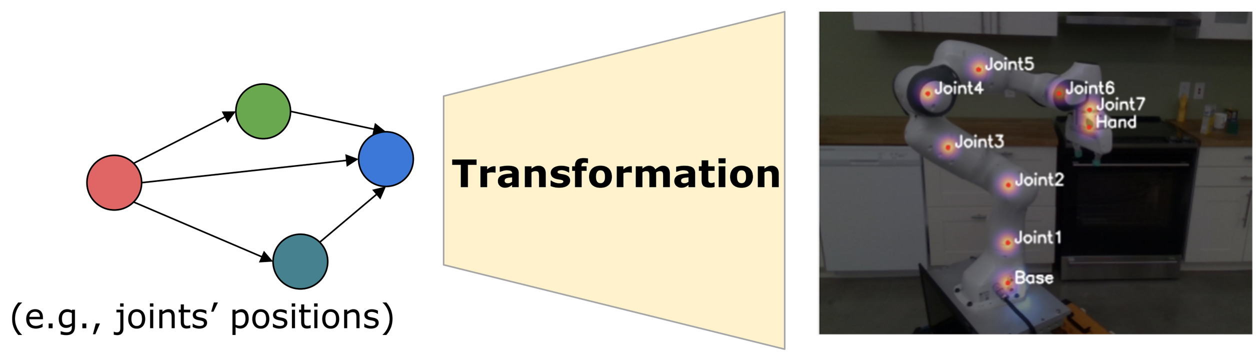
General Identifiability and Achievability for Causal Representation Learning

Burak Varıcı[1] Emre Acartürk[1] Karthikeyan Shanmugam[2] Ali Tajer[1]

[1]Rensselaer Polytechnic Institute [2]Google Research India

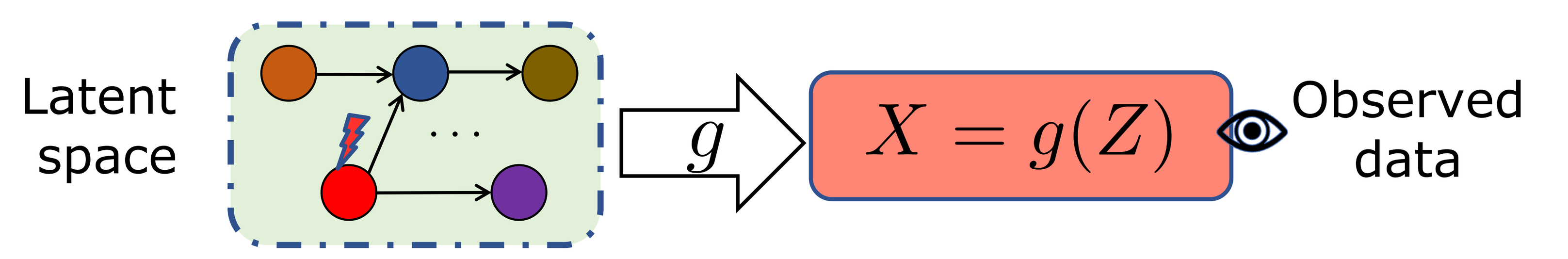
CRL from Interventions



".. learn a representation (partially) exposing the unknown causal structure, e.g., which variables describe the system, and their relations .." Schölkopf et al., 2021

Generic goal: Invert the unknown transformation to recover

1) latent representation and **2) the latent causal structure**



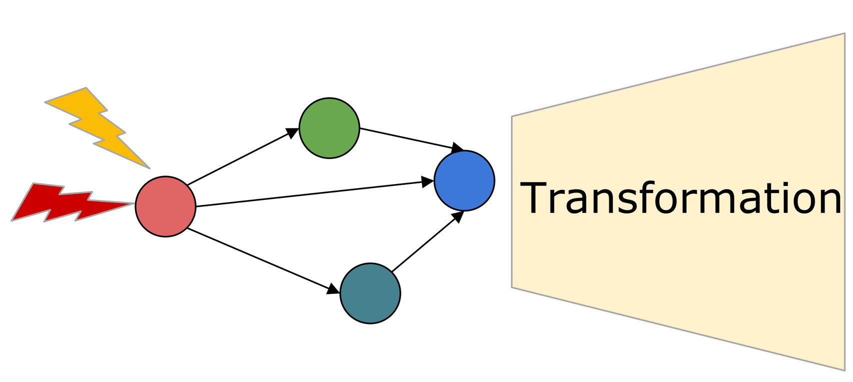
1. **Identifiability**: Conditions for uniquely recovering Z and $\mathcal{G}_Z$
2. **Achievability**: Provably correct algorithms to recover Z and $\mathcal{G}_Z$

