Structure Learning for Directed Trees

Jakobsen, M. E., Shah, R. D., Bühlmann, P., Peters, J. (2022)

Jieun Shin

December 7, 2022

Table of Contents

1. Introduction

1. Introduction

- In structural causal models, one assumes that there are (causal) functions $f_1, \ldots, f_p$ such that for all

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

  for subsets $\mathrm{PA}(i) \subset \{1, \ldots, p\}$ and jointly independent noise variable $N = (N_1, \ldots, N_p) \sim P_N$.
- For each variable X_i, one adds directed edges from its direct causes or parents $\mathrm{PA}(i)$ into i.
- Assumptions that guarantee identifiability of the causal graph.
  - linear additive Gaussian noise models with equal noise variance (Peters and Buhlmann, 2014).
  - linear additive non-Gaussian noise model (Shimizu et al., 2006).
  - nonlinear additive noise models (Hoyer et al., 2008; Peters et al., 2014).
  - post-nonlinear additive noise models (Zhang and Hyvarinen, 2009).
  - partially-linear additive Gaussian noise models (Rothenhausler et al., 2018).
  - discrete models (Peters et al., 2011).

Structure learning methods

- **Score-based learning** 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 (DAG), the true causal DAG $\mathcal{G}$ satisfy

$$\mathcal{G} \in \underset{\tilde{\mathcal{G}}:\tilde{\mathcal{G}} \text{ is a DAG}}{\operatorname{argmin}} \ell(\tilde{\mathcal{G}}). \tag{1}$$

- **Constraint-based learning** test for conditional independences in P_X and use these results to infer the causal structure.
- After finding the skeleton first, give it a direction.
- studied for polytrees.

polytree in constraint-based learning

- Allow for multiple root as well as nodes with multiple parents.
- Chow and Liu (1968) showed how to recover an undirected Markov tree from a given discrete joint pdf using the maximum weight spanning tree (MWST) algorithm.
- Rebane and Pearl (1987) propose a method to identify the causal basins, structures constructed from nodes with at least two parents.

polytree in constraint-based learning (Cont.)

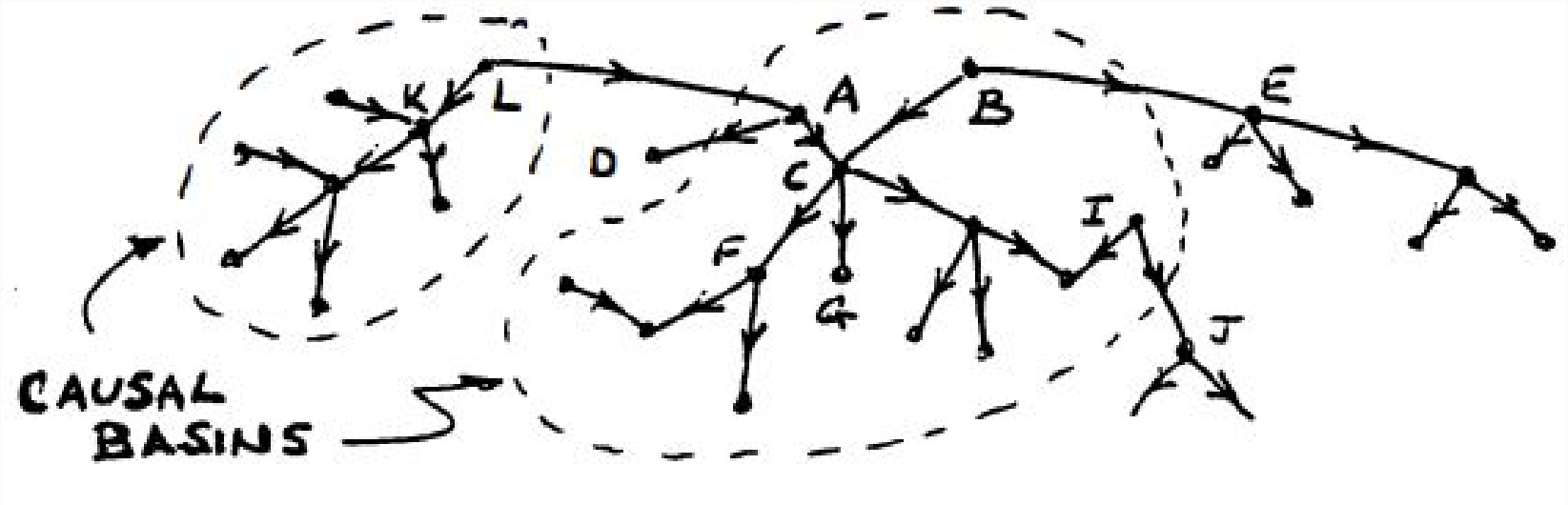

Figure 1: Example for basins Rebane and Pearl (1987).

2. Score-based Learning and Identifiability of Trees

graph terminology

- A directed graph $\mathcal{G} = (V, \mathcal{E})$ consists of $p \in \mathbb{N}_{>0}$ vertices (or nodes) $V = \{1, \ldots, p\}$ and a collection of directed edges $\mathcal{E} \subset \{(i \to j) \equiv (i, j) : i, j \in V, i \neq j\}$
- The unique node of a directed tree $\mathcal{G}$ with no parents is called the root node and is denoted by $\mathrm{rt}(\mathcal{G})$.
- 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

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

$$X_i := f_i(X_{\mathrm{pa}_{\mathcal{G}}(i)}) + N_i, \quad \forall 1 \leq i \leq p,$$

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.

- $\mathcal{M}$ denotes all measurable functions,
- $\mathcal{D}_k$ denotes the set of all $k \in \mathbb{N}$ times differentiable functions.
- $\mathcal{P}$ denotes the set of mean zero probability measures on $\mathbb{R}$.
- For any set $\mathcal{P}$ of probability measures, $\mathcal{P}^p$ denotes all p-dimensional product measures on $\mathbb{R}^p$ with marginals in $\mathcal{P}$.

