Differentiable Causal Discovery for Latent Hierarchical Causal Models

Parjanya Prajakta Prashant, Ignavier Ng, Kun Zhang, Biwei Huang

ICLR 2025

Overview

Latent Hierarchical Causal Models

Latent Hierarchical Causal Models

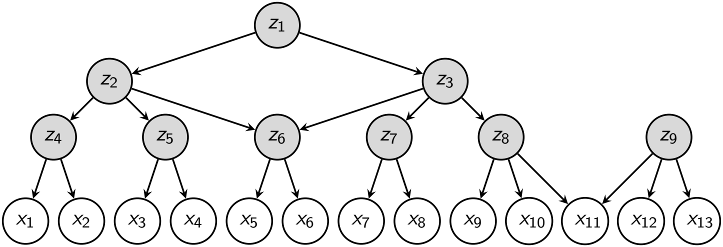


Figure: Example of a Latent Hierarchical Causal Model

Latent Hierarchical Causal Models

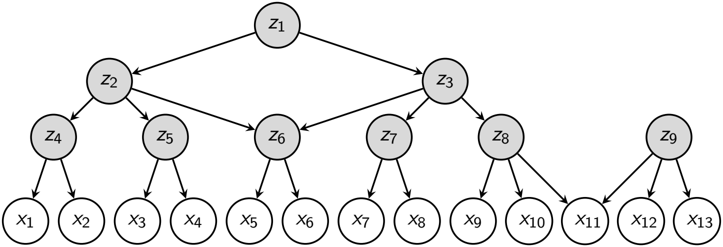


Figure: Example of a Latent Hierarchical Causal Model

Condition 1 (Structural Conditions)

1. Each latent variable has at least two pure children.
2. For any latent variable $z_i \in \mathbb{Z}$, let $\mathcal{D}_i = \text{De}(z_i) \cap \mathbb{X}$ be the set of measured descendants of z_i where $\text{De}(\cdot)$ denotes the descendants. Then, for all $x_j, x_k \in \mathcal{D}_i$, $d(z_i, x_j) = d(z_i, x_k)$.

Identifiability: Intuition

Identifiability: Intuition

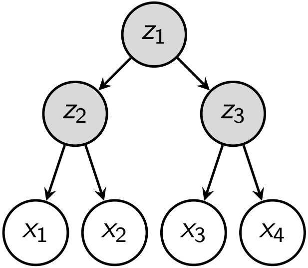


Figure: Latent hierarchical model

Identifiability: Intuition

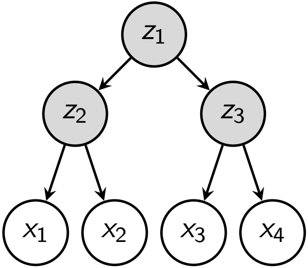


Figure: Latent hierarchical model

- $\{x_1, x_2\} \perp \{x_3, x_4\} \mid z_1$

Identifiability: Intuition

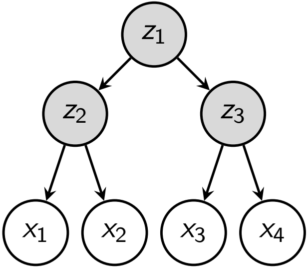


Figure: Latent hierarchical model

- $\{x_1, x_2\} \perp \{x_3, x_4\} \mid z_1$
- $P(x_1, x_2 \mid x_3, x_4) = \int P(x_1, x_2 \mid z_1) P(z_1 \mid x_3, x_4) dz_1$

Identifiability: Intuition

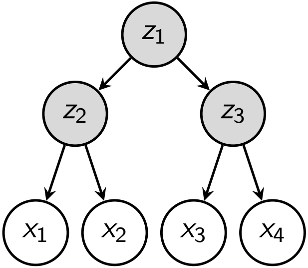


Figure: Latent hierarchical model

- $\{x_1, x_2\} \perp \{x_3, x_4\} \mid z_1$
- $P(x_1, x_2 \mid x_3, x_4) = \int P(x_1, x_2 \mid z_1) P(z_1 \mid x_3, x_4) dz_1$
- This places a constraint on the measured distribution $P(x_1, x_2, x_3, x_4)$

