes $V = \{1, .., p\}$ it always holds that $L \perp\!\!\!\perp W \,|\, A$. By arguments similar to those in the proof of Lemma D.7, we note that

$$\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) = \inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p} h(P_X, Q) - h(P_X),$$

and that

$$\inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p} h(P_X, Q) \geq h(P_X, Q^*),$$

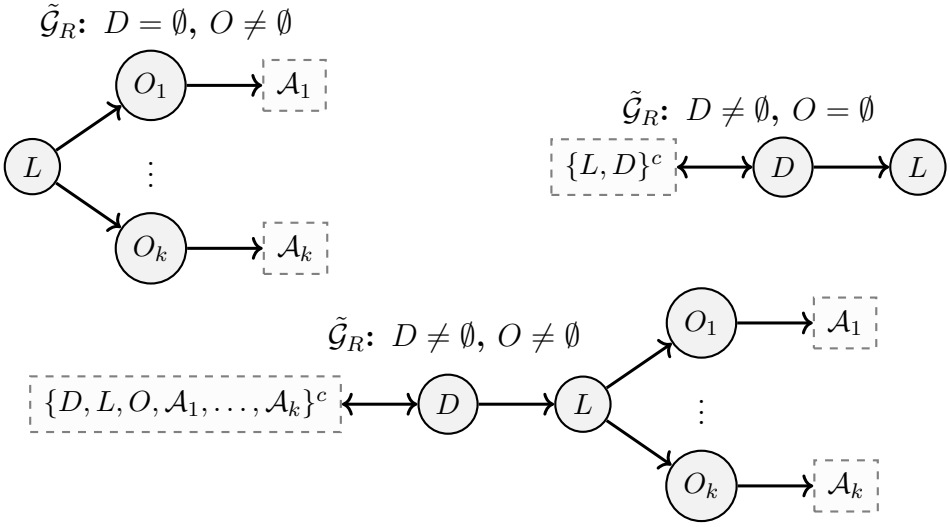

Figure 18: Illustrations of the possible $\tilde{\mathcal{G}}_R$ subgraphs for $\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, W)$.

for $P_X = P_{K|W,L,A} P_{W,L|A} P_A$ and $Q^* = P_{K|W,L,A} P_{L|A} P_{W|A} P_A$ where $K = V \setminus \{W, L, A\}$. To that end, we now have that

$$\begin{aligned}
\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) &\geq h(P_X, Q^*) - h(P_X) \\
&= D_{\mathrm{KL}}(P_X \| Q^*) \\
&= D_{\mathrm{KL}}(P_{K|W,L,A} P_{W,L|A} P_A \| P_{K|W,L,A} P_{L|A} P_{W|A} P_A) \\
&= D_{\mathrm{KL}}(P_{W,L|A} P_A \| P_{L|A} P_{W|A} P_A) \\
&= D_{\mathrm{KL}}(P_{W,L|A} \| P_{L|A} P_{W|A} | P_A) \\
&= I(W; L | A).
\end{aligned}$$

Let $\hat{\Pi}_W(\mathcal{G})$ denote all tuples $(w, l, a) \in V^3$ of adjacent nodes $(w \to l) \in \mathcal{E}$ for which there exists a node $a \in \mathrm{nd}^{\mathcal{G}}(l) \setminus \{w\}$. Now note that for any graph $\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, W)$ we can, by the above considerations, find a tuple $(w, l, a) \in \hat{\Pi}_W(\mathcal{G})$ such that

$$\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) \geq I(X_w; X_l \,|\, X_a). \tag{44}$$

(Conversely for any tuple $(w, l, a) \in \hat{\Pi}_W(\mathcal{G})$ we can construct a graph $\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, W)$ such that (44) holds. To see this, fix $(w, l, a) \in \hat{\Pi}_W(\mathcal{G})$ and construct $\tilde{\mathcal{G}}$ such that the subtree with root node l is identical in both $\mathcal{G}$ and $\tilde{\mathcal{G}}$ and a blocks the path between l and w in $\tilde{\mathcal{G}}$.) Therefore, the following lower bound holds (and it is not unnecessarily small).

$$\min_{\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, W)} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) \geq \min_{(w,l,a) \in \hat{\Pi}_W(\mathcal{G})} I(X_w; X_l \,|\, X_a).$$