2-1. Identifiability of Causal Additive Tree Models

- $\mathcal{C}_k$ denotes the k times continuously differentiable functions.
- $\mathcal{P}_+ \subset \mathcal{P}$ denotes the subset for which a density is strictly positive.
- For any class $\mathcal{F} \subseteq \{f | f : \mathbb{R} \to \mathbb{R}\}$, $\mathcal{P}_+ \subset \mathcal{P}$ denotes the subset with a density function in $\mathcal{F}$.
- As a special case, we let $\mathcal{P}_G \subset \mathcal{P}_{+\mathcal{C}_\infty} := \mathcal{P}_+ \cap \mathcal{P}_{\mathcal{C}_\infty}$ denotes the subset of Gaussian probability measures.

Definition (Restricted structural causal additive tree models)

The collection of restricted structural 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 $i \in \{1, \ldots, p\} \setminus \{\mathrm{rt}(\mathcal{G})\}$:

1. f_i is nowhere constant, i.e., it is not constant on any non-empty open set
2. 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

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

2-1. Identifiability of Causal Additive Tree Models

For example, consider causal additive tree model with Gaussian noise. Then, the second condition is that f_i is not linear.

To show this, assume that

- (1) of Definition 2 is satisfied (let f_i be nowhere constant)
- (2) of Definition 2 is not satisfied (f_i is linear function.)

The log density of N_i, for all $i \in \{1, \ldots, p\}$ is given by

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

And there exists an $i \in \{1, \ldots, p\} \setminus \{\mathrm{rt}(\mathcal{G})\}$ such that for all

$$(x, y) \coloneqq \{(x, y) \in \mathbb{R}^2 : \nu_i^{''}(y - f_i(x))f_i^{'}(x) \neq 0\}$$
$$\{(x, y) \in \mathbb{R}^2 : f_i^{'}(x) \neq 0\},$$

it holds that

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

Then, it hold that

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

i.e., $f_i^{''}(x)/f_i^{'}(x)$ is constant.

Thus, the differential equation holds that

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

by the division with $f_i^{'}(x)$. By integration this implies that
$0 = \xi^{''}(x)/f_i^{'}(x) - 2f_i^{'}(x)/\sigma^2 + c_3$ such that
$\xi^{''}(x) = 2\exp(2c_1x + 2c_2)/\sigma^2 - c_3\exp(c_1x + c_2)$ and
$\xi^{'}(x) = \exp(2c_1x + 2c_2)/c_1\sigma^2 - c_e\exp(c_1x + c_2)/c_1 + c_4$ 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 $\xi \to \infty \Longleftrightarrow p_{X_{\mathrm{pa}_{\mathcal{G}}(i)}} \to \infty$ as $x \to (c_1)\cdot\infty$. it is contradiction. ■

2-1. Identifiability of Causal Additive Tree Models

Existing indentifiability results for causal graphs in restricted SCM (Hoyer et al., 2008, Peters et al., 2014), for all $\theta \in \Theta_R$ and $\tilde{\theta} \in \tilde{\Theta}_R$ prove that

$$\mathcal{G} \neq \tilde{\mathcal{G}} \Rightarrow \mathcal{L}(X_\theta) \neq \mathcal{L}(X_{\tilde{\theta}}),$$

where $\mathcal{L}$ denotes the distribution of a random variable.

The proposition is extended to result that does not assume that $\tilde{\theta}$ is a restricted causal model as below.

Proposition (Identifiability of causal additive tree models)

Suppose that $\theta \in \Theta_R$ and $\tilde{\theta} \in (\mathcal{T}, \mathcal{D}_1^p, \mathcal{P}_{\mathcal{C}_0}^p)$ respectively, it holds that

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

2-2. Score functions

- For $\forall i \neq j$, assume that that $X_i - \mathbb{E}[X_i|X_j]$ has a density.
- For any graph $\tilde{\mathcal{G}} \in \mathcal{T}_p$, we define for each node $i \in V$, Gaussian score as

$$\ell_G(\tilde{\mathcal{G}}) = \sum_{i=1}^{p} \ell_G(\tilde{\mathcal{G}}, i),$$

where

$$\ell_G(\tilde{\mathcal{G}}, i) := \log\left(\mathrm{Var}\left(X_i - \mathbb{E}\left[X_i \mid X_{\mathrm{pa}^{\mathcal{G}}(i)}\right]\right)\right)/2,$$

- The Gaussian score is seen as

$$\ell_G(\tilde{\mathcal{G}}) = \inf_{Q \in \{\tilde{\mathcal{G}}\} \times \mathcal{D}_1^p \times \mathcal{P}_{\mathrm{G}}^p} h\left(P_X, Q\right) - p\log(\sqrt{2\pi e}), \tag{3}$$

where $Q \in \mathcal{T}_p \times \mathcal{M}^p \times \mathcal{P}^p$ and h is cross-entropy.

proof of (3)

Consider an SCM $\tilde{\theta} \in \{\tilde{\mathcal{G}}\} \times \mathcal{D}_1^p \times \mathcal{P}_{\mathrm{G}}^p$ and $Q_{\tilde{\theta}}$ is induced distribution.

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

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

As $\tilde{N} \sim \mathcal{N}(0, \tilde{\sigma}_i^2)$ for some $\tilde{\sigma}_i > 0$,

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

$$= \log(\sqrt{2\pi}\sigma_i) + \frac{\mathbb{E}\left(X_i - \tilde{f}(X_{\mathrm{PA}_{\tilde{\mathcal{G}}}(i)})\right)^2}{2\tilde{\sigma}_i^2}.$$

proof of (3) (Cont.)

