Causal Representation Learning:

General Identifiability and Achievability

Ali Tajer

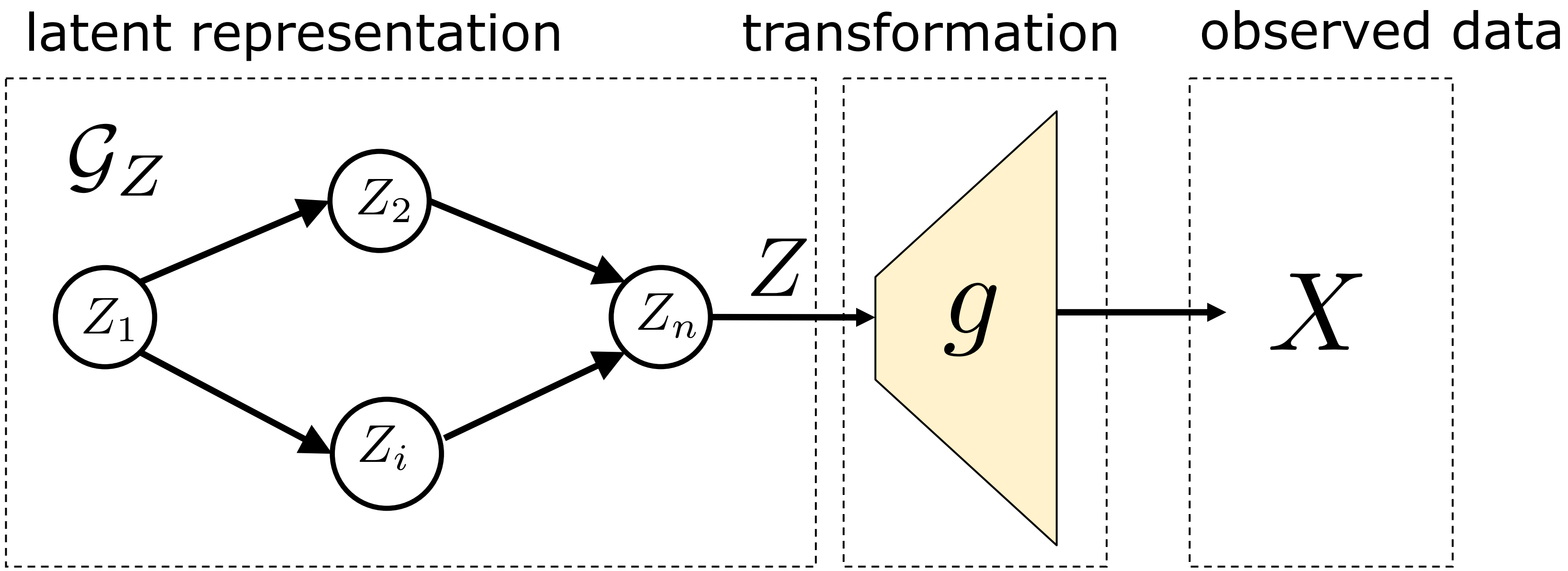
Rensselaer Polytechnic Institute

Burak Varıcı
RPI

Emre Acartürk
RPI

Karthikeyan Shanmugam
Google

Causal Representation Learning (CRL)