For any $(w, l, a) \in \hat{\Pi}_W(\mathcal{G})$ it either holds that $a \in (\mathrm{ch}^{\mathcal{G}}(w) \setminus \{l\}) \cup \mathrm{pa}^{\mathcal{G}}(w)$ or that there exists an $o \in (\mathrm{ch}^{\mathcal{G}}(w) \setminus \{l\}) \cup \mathrm{pa}^{\mathcal{G}}(w)$ blocking the path between a and l in $\mathcal{G}$ such that

$X_l \perp\!\!\!\perp X_a | X_o$. Furthermore, we note that as $X_l \perp\!\!\!\perp (X_o, X_a) \,|\, X_w$ we have that

$$\begin{aligned} I(X_w; X_l | X_a) &= h(X_l|X_a) - h(X_l|X_a, X_w) \\ &= h(X_l|X_a) - h(X_l|X_w) \\ &= h(X_l|X_a) - h(X_l|X_o, X_w) \\ &\geq h(X_l|X_a, X_o) - h(X_l|X_o, X_w) \\ &= h(X_l|X_o) - h(X_l|X_o, X_w) \\ &= I(X_w; X_l \,|\, X_o), \end{aligned}$$

as further conditioning reduces conditional entropy. Let $\Pi_W(\mathcal{G})$ denote all tuples $(w, l, o) \in V^3$ of adjacent nodes $(w \to l) \in \mathcal{E}$ and $o \in (\mathrm{ch}^{\mathcal{G}}(w) \setminus \{l\}) \cup \mathrm{pa}^{\mathcal{G}}(w)$. By the above considerations we conclude that

$$\min_{\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, W)} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) \geq \min_{(w,l,o) \in \Pi_W(\mathcal{G})} I(X_w; X_l \,|\, X_o).$$

■

Proof of Lemma D.9. Fix $\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, Y)$ and L such that $Y \neq \emptyset$. It holds that $W = Z = \emptyset$. We have illustrated the $\mathcal{G}_R$ in Figure 19 and the three possible subgraphs $\tilde{\mathcal{G}}_R$ in Figure 20.

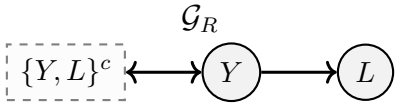

Figure 19: Illustrations of the $\mathcal{G}_R$ subgraph for $\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, Y)$.

Note that for any of the three possible local graph structures of $\tilde{\mathcal{G}}_R$ illustrated in Figure 20 we have that for all probability measures $Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p$ factorizes as $Q_{A|L,Y} Q_{L,Y}$, where $A = V \setminus \{L, Y\}$. It always holds that $Q_{L,Y}$ is the distribution of $(\tilde{L}, \tilde{Y})$ generated in accordance with a structural causal model of the form

$$\tilde{Y} := \tilde{f}_Y(\tilde{L}) + \tilde{N}_Y, \tag{45}$$

where $\tilde{f}_Y(l) = \mathbb{E}[Y|L = l]$ for all $l \in \mathbb{R}$, and any $\mathcal{L}(\tilde{N}_Y), \mathcal{L}(\tilde{L}) \in \mathcal{P}$ with $\tilde{N}_Y \perp\!\!\!\perp \tilde{L}$. Now recall that

$$\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) = \inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p} h(P_X, Q) - h(P_X),$$

and notice that by arguments similar to those in the proof of Lemma D.7 we get

$$\begin{aligned} h(P_X, Q) &= h(P_X, Q_{A|L,Y} Q_{L,Y}) \\ &= \mathbb{E}[-\log(q_{A|L,Y}(A|L,Y))] + h(P_{L,Y}, Q_{L,Y}) \\ &\geq \mathbb{E}[-\log(p_{A|L,Y}(A|L,Y))] + h(P_{L,Y}, Q_{L,Y}), \end{aligned}$$

