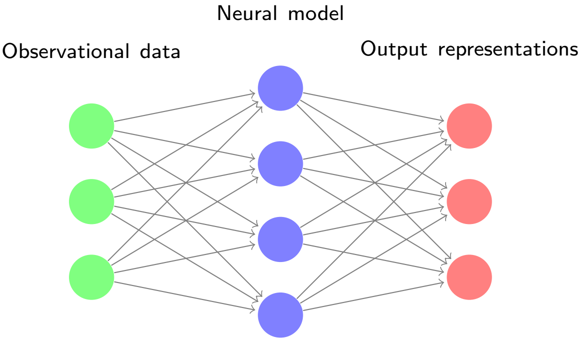
Learning Linear Causal Representations from Interventions under General Nonlinear Mixing

Simon Buchholz* (MPI), **Goutham Rajendran*** (CMU), Elan Rosenfeld (CMU), Bryon Aragam (UChicago), Bernhard Schölkopf (MPI), Pradeep Ravikumar (CMU)

NeurIPS 2023

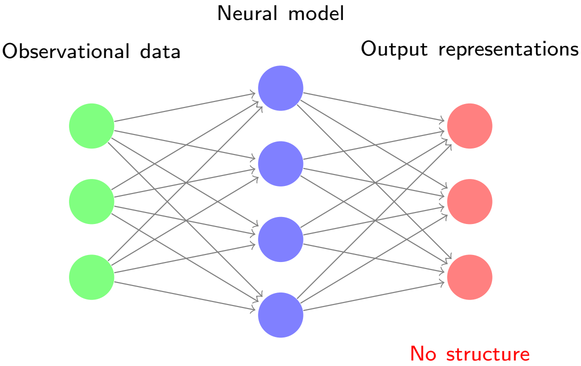
Representation Learning

- Traditional representation learning, used for generative modeling:



Representation Learning

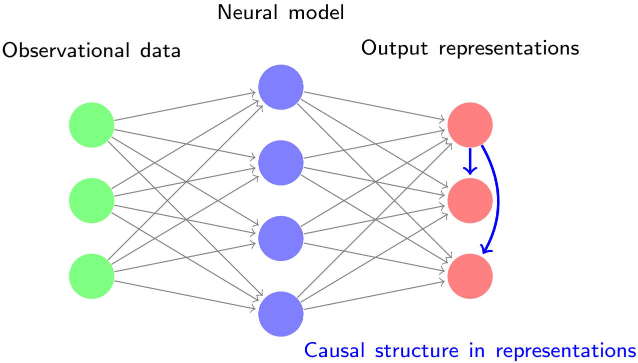
- Traditional representation learning, used for generative modeling:



- Drawbacks:
  - No structure in representations
  - Representations are not interpretable or controllable
  - Susceptibility to bias, poor generalization capabilities

Causal Representation Learning

- Causal representation learning, an emerging field aiming to resolve this issue:



- Causal representations will be more robust, interpretable and also enable alignment

Causal Representation Learning

- Observed data $X = f(Z)$ - complex, high-dimensional
- Z – simple, low-dimensional, e.g. Gaussian
- f – mixing function

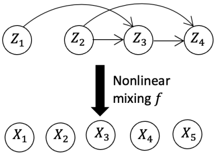


Figure: Generative model

Causal Representation Learning

- Observed data $X = f(Z)$ - complex, high-dimensional
- Z – simple, low-dimensional, e.g. Gaussian
- f – mixing function

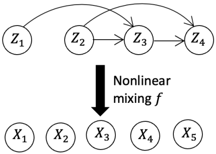


Figure: Generative model

- Example:
  - Z – position, type, and size of objects
  - f – rendering of image
  - X – image

- Goal: Identify f as well as Z

Causal Representation Learning

- Learning ground truth Z, f leads to
  - Recovery of causal structure
  - OOD generalization
  - Robustness
  - Reliability
- Special case - Causal disentanglement (independent latents)