Identifiability: Intuition

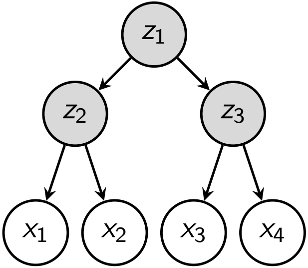


Figure: Latent hierarchical model

- $\{x_1, x_2\} \perp \{x_3, x_4\} \mid z_1$
- $P(x_1, x_2 \mid x_3, x_4) = \int P(x_1, x_2 \mid z_1) P(z_1 \mid x_3, x_4) dz_1$
- This places a constraint on the measured distribution $P(x_1, x_2, x_3, x_4)$
- Size of d-separating set $=$ minimum dimension of z s.t. $P(x_1, x_2 \mid x_3, x_4) = \int P(x_1, x_2 \mid z) P(z \mid x_3, x_4) dz$

Identifiability: Conditions

Identifiability: Conditions

Condition 2 (Generalized Faithfulness)

A probability distribution P is faithful to a DAG $\mathcal{G}$ if every rank Jacobian constraint on a pair of set of measured variables that holds in P is entailed by every structural equation model with respect to $\mathcal{G}$.

Identifiability: Conditions

Condition 2 (Generalized Faithfulness)

A probability distribution P is faithful to a DAG $\mathcal{G}$ if every rank Jacobian constraint on a pair of set of measured variables that holds in P is entailed by every structural equation model with respect to $\mathcal{G}$.

Condition 3 (Differentiability)

1. For every pair of measured sets $\mathbb{X}$ and $\mathbb{Y}$, the function $f : \mathbb{R}^{|\mathbb{X}|} \to \mathbb{R}^{|\mathbb{Y}|}$ defined as $f(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{y}|\boldsymbol{x}]$ is continuously differentiable.
2. For every pair of measured set $\mathbb{X}$ and latent set $\mathbb{Z}$, there exists a continuous differentiable function $g : \mathbb{R}^{|\mathbb{X}|} \to \mathbb{R}^{|\mathbb{Z}|}$ such that $p(\boldsymbol{z}|\boldsymbol{x}) = p(\boldsymbol{z}|g(\boldsymbol{x}))$.

Identifiability: Jacobian Indicator

Identifiability: Jacobian Indicator

Theorem 1

Let the causal model $\mathcal{G}$ satisfy Conditions 1-3. For any two sets of measured variables $\mathbb{X}$ and $\mathbb{Y}$ in $\mathcal{G}$, let $f(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{y}|\boldsymbol{x}]$. For any $r < |\mathbb{X}|, |\mathbb{Y}|$, the rank of the Jacobian matrix $\mathbf{J}_f = \frac{\partial f}{\partial \boldsymbol{x}} = r$ if and only if the size of the smallest set of latent variables that d-separates $\mathbb{X}$ from $\mathbb{Y}$ is r. Formally,

$$rank(\mathbf{J}_f) = \min_{\mathbb{Z}} |\mathbb{Z}| \quad \text{such that} \quad \mathbb{X} \perp\!\!\!\perp_{\mathcal{G}} \mathbb{Y} | \mathbb{Z} \tag{1}$$

Identifiability: Structural Lemmas

Identifiability: Structural Lemmas

Lemma 1: Pure Children

A set of measured variables $\mathbb{S}$ are pure children of the same parent if and only if for any subset $\mathbb{T} \subseteq \mathbb{S}$, $r(\mathbb{T}, \mathbb{X} \setminus \mathbb{T}) = 1$.