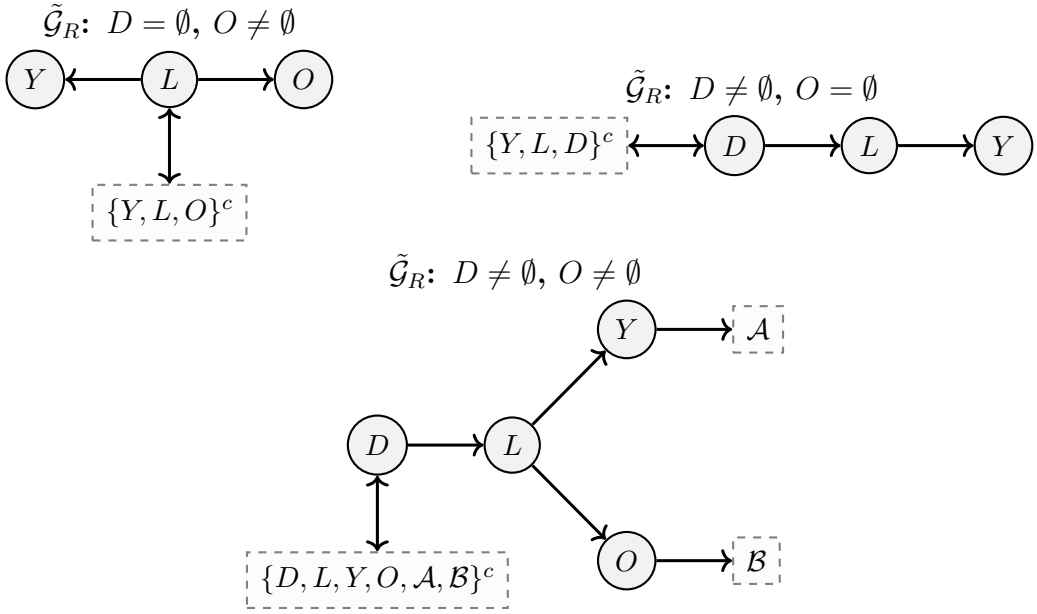

Figure 20: Illustrations of the possible $\tilde{\mathcal{G}}_R$ subgraphs for $\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, Y)$.

and that $h(P_X) = \mathbb{E}[-\log(p_{A|L,Y}(A|L,Y))] + h(P_{L,Y})$. Thus, we have that

$$\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) \geq \inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p} h(P_{L,Y}, Q_{L,Y}) - h(P_{L,Y}).$$

For any $Q = Q_{A|L,Y} Q_{L,Y} \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p$ we have that $Q_{L,Y}$ is uniquely determined by a marginal distribution $Q_L \in \mathcal{P}$ and the noise distribution of $\tilde{N}_Y \sim q_{\tilde{N}_Y} \cdot \lambda \in \mathcal{P}$ from the additive noise structural assignment in Equation (45) for $\tilde{Y}$ and the causal function $\tilde{f}_Y$. Thus, the density $q_{L,Y}$ of $Q_{L,Y}$ is given by

$$q_{L,Y}(l, y) = q_{Y|L}(y|l) q_L(l) = q_{\tilde{N}_Y}(y - \tilde{f}_Y(l)) q_L(l) = q_{\tilde{N}_Y}(y - \mathbb{E}[Y|L = l]) q_L(l).$$

Hence,

$$\begin{aligned}
h(P_{L,Y}, Q_{L,Y}) &= \mathbb{E}\left[-\log\left(q_{L,Y}(L, Y)\right)\right] \\
&= \mathbb{E}\left[-\log\left(q_{Y|L}(Y|L)\right)\right] + \mathbb{E}\left[-\log\left(q_L(L)\right)\right] \\
&= \mathbb{E}\left[-\log\left(q_{\tilde{N}_Y}(Y - \mathbb{E}[Y|L])\right)\right] + h(P_L, Q_L) \\
&= h(Y - \mathbb{E}[Y|L], \tilde{N}_Y) + h(P_L, Q_L) \\
&\geq h(Y - \mathbb{E}[Y|L]) + h(L),
\end{aligned}$$

where we used that $h(P, Q) = D_{\mathrm{KL}}(P, Q) + h(P) \geq h(P)$. Thus, we have that

$$\begin{aligned}
\ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) &\geq \inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p} h(P_{L,Y}, Q_{L,Y}) - h(P_{L,Y}) \\
&\geq h(Y - \mathbb{E}[Y|L]) + h(L) - h(L - \mathbb{E}[L|Y]) - h(Y) \\
&= \Delta \ell_{\mathrm{E}}(Y \xleftarrow{\cdot} L).
\end{aligned}$$

We conclude that

$$\min_{\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G},Y)} \ell_{\mathrm{E}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{E}}(\mathcal{G}) \geq \min_{(i \to j) \in \mathcal{E}} \Delta \ell_{\mathrm{E}}(j \overset{\leftarrow\!\!-\!\!\rightarrow}{} i).$$

■

D.4.3 Remaining proof of Section 5

Proof of Theorem 17.