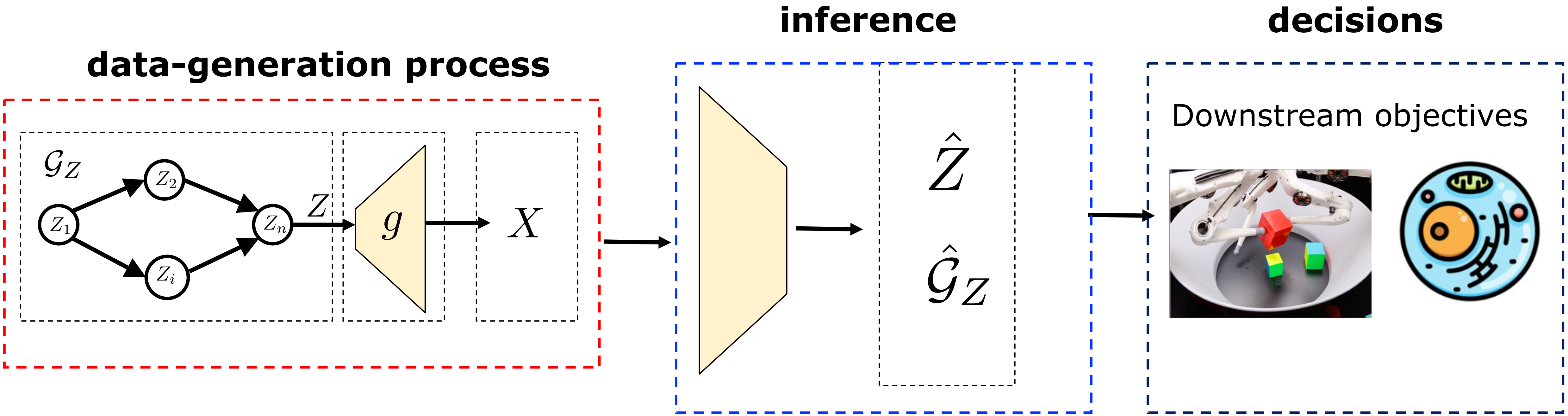


$$X = g(Z)$$

use data X to find:

1. Z: latent random variables
2. $\mathcal{G}_Z$: latent causal graph
3. g: transformation

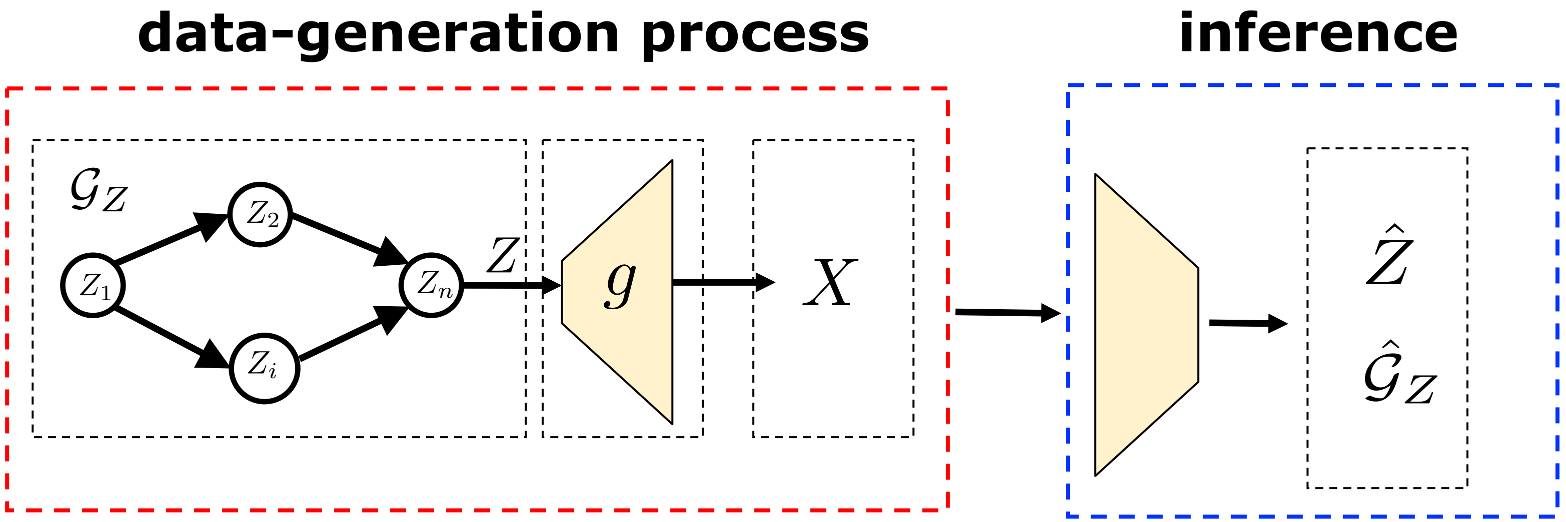
Why CRL?