Identifiability: Structural Lemmas

Lemma 1: Pure Children

A set of measured variables $\mathbb{S}$ are pure children of the same parent if and only if for any subset $\mathbb{T} \subseteq \mathbb{S}$, $r(\mathbb{T}, \mathbb{X} \setminus \mathbb{T}) = 1$.

Lemma 2: Non-Pure Children

c is a child of exactly the variables in $\mathbb{P}$ if and only if:

1. For each $\mathbb{S} \subseteq \mathbb{X}$ such that $|\mathbb{S} \cap \mathrm{Ch}(z_i)| = 1$ for each $z_i \in \mathbb{P}$:
$$r(\mathbb{S}, \mathbb{X} \setminus (\mathbb{S} \cup \{c\})) = r(\mathbb{S} \cup \{c\}, \mathbb{X} \setminus (\mathbb{S} \cup \{c\}))$$
2. The equality in condition (1) does not hold for any proper subset of $\mathbb{P}$.

Identifiability: Structural Lemmas

Lemma 1: Pure Children

A set of measured variables $\mathbb{S}$ are pure children of the same parent if and only if for any subset $\mathbb{T} \subseteq \mathbb{S}$, $r(\mathbb{T}, \mathbb{X} \setminus \mathbb{T}) = 1$.

Lemma 2: Non-Pure Children

c is a child of exactly the variables in $\mathbb{P}$ if and only if:

1. For each $\mathbb{S} \subseteq \mathbb{X}$ such that $|\mathbb{S} \cap \mathrm{Ch}(z_i)| = 1$ for each $z_i \in \mathbb{P}$:

$$r(\mathbb{S}, \mathbb{X} \setminus (\mathbb{S} \cup \{c\})) = r(\mathbb{S} \cup \{c\}, \mathbb{X} \setminus (\mathbb{S} \cup \{c\}))$$

2. The equality in condition (1) does not hold for any proper subset of $\mathbb{P}$.

Lemma 3: Independent Child

A measured variable c has no parent if and only if $r(\{c\}, \mathbb{X} \setminus \{c\}) = 0$.

Differentiable Causal Discovery Approach: Matching Distributions

Differentiable Causal Discovery Approach: Matching Distributions

$$z_j^i = f_j^i(\boldsymbol{M}^{i+1} \odot \boldsymbol{z}^{i+1}, \varepsilon_{z_j^i}), \tag{2}$$

$$x_j = g_j(\boldsymbol{M}^1 \odot \boldsymbol{z}^1, \varepsilon_{x_j}). \tag{3}$$

Differentiable Causal Discovery Approach: Matching Distributions

$$z_j^i = f_j^i(\boldsymbol{M}^{i+1} \odot \boldsymbol{z}^{i+1}, \varepsilon_{z_j^i}), \quad (2)$$

$$x_j = g_j(\boldsymbol{M}^1 \odot \boldsymbol{z}^1, \varepsilon_{x_j}). \quad (3)$$

Variational Approach

Maximize the evidence lower bound (ELBO):

$$\log p(\boldsymbol{x}; \theta, \boldsymbol{M}) \geq -\mathrm{KL}(q(\boldsymbol{\epsilon}|\boldsymbol{x})||p(\boldsymbol{\epsilon}; \theta)) + \mathbb{E}_q[\log p(\boldsymbol{x}|\boldsymbol{\epsilon}; \theta, \boldsymbol{M})] \quad (4)$$

- Encoder: Models approximate posterior $q(\boldsymbol{\epsilon}|\boldsymbol{x})$
- Decoder: Models conditional likelihood $p(\boldsymbol{x}|\boldsymbol{\epsilon}; \theta, \boldsymbol{M})$ according to SEM
- $\boldsymbol{\epsilon}$ represents all noise terms combined