Causal Representation Learning

- Learning ground truth Z, f leads to
  - Recovery of causal structure
  - OOD generalization
  - Robustness
  - Reliability
- Special case - Causal disentanglement (independent latents)

- Issue: Impossible!, for any X a huge class of Z and f
- Prior works:
  - Parametric assumptions: [Hyvarinen-Oja 2000]
  - Semi-supervised: [Khemakhem et al. 2020]
  - Functional assumptions: [Kivva et al. 2022], [Buchholz et al. 2022]
  - Interventional data - [Lippe et al. 2022, Squires et al. 2023]

Interventional Causal Representation Learning

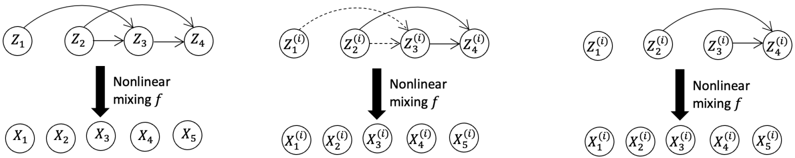
- Interventional data: A bunch of additional datasets (environments)
- Example: Images of rooms with and without lights
- Predominant in robotics: Agent explores environment via interventions

Interventional Causal Representation Learning

- Interventional data: A bunch of additional datasets (environments)
- Example: Images of rooms with and without lights
- Predominant in robotics: Agent explores environment via interventions

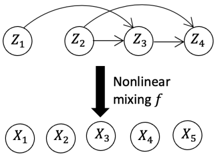
- Long line of prior works
  - All variables observed: Hauser et al. 2012, Peters et al. 2015, Squires et al. 2020, Jaber et al. 2020, Eberhardt et al. 2012, ...
  - Latent variables present: Zimmermann et al. 2021, Rosenfeld et al. 2021, Lippe et al. 2022, Lachapelle et al. 2022, Brehmer et al. 2022, Ahuja et al. 2022, Seigal et al. 2022, Ahuja et al. 2022, Rosenfeld et al. 2022, Chen et al. 2022, Varici et al. 2023

Our setting

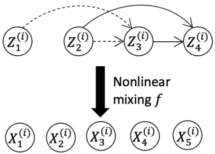


(a) No interventions (b) An imperfect intervention (c) A perfect intervention

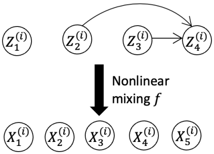
Our setting



(a) No interventions



(b) An imperfect intervention



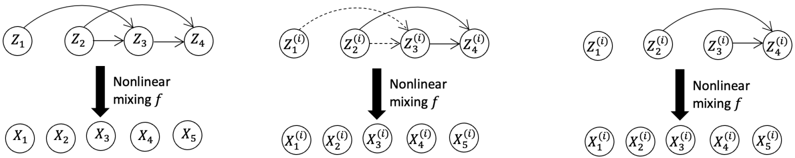
(c) A perfect intervention

- Linear Gaussian priors with non-linear mixing

$$Z = AZ + D^{1/2}\epsilon, \qquad A \text{ is a DAG, } D \text{ diagonal, } \epsilon \sim N(0, I)$$

$$X = f(Z), \qquad f \text{ injective, differentiable}$$

Our setting



(a) No interventions (b) An imperfect intervention (c) A perfect intervention

- Linear Gaussian priors with non-linear mixing

$$Z = AZ + D^{1/2}\epsilon, \qquad A \text{ is a DAG, } D \text{ diagonal, } \epsilon \sim N(0, I)$$
$$X = f(Z), \qquad f \text{ injective, differentiable}$$

- Single-node interventions: For target node t_i, change mean and var and dependence on parents (perfect intervention = no dependence).

$$Z_{t_i}^{(i)} = \underbrace{(A^{(i)} Z^{(i)})_{t_i}}_{\substack{\text{weights to parents change} \\ \text{0 for perfect}}} + \underbrace{(D^{(i)})^{1/2}_{t_i,t_i}}_{var}(\epsilon_{t_i} + \underbrace{\eta^{(i)}}_{mean})$$