- Robotics: learning a robot's dynamics from images
- Genomics: learning the causal variables from gene-expressions

Schölkopf, Locatello, Bauer, Ke, Kalchbrenner, Goyal, Bengio. "Toward causal representation learning" Proceedings of the IEEE 109, no. 5 (2021): 612-634.

CRL Objectives

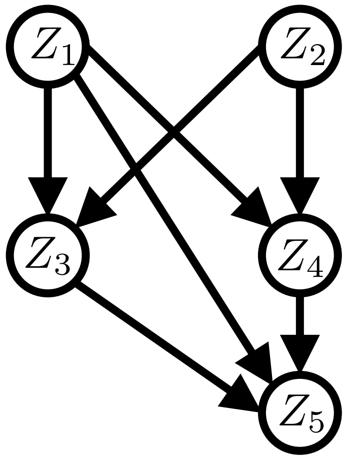


1. Identifiability: Conditions for uniquely recovering Z and $\mathcal{G}_Z$
2. Achievability: Provably correct algorithms for recovering Z and $\mathcal{G}_Z$

Interventions

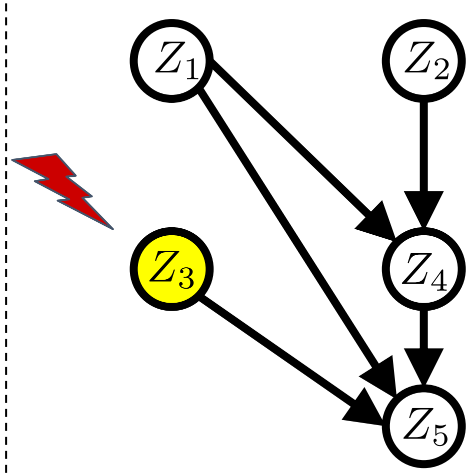
CRL is impossible without sufficient statistical diversity

observational



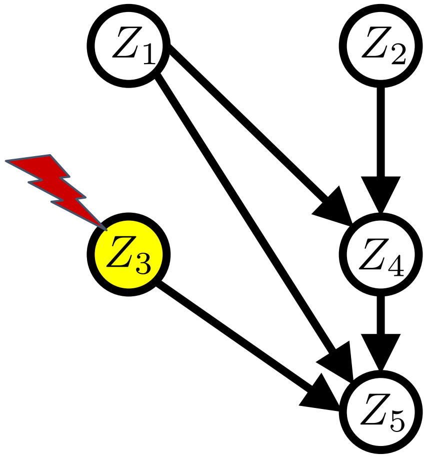
$$p(Z_3|Z_1, Z_2)$$

do



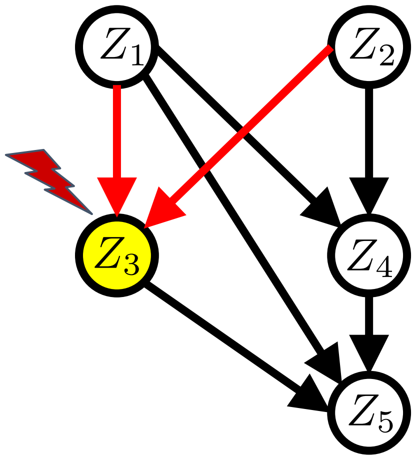
$$\begin{cases} 1 & \text{for} \quad Z_3 = Z \\ 0 & \text{for} \quad Z_3 \neq Z \end{cases}$$

hard (perfect)



$$q(Z_3)$$

soft (imperfect)



$$q(Z_3|Z_1, Z_2)$$

5

State of CRL under Soft Interventions

	Transform Parametric	Transform General
Latent Models Parametric	**Partial ID + Algo. [3,6]** Hard: Perfect ID + Algo [3]	**Partial ID + Algo. [4]** Hard: Perfect ID + Algo [4]
Latent Models General	**Partial ID + Algo. [2,5]** do: Perfect ID + Algo [2]	**----** Hard: Perfect ID [1]