Differentiable Causal Discovery Approach: Enforcing Structural Constraints

Differentiable Causal Discovery Approach: Enforcing Structural Constraints

Pure Children Constraint

$$\left\| \boldsymbol{M}_{i,:} \odot \prod_{j \neq i} (1 - \boldsymbol{M}_{j,:}) \right\|_1 \geq 2 \quad \forall i. \tag{5}$$

Differentiable Causal Discovery Approach: Enforcing Structural Constraints

Pure Children Constraint

$$\left\| \boldsymbol{M}_{i,:} \odot \prod_{j \neq i} (1 - \boldsymbol{M}_{j,:}) \right\|_1 \geq 2 \quad \forall i. \tag{5}$$

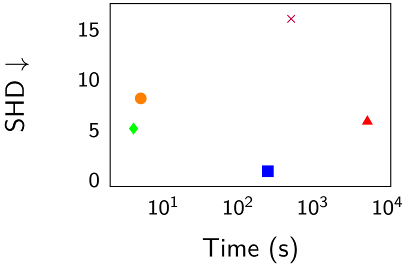
Final Objective

$$\mathcal{L}_{\text{final}} = -\mathbb{E}_{\boldsymbol{M} \sim \sigma(\gamma)} [\text{ELBO}(\theta, \boldsymbol{M})]) + \lambda_1 \mathcal{L}_{\text{ind}}(\epsilon) + \lambda_2 \|\sigma(\gamma)\|_1 \tag{6}$$

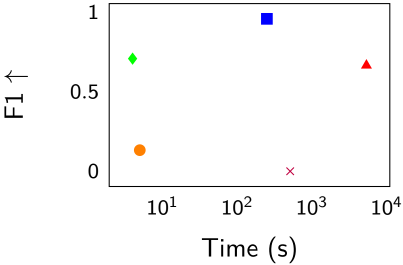
$$+ \lambda_3 \Big(\sum_i \max(0, \|\boldsymbol{M}_{i,:}\|_1 (2 - \|\boldsymbol{M}_{i,:} \odot \prod_{j \neq i} (1 - \boldsymbol{M}_{j,:})\|_1)) \Big)^2. \tag{7}$$

Experiments: Causal Discovery

Experiments: Causal Discovery



(a) Structural Hamming Distance (SHD) vs. Time



(b) F1 Score vs. Time

Figure: Performance vs. Time. Methods compared: Ours (■), KONG (▲), HUANG (♦), GIN (●), DeCAMFounder (×).

Experiments: Images

Experiments: Images

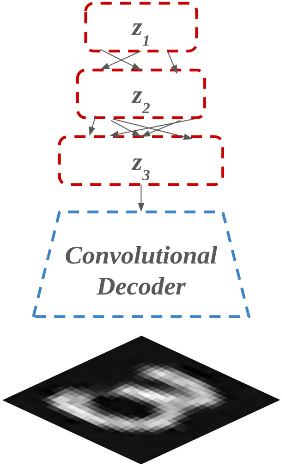


Figure: Architecture

Experiments: Images

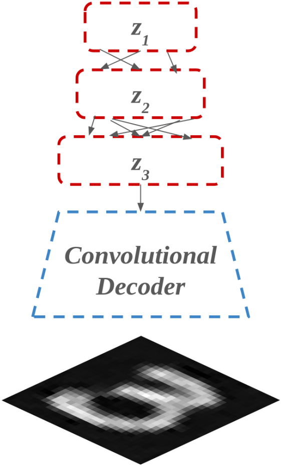


Figure: Architecture

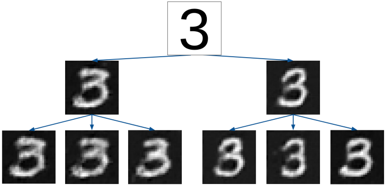


Figure: Discovered latent structure

Conclusion

- We establish identifiability of latent hierarchical causal models in the general case.
- We formulate a scalable differentiable approach.
- Extensive experiments validate the performance and scalability of our approach.