We thus have

$$\inf_{\tilde{\sigma}_i>0}\left\{\log(\sqrt{2\pi}\sigma_i)+\frac{\mathbb{E}\left(X_i-\tilde{f}(X_{\mathrm{PA}^{\tilde{\mathcal{G}}}(i)})\right)^2}{2\tilde{\sigma}_i^2}\right\}$$
$$=\log(\sqrt{2\pi})+\frac{1}{2}\log\mathbb{E}\left(X_i-\tilde{f}(X_{\mathrm{PA}^{\tilde{\mathcal{G}}}(i)})\right)^2+\frac{1}{2}.$$

We conclude that

$$\inf_{Q\in\mathcal{T}_p\times\mathcal{M}^p\times\mathcal{P}^p} h(P_X,Q)=p\log(\sqrt{2\pi})+\frac{p}{2}+\sum_{i=1}^{p}\frac{1}{2}\log\left(\inf_{\tilde{f}_i\in\mathcal{D}_1}\mathbb{E}\left(X_i-\tilde{f}(X_{\mathrm{PA}^{\tilde{\mathcal{G}}}(i)})\right)^2\right).$$

Finally, we have that

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

2-2. Score functions

- To choosing the best scoring graph, we use the identifiability gap between $\tilde{\mathcal{G}}$ and causal graph $\mathcal{G}$ formed as

$$\underset{\mathcal{G}\in\mathcal{T}\setminus\{\mathcal{G}\}}{\operatorname{argmin}}\ \ell(\tilde{\mathcal{G}}) - \ell(\mathcal{G}). \tag{4}$$

- In Gaussian noise, the Gaussian score gap equals to

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

where D_{KL} denotes the Kullback-Leibler divergence. And it implies that

$$D_{\mathrm{KL}}(P_X||Q) > 0.$$

- From assumption the gap is strictly positive, the score functions can be used to identify the true causal graph of a restricted structural model:

$$\mathcal{G} = \underset{\tilde{\mathcal{G}}\in\mathcal{T}_p}{\operatorname{argmin}}\, \ell(\tilde{\mathcal{G}}). \tag{5}$$

3. Causal Additive Trees (CAT)

Chu-Liu-Edmonds' Algorithm

- Based on a maximum spanning tree algorithm

1. Build a complete graph with directed and weighted edges.
2. Keep only incoming edges with the maximum scores.
3. If there is no cycle, go to #5.
4. If there is a cycle, pretend vertices in the cycle as one vertex and update scores for all incoming edges to the cycle; goto #2.
5. Break all cycles by removing inappropriate edges in the cycle.

Figure 2: Chu-Liu-Edmonds' algorithm

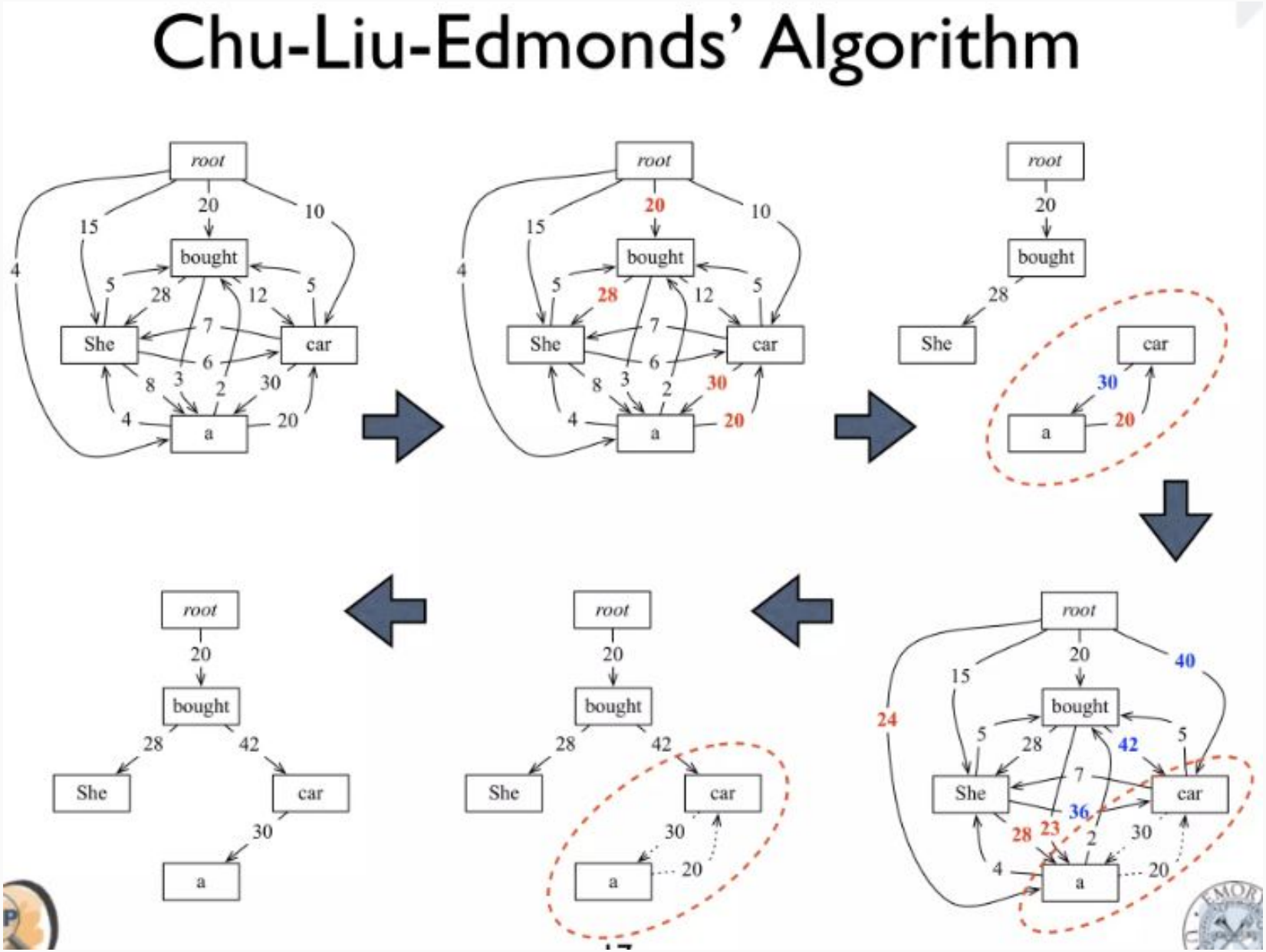

Figure 3: Chu-Liu-Edmonds' algorithm

3. Algorithm

- $\mathcal{T}_p$ denotes directed spannig trees that is recovered by Chu-Liu-Edmonds' algorithm.
- Minimizing the Gaussian score is equivalent to minimizing a translated version of the Gaussian score function:

$$\operatorname*{argmin}_{\tilde{\mathcal{G}} \in \mathcal{T}_p} \ell_G(\tilde{\mathcal{G}}) = \operatorname*{argmin}_{\tilde{\mathcal{G}} \in \mathcal{T}_p} \sum_{i=1}^{p} \frac{1}{2} \log(\mathrm{Var}(X_i - \mathbb{E}[X_i \mid X_{\mathrm{PA}_{\tilde{\mathcal{G}}}(i)}])) - \sum_{i=1}^{p} \frac{1}{2} \log\left(\mathrm{Var}\left(X_i\right)\right) \tag{6}$$

$$= \operatorname*{argmin}_{\tilde{\mathcal{G}} \in \mathcal{T}_p} \sum_{i=1}^{p} \frac{1}{2} \log\left(\frac{\mathrm{Var}(X_i - \mathbb{E}[X_i \mid X_{\mathrm{PA}_{\tilde{\mathcal{G}}}(i)}])}{\mathrm{Var}(X_i)}\right). \tag{7}$$

- Then, from (7), the Gaussian edge weights $w^G := (w^G_{ji})_{j \neq i}$

$$w^G_{ji} := \frac{1}{2} \log\left(\frac{\mathrm{Var}(X_i - \mathbb{E}[X_i \mid X_j])}{\mathrm{Var}(X_i)}\right). \tag{8}$$

- Solving the problem (5) is equal to

$$\mathcal{G} = \operatorname*{argmin}_{\tilde{\mathcal{G}} = (V, \tilde{\mathcal{E}}) \in \mathcal{T}_p} \sum_{(j \to i) \in \tilde{\mathcal{E}}} w^G_{ji}. \tag{9}$$

3. Algorithm

- Given the observed nodes $X = (X_1, \ldots, X_p)$, we estimate the edge weights by simple plug-in estimators.
- Let us denote the conditional expectation function and its estimated by

$$\varphi_{ji}(x) := \mathbb{E}[X_i | X_j = x], \quad \hat{\varphi}_{ji}(x) := \hat{\mathbb{E}}[X_i | X_j = x] \tag{10}$$

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

$$\hat{w}^G_{ji} := \frac{1}{2} \log \left(\frac{\widehat{\mathrm{Var}}(X_i - \hat{\varphi}_{ji}(X_j))}{\widehat{\mathrm{Var}}(X_i)} \right), \tag{11}$$

  for all $j \neq i$, where $\widehat{\mathrm{Var}}$ denotes a variance estimator.

Algorithm Causal additive trees (CAT)

1: **procedure** CAT (X_n, regression method)
2: Run regression method to obtain $\hat{\varphi}_{ji}$ for all $j \neq i$.
3: Compute empirical edge weights $\hat{w}^G$, see Equation (11).
4: Apply Chu-Liu-Edmonds' algorithm to find MWDST with respect to $\hat{w}^G$.
5: **return** MWDST $\hat{\mathcal{G}}$.
6: **end procedure**

Theorem (Pointwise consistency)

Suppose that for all $j \neq i$ the following two conditions hold:
(a) if $(j \to i) \in \mathcal{E}, \mathbb{E}\left[\left(\hat{\varphi}_{ji}(X_j) - \varphi_{ji}(X_j)\right)^2 \mid \tilde{\mathbf{X}}_n\right] \xrightarrow{P}_n 0$.
(b) if $(j \to i) \notin \mathcal{E}, \mathbb{E}\left[\left(\hat{\varphi}_{ji}(X_j) - \tilde{\varphi}_{ji}(X_j)\right)^2 \mid \tilde{\mathbf{X}}_n\right] \xrightarrow{P}_n 0$ 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).

4. Hypothesis Testing

4-1. Confidence Region for the Causal Tree

- The joint distribution of the estimated edge weights should be asymptotically Gaussian:
  - For all $k \in \{1, \dots, n\}$, let the vectors of squared residuals be given by
$$\hat{M}_k := \{(X_{k,i} - \hat{\varphi}(X_{k,j}))^2\}_{j \neq i} \in \mathbb{R}^{p(p-1)}$$
and let squared centered observations be given by
$$\hat{V}_k = \left\{ \left(X_{k,i} - \frac{1}{n} \sum_{m=1}^{n} X_{m,i} \right)^2 \right\}_{1 \leq i \leq p} \in \mathbb{R}^p.$$
  - Further let
$$\hat{\mu} := \frac{1}{n} \sum_{k=1}^{n} \hat{M}_k, \quad \hat{\nu} := \frac{1}{n} \sum_{k=1}^{n} \hat{V}_k.$$
  - The empirical Gaussian edge weight for $j \to i$ is given by $\log(\hat{\mu}_{ji}/\hat{\nu}_i)/2$.
  - Let $\hat{\Sigma}_M \in \mathbb{R}^{p(p-1) \cdot p(p-1)}$ denote the empirical variances of $\hat{M}_k$.
  - Let $\hat{\Sigma}_V \in \mathbb{R}^{p \cdot p}$ denote the empirical variances of $\hat{V}_k$.
  - Let $\hat{\Sigma}_{MV} \in \mathbb{R}^{p(p-1) \cdot p}$ denote the empirical covariance of $\hat{M}_k$ and $\hat{M}_k$.
  - $$\hat{\Sigma} := \begin{pmatrix} \hat{\Sigma}_M & \hat{\Sigma}_{MV} \\ \hat{\Sigma}_{MV} & \hat{\Sigma}_V \end{pmatrix} := \begin{pmatrix} \hat{M}_k \hat{M}_k^T - \hat{\mu}\hat{\mu}^T & \hat{M}_k \hat{V}_k^T - \hat{\mu}\hat{\nu}^T \\ \hat{V}_k \hat{M}_k^T - \hat{\nu}\hat{\mu}^T & \hat{V}_k \hat{V}_k^T - \hat{\nu}\hat{\nu}^T \end{pmatrix}$$