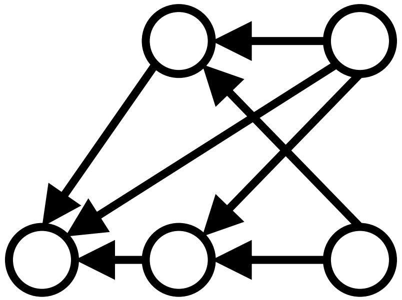
[1] von Kügelgen et al. *Nonparametric identifiability of causal representations from unknown interventions"*. NeurIPS 2023
[2] Ahuja et al. *"Interventional causal representation learning".* ICML 2023
[3] Squires et al. *"Linear causal disentanglement via interventions".* ICML 2023
[4] Buchholz et al. *"Learning linear causal representations from interventions under general nonlinear mixing"*. NeurIPS 2023
[5] Zhang et al. *"Identifiability guarantees for causal disentanglement from soft interventions".* NeurIPS 2023
[6] Jin and Syrgkanis. *"Learning Causal Representations from General Environments: Identifiability and Intrinsic Ambiguity"*

What Score-based CRL can do?

Latent model	Transform	Interventions	Main results
General	General	Two hard	1. perfect ID 2. provably correct algo
General	Linear	One hard	1. perfect ID 2. provably correct algo
General	Linear	One soft	1. ID up to ancestors 2. provably correct algo

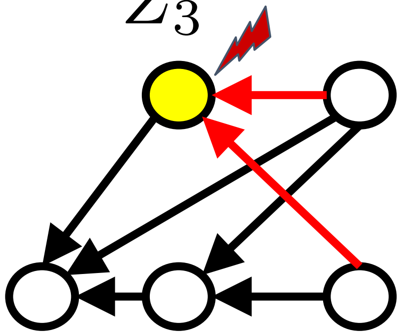
Connection to Score Difference

$$p_3(z_3 \mid z_{\mathrm{pa}(3)})$$



$$p(Z) = p_3(z_3 \mid z_{\mathrm{pa}(3)}) \prod_{i \neq 3} p_i(z_i \mid z_{\mathrm{pa}}(i))$$

$$q_3(z_3 \mid z_{\mathrm{pa}(3)})$$



$$p^3(Z) = q_3(z_3 \mid z_{\mathrm{pa}(3)}) \prod_{i \neq 3} p_i(z_i \mid z_{\mathrm{pa}}(i))$$

$$\underbrace{\nabla_z \log p(Z)}_{s(Z)} - \underbrace{\nabla_z \log p^3(Z)}_{s^3(Z)} = \begin{bmatrix} 0 \\ 0 \\ \times \\ 0 \\ \times \\ 0 \end{bmatrix}$$