Our contributions

Related work for perfect ID	Transform	Requirements	Provable Algorithm
Varıcı et al. 2024	Linear	1 int/node	✓
von Kügelgen et al. 2023	General	2 (coupled) int/node + faithfulness	✗
This work	**General**	**2 (uncoupled) int/node**	✓

Algorithm Overview

Sufficient Interventional Diversity**:**
2 different hard interventions per node in the latent space

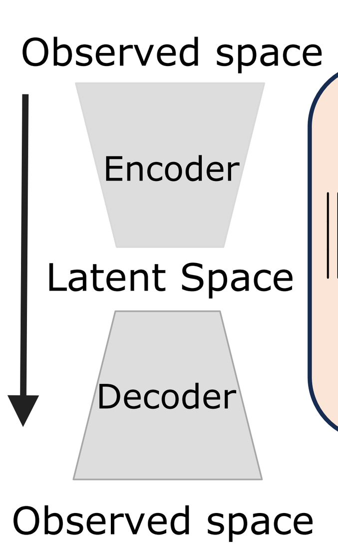


Intervention 1 → $\nabla \log p^{\text{Int}_1}(x) = s^1(x)$

Intervention 2 → $\nabla \log p^{\text{Int}_2}(x) = s^2(x)$

⋮

$\nabla \log p^{\text{Int}_{2n}}(x) = s^{2n}(x)$



Minimize

$$\left\| \mathbb{E}\big[|\text{Jac}_{\text{dec}}(x)(s^1(x) - s^2(x))|\big] \ \cdots \ \mathbb{E}\big[|\text{Jac}_{\text{dec}}(x)(s^{2n-1}(x) - s^{2n}(x))|\big] \right\|_0$$

\+ Reconstruction Loss (ensures injectivity)

Provably correct algorithm for unsupervised learning (small variations for different settings)

Experiments

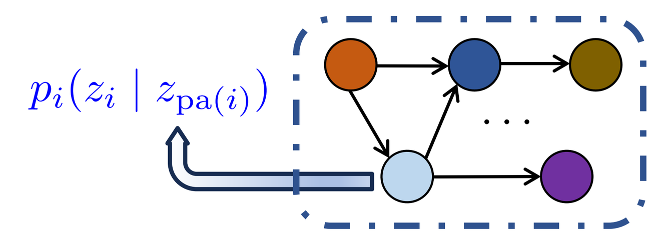
Non-linear latent model: $Z_i = \sqrt{Z_{\text{pa}(i)}^\top A_{p,i} Z_{\text{pa}(i)}} + N_{p,i}$ n=8 latent variables

Input score differences $(s_X - s_X^m)$: Perfect score oracle or Sliced Score Matching

Non-linear transform: $X = \tanh(T.Z)$

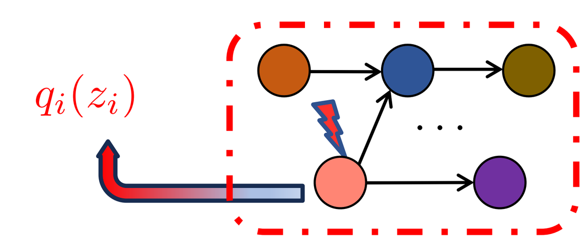
Obs. dim	Norm. Z error	DAG error (SHD)	Norm. Z error	DAG error (SHD)
8	0.16	1.56	0.70	11.9
25	0.20	1.55	0.68	10.5
40	0.21	1.14	0.71	11.8
	score oracle		noisy scores	

Why score functions?



$p(z) = p_i(z_i \mid z_{\text{pa}(i)}) \prod_{j \neq i} p_j(z_j \mid z_{\text{pa}(j)})$

$s(z) \triangleq \nabla_z \log p(z)$



$p^m(z) = q_i(z_i) \prod_{j \neq i} p_j(z_j \mid z_{\text{pa}(j)})$

$s^m(z) \triangleq \nabla_z \log p^m(z)$

$$s(z) - s^m(z) = \nabla_z \log p_i(z_i \mid z_{\text{pa}(i)}) - \nabla_z \log q_i(z_i)$$