4-1. Confidence Region for the Causal Tree

- Set the confidence interval

$$\hat{u}_{ji}, \hat{l}_{ji} := \frac{1}{2} \log \left(\frac{\hat{\mu}_{ji}}{\hat{\nu}_i} \right) \pm z_\alpha \frac{\hat{\sigma}_{ji}}{2\sqrt{n}} = \hat{w}_{ji}^G \pm z_\alpha \frac{\hat{\sigma}_{ji}}{2\sqrt{n}}, \tag{12}$$

  where $\hat{\sigma}_{ji}^2 := \frac{\hat{\Sigma}_{M,ji,ji}}{\hat{\mu}_{ji}^2} + \frac{\hat{\Sigma}_{V,i,i}}{\hat{\nu}_i^2} - \frac{\hat{\Sigma}_{MV,ji,i}}{\hat{\mu}_{ji}\hat{\nu}_i}$.
  By Bonferroni correction, z_α denotes the upper $\alpha/\{2p(p-1)\}$ quantile of a standard normal distribution.

- The region of directed trees formed of minimizers of the score with edge weights in the confidence hyper-rectangle:

$$\hat{C}_{\text{Bon}} := \hat{C}(\hat{l}, \hat{u}) := \left\{ \underset{\tilde{\mathcal{G}}=(V,\tilde{\mathcal{E}}) \in \tilde{T}_p}{\text{argmin}} \sum_{(j \longrightarrow i)} w_{ji}^{'} : \forall j \neq i, w_{ji}^{'} \in [\hat{l}_{ji}, \hat{u}_{ji}] \right\}.$$

Theorem (Confidence region)

Suppose the following conditions hold:

1. here exists $\xi > 0$ such that $\mathbb{E}|X|^{4+\xi} < \infty$
2. there exists $\xi > 0$ such that for all $j \neq i, \mathbb{E}[|\hat{\varphi}_{ji}(X_j) - \varphi_{ji}(X_j)|^{4+\xi} \mid \tilde{\mathbf{X}}_n] = O_p(1)$
3. $\text{Var}\left(\left(\hat{M}_1^T, \hat{V}_1^T\right)^T \mid \tilde{\mathbf{X}}_n\right) \xrightarrow{P} {}_n\Sigma$, where Σ is constant with strictly positive diagonal
4. for $(j \to i) \in \mathcal{E}, \sqrt{n}\mathbb{E}\left[\left(\hat{\varphi}_{ji}(X_{k,j}) - \varphi_{ji}(X_{k,j})\right)^2 \mid \tilde{\mathbf{X}}_n\right] \xrightarrow{P} {}_n0$

Then

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

But, the confidence region $\hat{C}_{\text{Bon}}$ will not be possible to compute it in practice.. (due to the ranges of $w_{ji}^{'}$)

A substructure restriction $\mathcal{R} = (\mathcal{E}_{\mathcal{R}}, \mathcal{E}_{\mathcal{R}}^{\text{miss}}, r)$ exists on the node V, where

- $\mathcal{E}_{\mathcal{R}}$ denotes existing edges
- $\mathcal{E}_{\mathcal{R}}^{\text{miss}}$ denotes missing edges
- r denotes a specific root node r

The null hypothesis for testing substructure:

$$\mathcal{H}_0(\mathcal{R}) : \mathcal{E}_{\mathcal{R}} \setminus \mathcal{E} = \emptyset, \mathcal{E} \setminus \mathcal{E}_{\mathcal{R}}^{\text{miss}} = \mathcal{E}, r = \text{rt}(\mathcal{G}), \tag{13}$$

where $\mathcal{G} = (V, \mathcal{E})$ is the true graph.