coordinates of parents of node i

Score Difference Properties

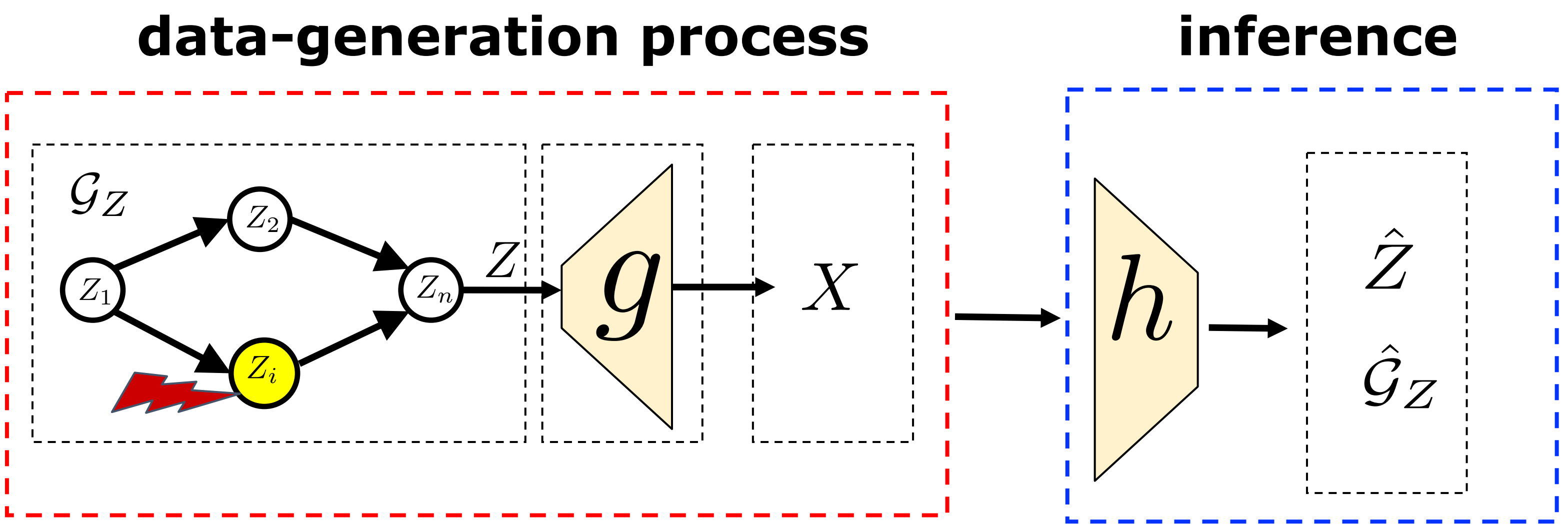
$$s(z) - s^i(z) = \nabla_z p_i(z_i \mid z_{\mathrm{pa}(i)}) - \nabla_z q_i(z_i \mid z_{\mathrm{pa}(i)}) \quad = \text{function of only } z_{\mathrm{pa}(i)}$$

Two properties:

non-zero coordinates of score difference = parents of intervention target

estimated score differences cannot be sparser than true score differences

Score-based CRL

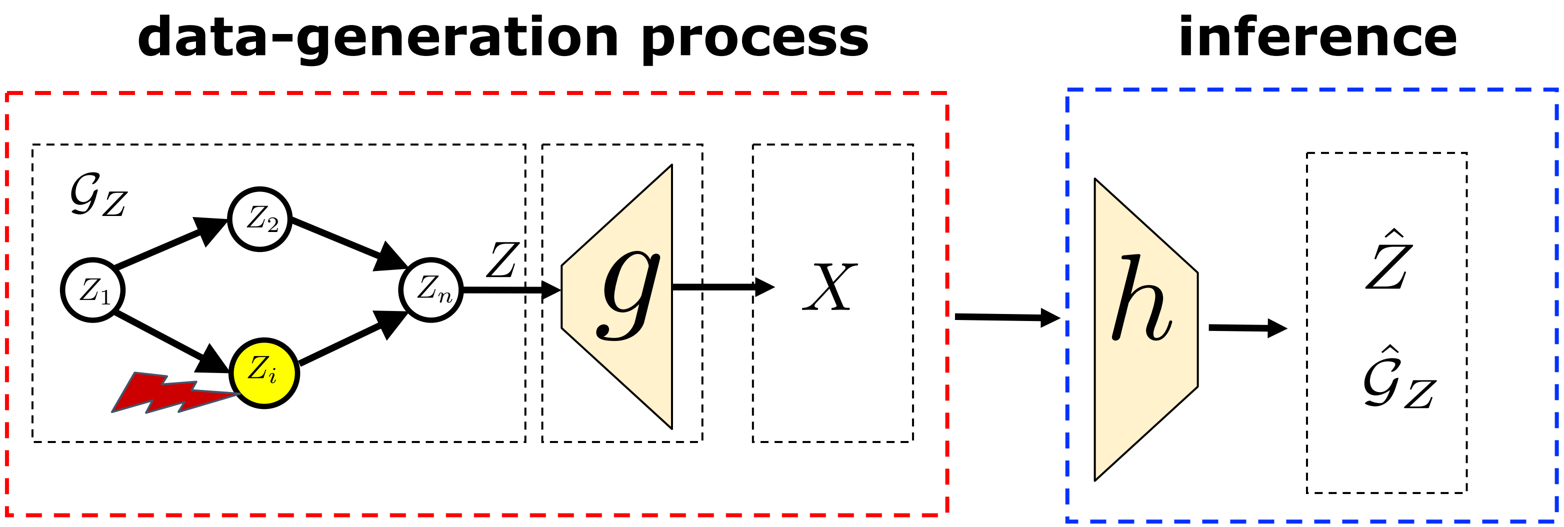


$$Z \xrightarrow{g} X \xrightarrow{h} \hat{Z}(h)$$

correct encoder, i.e., $h(g(Z)) = Z \quad \rightarrow \quad s(\hat{z}) - s^i(\hat{z})$ function of only $z_{\overline{\mathrm{pa}}(i)}$

incorrect encoder, i.e., $h(g(Z)) \neq Z \quad \rightarrow \quad s(\hat{z}) - s^i(\hat{z})$ **not** a function of only $z_{\overline{\mathrm{pa}}(i)}$

Score-based CRL



$$Z \xrightarrow{g} X \xrightarrow{h} \hat{Z}(h)$$

$$\min_h \; \left\| \text{estimated score difference vector} \right\|_0$$

Provably correct algorithm for unsupervised learning
(small variations for each setting)

Score Estimation

Q: How to compute $[s(\hat{z}) - s^i(\hat{z})]$ using X?

$$s(\hat{z}) - s^i(\hat{z}) = [J_{\text{decoder}}(\hat{z})]^\top [s_X(x) - s_X^i(x)]$$

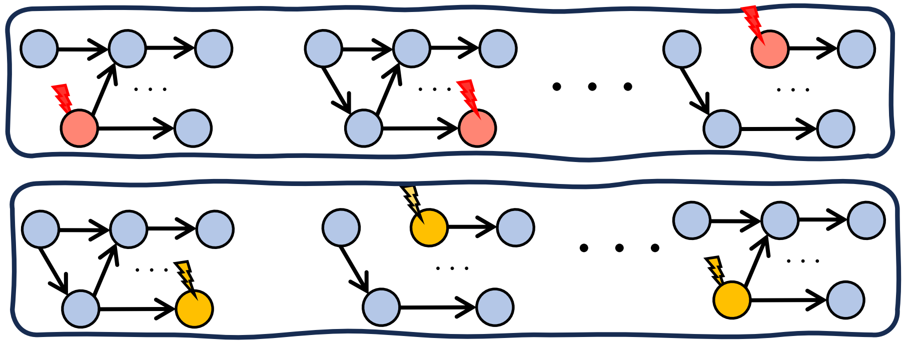
Q: Is knowing intervention-node assignments needed?

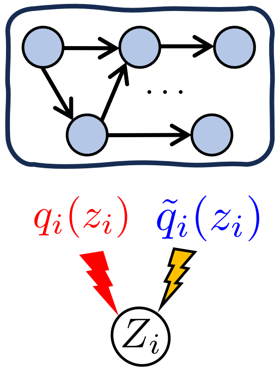
A: No – knowing that each node is intervened in one environment is sufficient

Q: What score estimation algorithm to use?