Consider a graph $\tilde{\mathcal{G}} \in \mathcal{T}_p(\mathcal{G}, Z)$ and let $\mathcal{G}_{R,1} = (\mathcal{E}_{R,1}, V_{R,1})$ and $\tilde{\mathcal{G}}_{R,1} = (\tilde{\mathcal{E}}_{R,1}, V_{R,1})$ be the reduced graphs after the initial edge and node deletion procedure of Section 5.3. The deletion procedure does not change the score gap, that is,

$$\ell_{\mathrm{G}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{G}}(\mathcal{G}) = \ell_{\mathrm{G}}(\tilde{\mathcal{G}}_{R,1}) - \ell_{\mathrm{G}}(\mathcal{G}_{R,1}).$$

For any $i \geq 1$ and fixed $\mathcal{G}_{R,i}$ and $\tilde{\mathcal{G}}_{R,i}$ we define

$$\mathbb{L}_{R,i} := \{L \in V_{R,i} : \mathrm{ch}^{\mathcal{G}_{R,i}}(L) = \emptyset \wedge (\mathrm{pa}^{\tilde{\mathcal{G}}_{R,i}}(L) \neq \mathrm{pa}^{\mathcal{G}_{R,i}}(L) \vee \mathrm{ch}^{\tilde{\mathcal{G}}_{R,i}}(L) \neq \emptyset)\}.$$

Now fix $L_1 \in \mathbb{L}_{R,1}$ such that $Z_1 \neq \emptyset$, where Y_1, Z_1, W_1, D_1 and O_1 are defined similarly to the variables in Section 5. Let $O_1 = \{O_{1,1}, \ldots, O_{1,k_1}\}$, for some $k_1 \in \mathbb{N}$.

Assume that there exists an $i \in \{1, \ldots, k_1\}$ such that $(Z_1 \to O_{1,i}) \in \mathcal{E}_{R,1}$ in which case we have the following two paths in $\mathcal{G}_{R,1}$ and $\tilde{\mathcal{G}}_{R,1}$

$$\mathcal{G}_{R,1} : O_{1,i} \leftarrow Z_1 \to L_1, \quad \text{and} \quad \tilde{\mathcal{G}}_{R,1} : Z_1 \to L_1 \to O_{1,i}.$$

Since $O_{1,i} \perp\!\!\!\perp_{\tilde{\mathcal{G}}_{R,1}} Z_1 \,|\, L_1$, an entropy score gap lower bound is given by

$$\begin{aligned}
\ell_{\mathrm{G}}(\tilde{\mathcal{G}}_{R,1}) - \ell_{\mathrm{G}}(\mathcal{G}_{R,1}) &\geq \ell_{\mathrm{E}}(\tilde{\mathcal{G}}_{R,1}) - \ell_{\mathrm{E}}(\mathcal{G}_{R,1}) \\
&= \inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{F}(\tilde{\mathcal{G}}) \times \mathcal{P}^p} h(P_X, Q) - P_X \\
&\geq D_{\mathrm{KL}}(P_X \| Q^*) \\
&= I(O_{1,i}; Z_1 | L_1),
\end{aligned}$$

with $P_X = P_{K|O,Z,L} P_{O,Z|L} P_L$ and $Q^* = P_{K|O,Z,L} P_{Z|L} P_{O|L} P_L$ for $K = V \setminus \{O, Z, L\}$, by arguments similar to those from the proof of Lemma D.8. Now note that $(Z_1, O_{1,i}, L_1) \in \Pi_W(\mathcal{G}_{R,1}) \subseteq \Pi_W(\mathcal{G})$ as $(Z_1 \to O_{1,i}) \in \mathcal{E}_{R,1}$ and $L_1 \in \mathrm{ch}^{\mathcal{G}_{R,1}}(Z_1) \setminus \{O_{1,i}\} \subseteq (\mathrm{ch}^{\mathcal{G}_{R,1}}(Z_1) \setminus \{O_{1,i}\}) \cup \mathrm{pa}^{\mathcal{G}_{R,1}}(Z_1)$. Hence,