4-2. Testing of Substructures

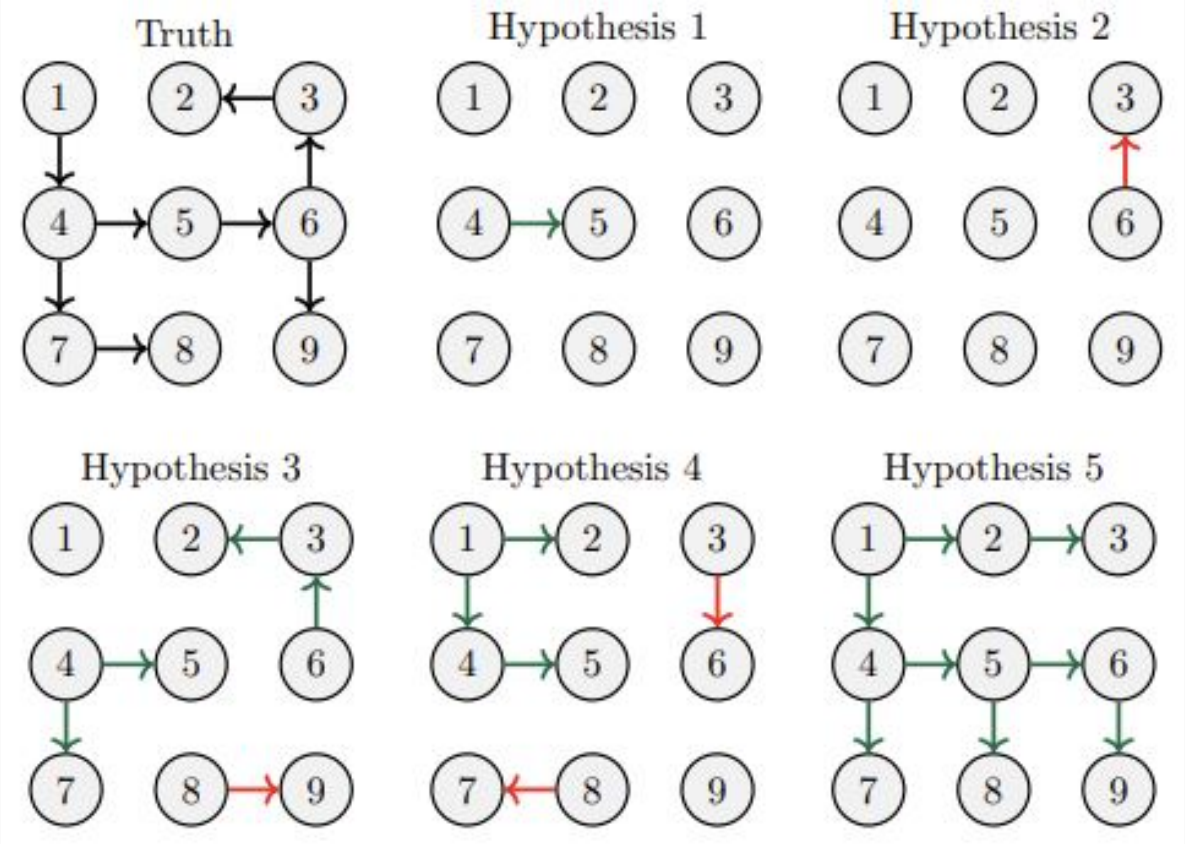

Figure 4: Illustration for hypothesis testing (green is $\mathcal{E}_{\mathcal{R}}$, red is $\mathcal{E}_{\mathcal{R}}^{\text{miss}}$).

- Hypothesis 1 consists of the restriction $\mathcal{R} = \mathcal{E}_{\mathcal{R}}$, where $\mathcal{E}_{\mathcal{R}} := \{(X_4 \to X_5)\}$. (*true*)
- Hypothesis 2 consists of the restriction $\mathcal{R} = \mathcal{E}_{\mathcal{R}}^{\text{miss}}$, where $\mathcal{E}_{\mathcal{R}}^{\text{miss}} := \{(X_6 \to X_3)\}$. (*false*)

1. CheckC (Check confidence interval)

- Let $\mathcal{T}_p(\mathcal{R}) \subset \mathcal{T}_p$ be the set of all directed trees satisfying the substructure restriction $\mathcal{R}$.
- Suppose that the causal directed tree $\mathcal{G}$ satisfies $\mathcal{R}$, i.e., $\mathcal{G} \in \mathcal{T}_p(\mathcal{R})$.
- From Chu-Liu-Edmond's algorithm, there exists a graph in $w^{'} = (w^{'}_{ji})_{j \neq i}$, with $\hat{l}_{ji} \leq w^{'} \leq \hat{u}_{ji}$ for all $j \neq i$.
- Find $S_{\mathcal{T}_p(\mathcal{R})}(\hat{l})$ and $\mathcal{G}^*_{\mathcal{T}_p(\mathcal{R})}(\hat{l})$. Hence, it must hold that

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

- Define CheckC test function as

$$\psi^{\text{CheckC}}_{\mathcal{R}} := \begin{cases} 0, & \text{if } S_{\mathcal{T}_p(\mathcal{R})}(\hat{l}) \leq S_{\mathcal{T}_p(\mathcal{R})}(\hat{u}), \\ 1, & \text{otherwise.} \end{cases} \tag{14}$$

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

procedure CheckC $\left(\mathcal{R} = (\mathcal{E}_{\mathcal{R}}, \mathcal{E}_{\mathcal{R}}^{\text{miss}}, r)\right., \hat{l} = \left(\hat{l}_{ji}\right)_{j \neq i}, \hat{u} = (\hat{u}_{ji})_{j \neq i})$.

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 $\{(\mathrm{k} \to i) : k \in V \setminus \{j\}\} \cup \{i \to j\}$.

For each $(j \to i) \in \mathcal{E}_{\mathcal{R}^{\text{miss}}} from \mathrm{H} the edges \{(\mathrm{j} \to i)$.

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 $\mathcal{H}$.