A: the design is agnostic to the choice – we are using sliced score matching

General Transform + General Latent + 2 Hard





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

Theorem : Observational data and **two hard** interventions/node = **Perfect ID + Algorithm**

Note: von Kügelgen'2023: **coupled** two hard + **faithfulness** (for all candidates) = **Perfect ID**

Linear Transform + General Latent

Theorem : Linear transform + **1 hard**/node = **Perfect ID + Algorithm**

Note: this subsumes Squires'23, Buchholz'23 **results on linear** latent models

Theorem : Linear transform + **1 soft**/node = **ID up to ancestors + Algorithm**

Note: Squires'23: in linear latent models + soft: **ID up to ancestors is the best one can hope.**

Empirical Results

Nonlinear latent model:

$$Z_i = \sqrt{Z_{\mathrm{pa}(i)}^\top A_{p,i} Z_{\mathrm{pa}(i)}} + N_{p,i}$$

Nonlinear transform:

$$X = \tanh(G.Z)$$

Score estimation: sliced score matching

n=5 latent variables

Obs. dim	Norm. Z error	DAG error (SHD)	Norm. Z error	DAG error (SHD)
5	0.03	0.12	1.19	5.1
25	0.03	0.04	1.09	4.4
40	0.04	0.02	0.86	5.0

score oracle | noisy scores

Summary

- A general framework for establishing ID and Achievability guarantees
- Score difference functions contain all the information needed about latent DAGs
- Minimize score variations, constructive ID proof + provably correct algorithms
- General transform with 2 interventions/node: https://arxiv.org/abs/2310.15450
- Linear transform with 1 intervention/node: https://arxiv.org/abs/2402.00849

Poster Session 3 (Saturday), Number 4

References

B. Varıcı, E. Acartürk, K. Shanmugam, and A. Tajer. "*General identifiability and achievability for causal representation learning*". AISTATS 2024.

B. Varıcı, E. Acartürk, K. Shanmugam, A. Kumar, and A. Tajer. "*Score-based causal representation: Linear and general transformations*". arXiv: 2402.00849

B. Varıcı, E. Acartürk, K. Shanmugam, A. Kumar, and A. Tajer. "*Score-based causal representation with interventions*". arXiv: 2301.08230, 2023

K. Ahuja, D. Mahajan, Y. Wang, and Y. Bengio. *"Interventional causal representation learning".* ICML 2023

C. Squires, A. Seigal, S. S. Bhate, and C. Uhler. *"Linear causal disentanglement via interventions".* ICML 2023

J. von Kügelgen, M. Besserve, W. Liang, L. Gresele, A. Kekić, E. Bareinboim, D. M. Blei, and B. Schölkopf. *"Nonparametric identifiability of causal representations from unknown interventions"*. NeurIPS 2023

S. Buchholz, G. Rajendran, E. Rosenfeld, B. Aragam, B. Schölkopf, and P. Ravikumar. *"Learning linear causal representations from interventions under general nonlinear mixing"*. NeurIPS 2023

J. Zhang, C. Squires, K. Greenewald, A. Srivastava, K. Shanmugam, and C. Uhler. "*Identifiability guarantees for causal disentanglement from soft interventions".* NeurIPS 2023

W. Liang, A. Kekić, J. von Kügelgen, S. Buchholz, M. Besserve, L. Gresele, and B. Schölkopf. "*Causal component analysis"*. NeurIPS 2023

Y. Jiang and B. Aragam. "*Learning nonparametric latent causal graphs with unknown interventions*". NeurIPS 2023

S. Bing, U. Ninad, J. Wahl, and J. Runge. "*Identifying linearly-mixed causal representation learning from multi-node interventions*", CLeaR 2024.

J. Jin and V. Syrgkanis. "*Learning Causal Representations from General Environments: Identifiability and Intrinsic Ambiguity*." arXiv:2311.12267, 2023.