$$\ell_{\mathrm{G}}(\tilde{\mathcal{G}}_{R,1}) - \ell_{\mathrm{G}}(\mathcal{G}_{R,1}) \geq \min_{(w,l,o) \in \Pi_W(\mathcal{G})} I(X_w; X_l | X_o). \tag{46}$$

Assume now that for all $i \in \{1, .., k_1\}$ we have $(Z_1 \to O_{1,i}) \notin \mathcal{E}_{R,1}$. Let $\hat{\mathcal{G}}_{R,1} = (\hat{\mathcal{E}}_{R,1}, V_{R,1})$ denote an intermediate graph where $\hat{\mathcal{E}}_{R,1}$ is identical to $\tilde{\mathcal{E}}_{R,1}$ except the edges $\{(L_1 \to O_{1,i}) : 1 \leq i \leq k_1\} \subset \tilde{\mathcal{E}}_{R,1}$ are replaced by the edges $\{(Z_1 \to O_{1,i}) : 1 \leq i \leq k_1\}$. It holds that

$$\begin{aligned}
\ell_{\mathrm{G}}(\tilde{\mathcal{G}}_{R,1}) - \ell_{\mathrm{G}}(\mathcal{G}_{R,1}) &= \ell_{\mathrm{G}}(\tilde{\mathcal{G}}_{R,1}) - \ell_{\mathrm{G}}(\hat{\mathcal{G}}_{R,1}) + \ell_{\mathrm{G}}(\hat{\mathcal{G}}_{R,1}) - \ell_{\mathrm{G}}(\mathcal{G}_{R,1}) \\
&\geq \ell_{\mathrm{G}}(\hat{\mathcal{G}}_{R,1}) - \ell_{\mathrm{G}}(\mathcal{G}_{R,1}).
\end{aligned}$$

Note that this score gap lower bound is still strictly positive as $\hat{\mathcal{G}}_{R,1} \neq \mathcal{G}_{R,1}$. To realize the last inequality, simply note that as $O_{1,i} \perp\!\!\!\perp L_1 \,|\, Z_1$ we have for all $i \in \{1, \dots, k_1\}$ that

$$\begin{aligned} 2\ell_{\mathrm{G}}(\tilde{\mathcal{G}}_{R,1}, O_{1,i}) &= \log \mathbb{E}[(O_{1,i} - \mathbb{E}[O_{1,i}|L_1])^2] \\ &\geq \log \mathbb{E}[(O_{1,i} - \mathbb{E}[O_{1,i}|Z_1, L_1])^2] \\ &= \log \mathbb{E}[(O_{1,i} - \mathbb{E}[O_{1,i}|Z_1])^2] \\ &= 2\ell_{\mathrm{G}}(\hat{\mathcal{G}}_{R,1}, O_{1,i}). \end{aligned} \tag{47}$$

Now since all edges in $\tilde{\mathcal{G}}_{R,1}$ and $\hat{\mathcal{G}}_{R,1}$ coincide except the incoming edges into $O_{1,1}, \dots, O_{1,k_1}$ we get that

$$\ell_{\mathrm{G}}(\tilde{\mathcal{G}}_{R,1}) - \ell_{\mathrm{G}}(\hat{\mathcal{G}}_{R,1}) = \sum_{i=1}^{k_1} \ell_{\mathrm{G}}(\tilde{\mathcal{G}}_{R,1}, O_{1,i}) - \ell_{\mathrm{G}}(\hat{\mathcal{G}}_{R,1}, O_{1,i}) \geq 0,$$

where the inequality follows from Equation (47). Now both $\hat{\mathcal{G}}_{R,1}$ and $\mathcal{G}_{R,1}$ have a childless node L_1 with the same parent Z_1, so we let $\tilde{\mathcal{G}}_{R,2}$ and $\mathcal{G}_{R,2}$ denote these two graphs where the node L_1 and its incoming edge are deleted. This deletion does not change the graph scores, i.e.,

$$\ell_{\mathrm{G}}(\hat{\mathcal{G}}_{R,1}) - \ell_{\mathrm{G}}(\mathcal{G}_{R,1}) = \ell_{\mathrm{G}}(\tilde{\mathcal{G}}_{R,2}) - \ell_{\mathrm{G}}(\mathcal{G}_{R,2}).$$