Our main guarantees

- Assume we are given interventional datasets with
  - Linear Gaussian latent variables with non-linear mixing
  - Perfect single node interventions
  - All nodes are intervened upon

Our main guarantees

- Assume we are given interventional datasets with
  - Linear Gaussian latent variables with non-linear mixing
  - Perfect single node interventions
  - All nodes are intervened upon

Theorem (informal)

Under these assumptions we can identify f, Z, and the causal graph (up to trivial transformations)

Our main guarantees

- Assume we are given interventional datasets with
  - Linear Gaussian latent variables with non-linear mixing
  - Perfect single node interventions
  - All nodes are intervened upon

Theorem (informal)

Under these assumptions we can identify f, Z, and the causal graph (up to trivial transformations)

- We also extend to imperfect interventions

Our main guarantees

- Assume we are given interventional datasets with
  - Linear Gaussian latent variables with non-linear mixing
  - Perfect single node interventions
  - All nodes are intervened upon

Theorem (informal)

Under these assumptions we can identify f, Z, and the causal graph (up to trivial transformations)

- We also extend to imperfect interventions
- We show our assumptions are necessary (via counterexamples)

Comparison to prior works

- Closely related prior works:

Paper	Setting	Our work
[Squires et al. 2023]	linear Z, f	Non-linear f
[Varici et al. 2023, Jiang et al. 2023]	non-linear Z, linear f	linear Z, non-linear f
[Ahuja et al. 2022]	polynomial f, do-interventions	non-linear f, soft interventions

Comparison to prior works

- Closely related prior works:

Paper	Setting	Our work
[Squires et al. 2023]	linear Z, f	Non-linear f
[Varici et al. 2023, Jiang et al. 2023]	non-linear Z, linear f	linear Z, non-linear f
[Ahuja et al. 2022]	polynomial f, do-interventions	non-linear f, soft interventions

- Concurrent works: [Zhang et al. 2023], [Liang et al. 2023], [von Kügelgen et al. 2023]

Comparison to prior works

- Closely related prior works:

Paper	Setting	Our work
[Squires et al. 2023]	linear Z, f	Non-linear f
[Varici et al. 2023, Jiang et al. 2023]	non-linear Z, linear f	linear Z, non-linear f
[Ahuja et al. 2022]	polynomial f, do-interventions	non-linear f, soft interventions

- Concurrent works: [Zhang et al. 2023], [Liang et al. 2023], [von Kügelgen et al. 2023]
- Other highlights of our work:
  - Non-paired data
  - Unknown targets
  - Can handle perfect/imperfect/soft interventions

Experimental methodology

- Usual approach: Use Variational Autoencoders to learn encoder $X \rightarrow Z$ and decoder $Z \rightarrow X$
- However, we don't know intervention targets, so not usable

Experimental methodology

- Usual approach: Use Variational Autoencoders to learn encoder $X \to Z$ and decoder $Z \to X$
- However, we don't know intervention targets, so not usable
- Our approach: Contrastive learning

Experimental methodology

- Usual approach: Use Variational Autoencoders to learn encoder $X \to Z$ and decoder $Z \to X$
- However, we don't know intervention targets, so not usable
- Our approach: Contrastive learning
- Train a deep neural network to distinguish
  - Observational samples $x \sim X^{(0)}$ from
  - Interventional samples $x \sim X^{(i)}$

Experimental methodology

- Usual approach: Use Variational Autoencoders to learn encoder $X \to Z$ and decoder $Z \to X$
- However, we don't know intervention targets, so not usable
- Our approach: Contrastive learning
- Train a deep neural network to distinguish
  - Observational samples $x \sim X^{(0)}$ from
  - Interventional samples $x \sim X^{(i)}$
- Choose the last layer to model Gaussian log-density
- Makes sense because optimal Bayes classifier should look like this