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

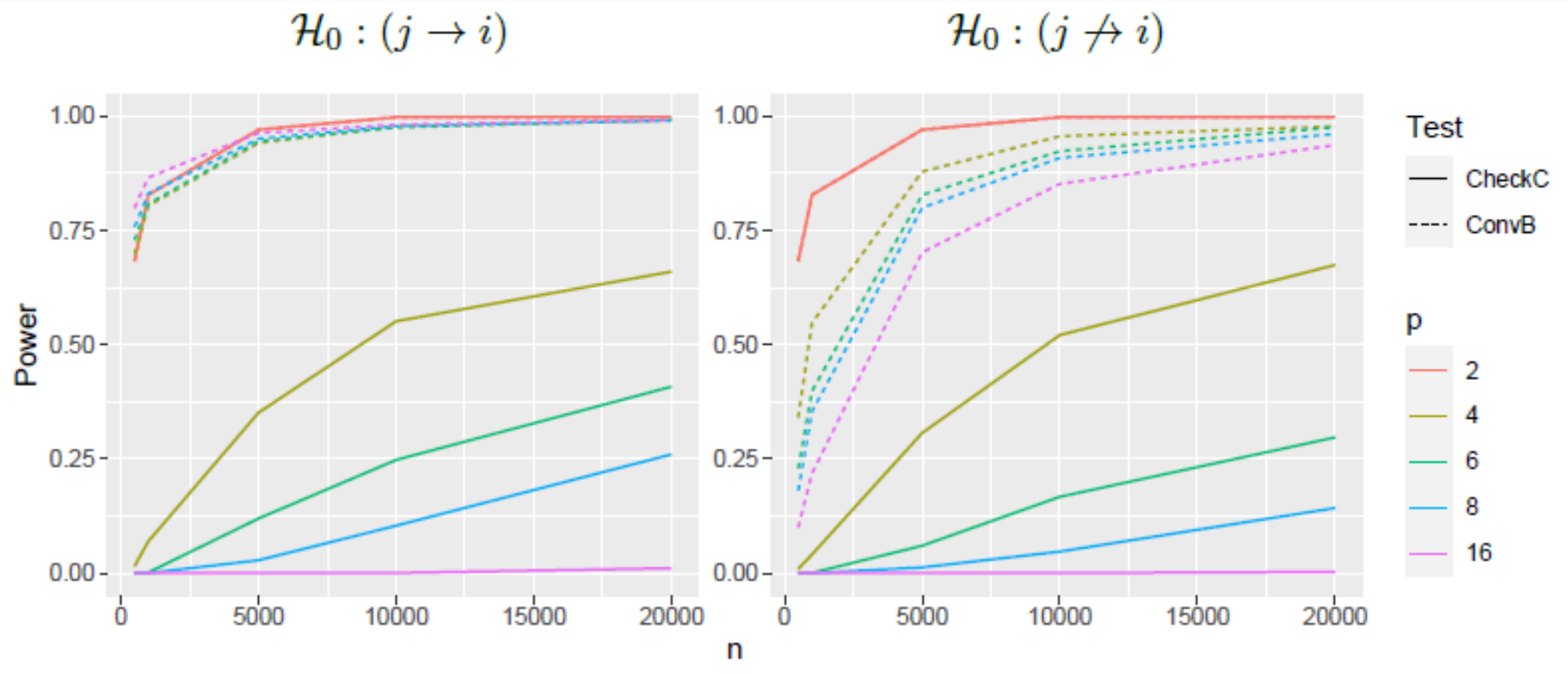

Figure 5: The power of the proposed testing procedure for simple hypotheses.

In simulation results, the power using the CheckC is lower than the alternative thing.

4-3. Converging Bounds

- 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 $\mathcal{G} \in \hat{\mathcal{C}}_{\text{Bon}}$, which implies exists a graph in $w^{'} = (w^{'}_{ji})_{j \neq i}$, with $\hat{l}_{ji} \leq w^{'} \leq \hat{u}_{ji}$ for all $j \neq i$.
- Define the edge weights as

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

- Define ConvB test function as

$$\psi^{\text{ConvB}}_{\mathcal{R}} = \begin{cases} 0, & \text{if } \mathcal{G}^{*}_{\mathcal{T}_p}(\check{w}) \text{ satisfies } \mathcal{R} \\ 1, & \text{otherwise.} \end{cases} \tag{15}$$

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

procedure CoNVB $\left(\mathcal{R} = \left(\mathcal{E}_{\mathcal{R}}, \mathcal{E}_{\mathcal{R}}^{\text{miss}}, r\right), \hat{l} = \left(\hat{l}_{ji}\right)_{j \neq i}, \hat{u} = \left(\hat{u}_{ji}\right)_{j \neq i}\right)$

Initialize $\check{w} := \hat{l}$.

For each $(j \to i) \in \mathcal{E}_{\mathcal{R}}$ and all $k \in V \backslash \{j\}$, set $\check{w}_{ki} := \hat{u}_{ki}$.

For each $(j \to i) \in \mathcal{E}_{\text{Riss}}$, 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 $\psi_{\mathcal{R}}^{\text{ConvB}} := 0$, otherwise set $\psi_{\mathcal{R}}^{\text{ConvB}} := 1$.

return $\psi_{\mathcal{R}}^{\text{ConvB}}$.