Now fix $L_2 \in \mathbb{L}_{R,2}$ and define Y_2, Z_2, W_2, D_2 and $O_2 = \{O_{2,1}, \dots, O_{2,k_2}\}$ accordingly.

If either Y_2 or W_2 is non-empty, we use the score gap lower bound previously discussed in Lemma D.8 and Lemma D.9. If Z_2 is non-empty, we can repeat the above procedure and iteratively move edges and delete nodes until we arrive at the first $i \in \mathbb{N}$ with $\tilde{\mathcal{G}}_{R,i}$ and $\mathcal{G}_{R,i}$ being the iteratively reduced graphs and $L_{R,i} \in \mathbb{L}_{R,i}$ where either

i) Y_i or W_i is non-empty, here, we get that $\ell_{\mathrm{G}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{G}}(\mathcal{G})$ is lower bounded by a bound similar to the form of Lemma D.8 or Lemma D.9. That is,

$$\begin{aligned} \ell_{\mathrm{G}}(\tilde{\mathcal{G}}_{R,i}) - \ell_{\mathrm{G}}(\mathcal{G}_{R,i}) &\geq \ell_{\mathrm{E}}(\tilde{\mathcal{G}}_{R,i}) - \ell_{\mathrm{E}}(\mathcal{G}_{R,i}) \\ &\geq \min\left\{ \min_{j \to i \in \mathcal{E}} \Delta\ell_{\mathrm{E}}(i \xleftrightarrow{} j), \min_{(w,l,o) \in \Pi_W(\mathcal{G})} I(X_w; X_l \,|\, X_o) \right\}. \end{aligned}$$

ii) Z_i is non-empty and there exists a $j \in \{1, \dots, k_i\}$ such that $(Z_i \to O_{i,j}) \in \mathcal{G}_{R,i}$. As previously argued, the score gap lower bound of Equation (46) applies. That is

$$\ell_{\mathrm{G}}(\tilde{\mathcal{G}}_{R,i}) - \ell_{\mathrm{G}}(\mathcal{G}_{R,i}) \geq \ell_{\mathrm{E}}(\tilde{\mathcal{G}}_{R,i}) - \ell_{\mathrm{E}}(\mathcal{G}_{R,i}) \geq \min_{(w,l,o) \in \Pi_W(\mathcal{G})} I(X_w; X_l \,|\, X_o).$$

Note that whenever we do not meet scenario i) or ii) we remove a node in both graphs that is a sink node in the reduced true causal graph $\mathcal{G}_{R,i}$ and the intermediate graph $\hat{\mathcal{G}}_{R,i}$. After at most $p-2$ graph reduction iterations of not encountering scenario i) or ii) we are left with two different graphs on two nodes, in which case the score gap is an edge reversal. We conclude that

$$\ell_{\mathrm{G}}(\tilde{\mathcal{G}}) - \ell_{\mathrm{G}}(\mathcal{G}) \geq \min\left\{ \min_{i \to j \in \mathcal{E}} \Delta\ell_{\mathrm{E}}(j \xleftrightarrow{} i), \min_{(w,l,o) \in \Pi_W(\mathcal{G})} I(X_w; X_l \,|\, X_o) \right\}.$$

■

References

T. B. Berrett and R. J. Samworth. Nonparametric independence testing via mutual information. *Biometrika*, 106(3):547–566, 2019.

T. B. Berrett, D. Grose, and R. J. Samworth. CRAN R-package 'IndepTest': Nonparametric independence tests based on entropy estimation, 2018. URL `https://cran.r-project.org/web/packages/IndepTest`.

T. B. Berrett, R. J. Samworth, and M. Yuan. Efficient multivariate entropy estimation via k-nearest neighbour distances. *The Annals of Statistics*, 47(1):288 – 318, 2019.