Score functions contain all information about latent DAGs

node i intervened: $s(z) - s^m(z)$ becomes a function of only $z_{\overline{\text{pa}}(i)}$

$$s(z) - s^m(z) = \begin{bmatrix} 0 & 0 & \boxed{\times} & 0 & \boxed{\times} & 0 \end{bmatrix}^\top$$

coordinates of parents of node i

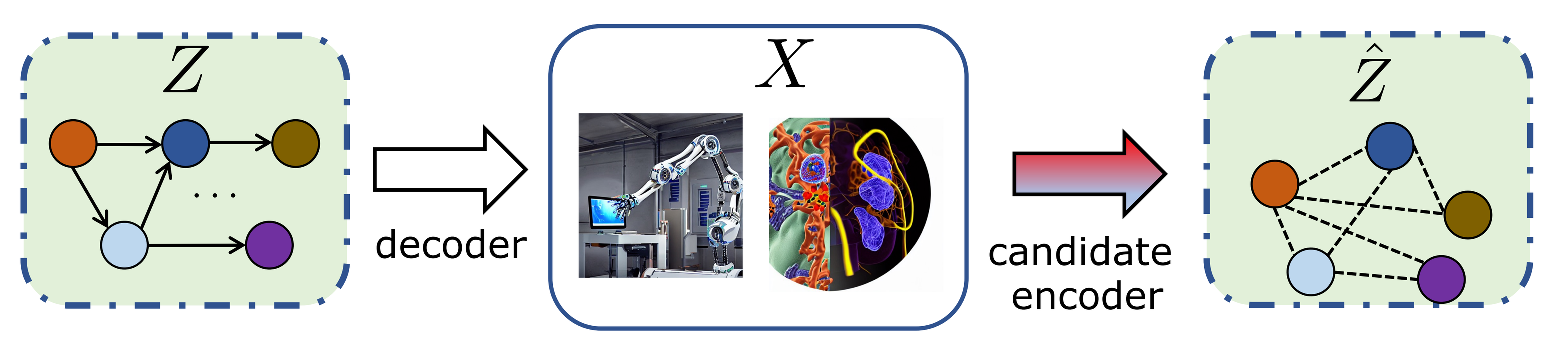
Coupled hard interventions have sparse score differences

node i intervened twice: $s^m(z) - \tilde{s}^m(z)$ becomes a function of only z_i

$$s^m(z) - \tilde{s}^m(z) = \begin{bmatrix} 0 & 0 & 0 & 0 & \boxed{\times} & 0 \end{bmatrix}^\top$$

coordinate of node i

Methodology



incorrect encoder $\rightarrow$ $s_{\hat{Z}}(\hat{z}) - s_{\hat{Z}}^m(\hat{z})$ **not** a function of only $z_{\overline{\text{pa}}(i)}$

estimated score differences cannot be sparser than true score differences

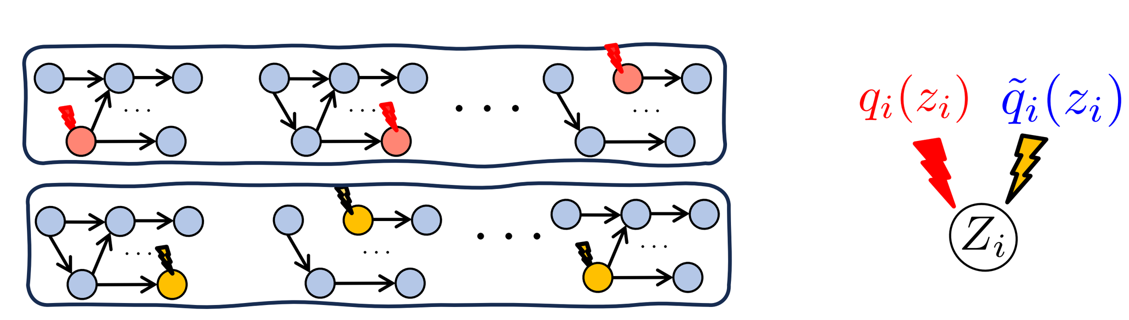
⇩

Min. score variations over environment pairs = correct encoder

$$s_{\hat{Z}}(\hat{z}) - s_{\hat{Z}}^m(\hat{z}) = [J_{\text{decoder}}(\hat{z})]^\top (s_X(x) - s_X^m(x))$$

Results

Nonparametric transform



Interventional discrepancy: $\dfrac{\partial}{\partial z_i} \dfrac{q_i(z_i)}{\tilde{q}_i(z_i)} \neq 0$ almost everywhere

Theorem : Observational data and **two hard** interventions/node suffice for **perfect ID:**

1. Latent graph recovery up to isomorphism
2. Latent variables recovery up to elementwise transform

- Achievable algorithm
- No faithfulness assumption for identifiability
- Uncoupled two hard interventions per node

correspondence: burakvarici@gmail.com

project page