end procedure

5. Simulation

5. Simulation

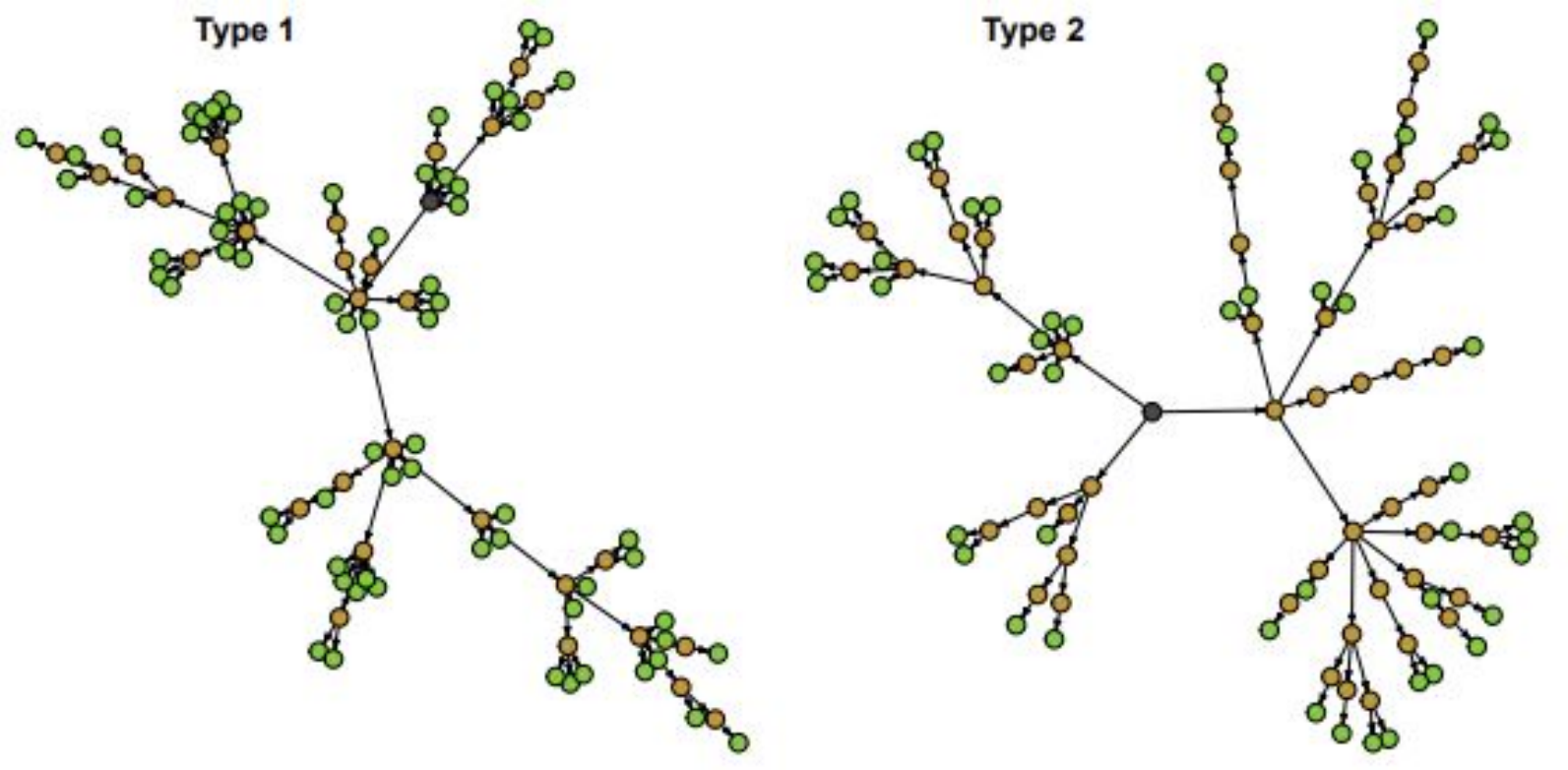

Figure 6: Illustration of Type 1 (many leaf nodes) and Type 2 (many branch nodes) directed trees over $p = 100$ nodes.

- green (leaf node), brown (branch node), black (root node)
- The Type 1 tree contains 70 leaf nodes, while the Type 2 tree only contains 49 leaf nodes.

5. Simulation

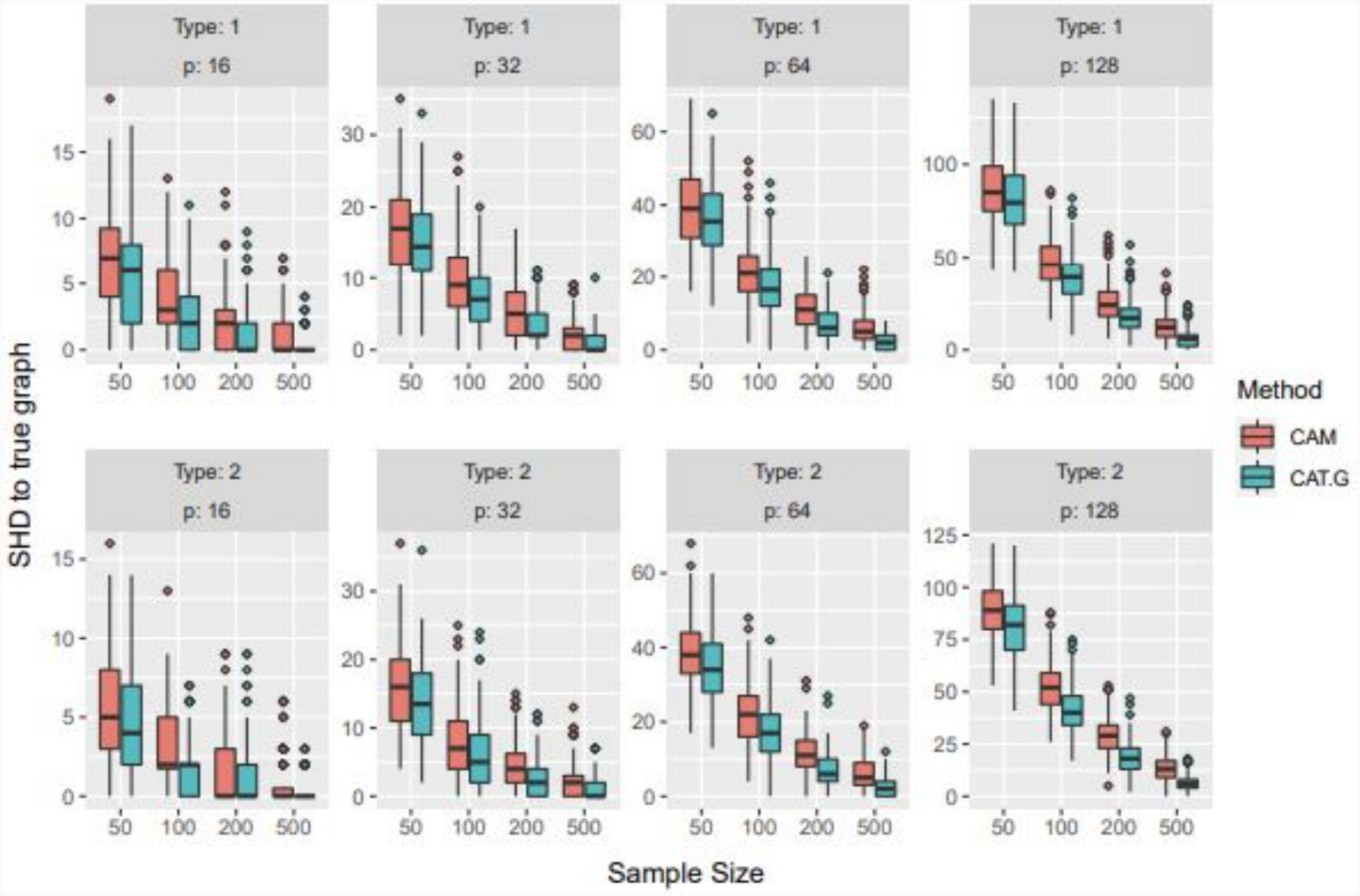

Figure 7: Causal additive tree models with Gaussian noise

Boxplots of the SHD (Structural Hamming Distance) performance of CAM (Buhlmann et al., 2014) and CAT.G