P. Bühlmann, J. Peters, and J. Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526 – 2556, 2014.

P. M. Camerini, L. Fratta, and F. Maffioli. A note on finding optimum branchings. *Networks*, 9(4):309–312, 1979. doi: 10.1002/net.3230090403.

V. Carey, L. Long, and R. Gentleman. Bioconductor R-package 'RBGL', 2021. URL `https://www.bioconductor.org/packages/release/bioc/html/RBGL.html`.

A. Cayley. A theorem on trees. *Quart. J. Math.*, 23:376–378, 1889.

D. M. Chickering. Learning Bayesian networks is NP-complete. In D. Fisher and H.-J. Lenz, editors, *Learning from Data: Artificial Intelligence and Statistics V*, pages 121–130. Springer New York, New York, NY, 1996.

D. M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–554, 2002.

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

Y. J. Chu and T. H. Liu. On the shortest arborescence of a directed graphs. *Science Sinica*, 14:1396–1400, 1965.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Hoboken, New Jersey, 2006.

J. Cussens. Bayesian network learning with cutting planes. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 153–160. AUAI Press, 2011.

J. Cussens, M. Järvisalo, J. H. Korhonen, and M. Bartlett. Bayesian network structure learning with integer programming: Polytopes, facets and complexity. *Journal of Artificial Intelligence Research*, 58:185–229, 2017.

I. S. Dominguez, A. H. Aguirre, and E. V. Diharce. The Gaussian polytree eda with copula functions and mutations. In *EVOLVE-A Bridge between Probability, Set Oriented Numerics and Evolutionary Computation*, pages 123–153. Springer, Berlin, Heidelberg, 2013.

J. Edmonds. Optimum branchings. *Journal of Research of the National Bureau of Standards B*, 71(4):233–240, 1967.

N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine learning*, 29(2):131–163, 1997.

H. N. Gabow, Z. Galil, T. Spencer, and R. E. Tarjan. Efficient algorithms for finding minimum spanning trees in undirected and directed graphs. *Combinatorica*, 6(2):109–122, 1986.

D. Geiger and D. Heckerman. Learning Gaussian networks. In *Proceedings of the Tenth Annual Conference on Uncertainty in Artificial Intelligence*, pages 235–243, 1994.

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-free Theory of Nonparametric Regression*, volume 1. Springer, Berlin, Germany, 2002.

A. Hagberg, P. Swart, and D. Schult. Python package 'NetworkX', 2022. URL `github.com/NetworkX/NetworkX`.

Y. Han, J. Jiao, T. Weissman, and Y. Wu. Optimal rates of entropy estimation over Lipschitz balls. *The Annals of Statistics*, 48(6):3228 – 3250, 2020.

D. Heckerman and D. Geiger. Learning Bayesian networks: A unification for discrete and Gaussian domains. In *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence*, pages 274–284, 1995.

P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in Neural Information Processing Systems*, 21: 689–696, 2008.

T. Jaakkola, D. Sontag, A. Globerson, and M. Meila. Learning Bayesian network structure using lp relaxations. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 358–365. JMLR Workshop and Conference Proceedings, 2010.

M. Kalisch and P. Bühlman. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8(22):613–636, 2007.

M. Kanagawa, P. Hennig, D. Sejdinovic, and B. K. Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

R. M. Karp. A simple derivation of edmonds' algorithm for optimum branchings. *Networks*, 1(3):265–272, 1971.

M. Koivisto and K. Sood. Exact Bayesian structure discovery in Bayesian networks. *The Journal of Machine Learning Research*, 5:549–573, 2004.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, Massachusetts, 2009.

C. Lee and P. v. Beek. Metaheuristics for score-and-search Bayesian network structure learning. In *Canadian Conference on Artificial Intelligence*, pages 129–141. Springer, 2017.

A. Marsiglietti and V. Kostina. A lower bound on the differential entropy of log-concave random vectors with applications. *Entropy*, 20(3):185, 2018.

J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016.

P. Nandy, A. Hauser, and M. H. Maathuis. High-dimensional consistency in score-based and hybrid structure learning. *The Annals of Statistics*, 46(6A):3151–3183, 2018.

M. Ouerd. *Learning in belief networks and its application to distributed databases.* PhD Thesis, University of Ottawa, Ottawa, Canada, 2000.

L. Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6): 1191–1253, 2003.

P. Parviainen and M. Koivisto. Exact structure discovery in Bayesian networks with less space. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, page 436–443. AUAI Press, 2009.

J. Pearl. *Causality.* Cambridge University Press, Cambridge, UK, 2009.

J. Peters and P. Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

J. Peters and P. Bühlmann. Structural intervention distance for evaluating causal graphs. *Neural Computation*, 27(3):771–799, 2015.

J. Peters, D. Janzing, and B. Schölkopf. Causal inference on discrete data using additive noise models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12): 2436–2450, 2011.

J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(58):2009–2053, 2014.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms.* MIT Press, Cambridge, Massachusetts, 2017.

Y. Polyanskiy and Y. Wu. Lecture notes on information theory, 2019. URL `people.lids.mit.edu/yp/homepage/`. last accessed, 9.3.2022.

G. Rebane and J. Pearl. The recovery of causal poly-trees from statistical data. In *Proceedings of the Third Annual Conference on Uncertainty in Artificial Intelligence*, pages 222–228, 1987.

A. Reisach, C. Seiler, and S. Weichwald. Beware of the simulated DAG! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34, 2021.

D. Rothenhäusler, J. Ernest, and P. Bühlmann. Causal inference in partially linear structural equation models. *Annals of Statistics*, 46(6A):2904–2938, 2018.

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*, 308(5721):523–529, 2005.

M. Scanagatta, C. P. de Campos, G. Corani, and M. Zaffalon. Learning Bayesian networks with thousands of variables. *Advances in neural information processing systems*, 28, 2015.

R. L. Schilling. *Measures, Integrals and Martingales*. Cambridge University Press, Cambridge, UK, 2017.

R. D. Shah and J. Peters. The hardness of conditional independence testing and the generalised covariance measure. *Annals of Statistics*, 48(3):1514–1538, 2020.

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

T. Silander and P. Myllymäki. A simple approach for finding the globally optimal Bayesian network structure. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 445–452. AUAI Press, 2006.

E. I. Silva. *A unified framework for the analysis and design of networked control systems*. PhD Thesis, University of Newcastle, Callaghan, Australia, 2009.

N. J. A. Sloane. The on-line encyclopedia of integer sequences, 2021. URL `oeis.org/A003024`. The OEIS Foundation Inc. (2021).

P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman. *Causation, Prediction, and Search*. MIT press, Cambridge, Massachusetts, 2000.

R. E. Tarjan. Finding optimum branchings. *Networks*, 7(1):25–35, 1977.

A. Tofigh and E. Sjölund. C++ implementation of Edmonds algorithm, 2007. URL `github.com/atofigh/edmonds-alg`.

I. Tsamardinos, L. E. Brown, and C. F. Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, 2006.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, Berlin, Germany, 2009.

C. Uhler, G. Raskutti, P. Bühlmann, and B. Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, 41(2):436–463, 2013. doi: 10.1214/12-aos1080.

P. van Beek and H.-F. Hoffmann. Machine learning of Bayesian networks using constraint programming. In G. Pesant, editor, *Principles and Practice of Constraint Programming*, pages 429–445, Cham, 2015. Springer International Publishing.

A. W. Van der Vaart. *Asymptotic statistics*. Cambridge University Press, Cambridge, UK, 3 edition, 2000.

S. Wood. CRAN R-package 'mgcv': Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation, 2022. URL `cran.r-project.org/web/packages/mgcv`.

S. N. Wood. Thin-plate regression splines. *Journal of the Royal Statistical Society (B)*, 65 (1):95–114, 2003.

C. Yuan and B. Malone. Learning optimal Bayesian networks: A shortest path perspective. *Journal of Artificial Intelligence Research*, 48:23–65, 2013.

C. Yuan, B. Malone, and X. Wu. Learning optimal Bayesian networks using a* search. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

J. Zhang and P. Spirtes. Strong faithfulness and uniform consistency in causal inference. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, page 632–639, 2002.

K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, page 647–655, 2009.

X. Zheng, B. Aragam, P. Ravikumar, and E. P. Xing. DAGs with no tears: Continuous optimization for structure learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, page 9492–9503, 2018.