Experimental methodology

- Gaussian log-odds: The log-odds of a sample $x \sim X^{(i)}$ over $x \sim X^{(0)}$ is given by

$$\ln p_X^{(i)}(x) - \ln p_X^{(0)}(x) = c_i - \frac{1}{2}\lambda_i^2((f^{-1}(x)_{t_i})^2 + \eta^{(i)}\lambda_i \cdot (f^{-1}(x))_{t_i} + \frac{1}{2}\langle f^{-1}(x), s^{(i)}\rangle^2$$

Experimental methodology

- Gaussian log-odds: The log-odds of a sample $x \sim X^{(i)}$ over $x \sim X^{(0)}$ is given by

$$\ln p_X^{(i)}(x) - \ln p_X^{(0)}(x) = c_i - \frac{1}{2}\lambda_i^2((f^{-1}(x)_{t_i})^2 + \eta^{(i)}\lambda_i \cdot (f^{-1}(x))_{t_i} + \frac{1}{2}\langle f^{-1}(x), s^{(i)}\rangle^2$$

- So pick last layer to be (h is deep network intended to be f^{-1})

$$g_i(x, \alpha_i, \beta_i, \gamma_i, w^{(i)}, \theta) = \alpha_i - \beta_i h_{t_i}^2(x,\theta) + \gamma_i h_{t_i}(x,\theta) + \langle h(x,\theta), w^{(i)}\rangle^2$$

- Loss function:

$$\mathcal{L} = \underbrace{\sum_{i \in I} \mathcal{L}_{\mathrm{CE}}^{(i)}}_{\text{Cross-Entropy loss}} + \tau_1 \underbrace{\mathcal{R}_{NOTEARS}(W)}_{\text{acyclicity regularizer}} + \tau_2 \underbrace{\mathcal{R}_{REG}(W)}_{\text{sparsity regularizer}}$$

Experiments - Synthetic data

- Sample random DAG and non-linear 3-layer MLP f

Setting	Method	SHD ↓	AUROC ↑	MCC ↑	R^2 ↑
Non-linear f	**Contrastive**	1.8 ± 0.5	0.97 ± 0.01	0.97 ± 0.00	0.96 ± 0.00
ER(5, 2) DAG, n = 10k	VAE	10.0 ± 0.0	0.50 ± 0.00	0.48 ± 0.03	0.57 ± 0.07
d = 5, d' = 20	Linear baseline	10.6 ± 1.9	0.48 ± 0.11	0.32 ± 0.03	0.34 ± 0.06
Non-linear f	**Contrastive**	1.6 ± 0.5	1.00 ± 0.00	0.98 ± 0.00	0.97 ± 0.00
ER(10, 2) DAG, n = 10k	VAE	18.6 ± 0.9	0.50 ± 0.00	0.62 ± 0.02	0.78 ± 0.01
d = 10, d' = 100	Linear baseline	28.4 ± 2.1	0.51 ± 0.04	0.17 ± 0.03	0.13 ± 0.03

- Metrics:
  - SHD - Structural Hamming Distance (a measure of distance between graphs)
  - MCC - Mean Correlation Coefficient (a measure of recovery of latent variables)

Experiments - Synthetic data

- Sample random DAG and non-linear 3-layer MLP f

Setting	Method	SHD ↓	AUROC ↑	MCC ↑	R^2 ↑
Non-linear f	**Contrastive**	1.8 ± 0.5	0.97 ± 0.01	0.97 ± 0.00	0.96 ± 0.00
ER(5, 2) DAG, n = 10k	VAE	10.0 ± 0.0	0.50 ± 0.00	0.48 ± 0.03	0.57 ± 0.07
d = 5, d' = 20	Linear baseline	10.6 ± 1.9	0.48 ± 0.11	0.32 ± 0.03	0.34 ± 0.06
Non-linear f	**Contrastive**	1.6 ± 0.5	1.00 ± 0.00	0.98 ± 0.00	0.97 ± 0.00
ER(10, 2) DAG, n = 10k	VAE	18.6 ± 0.9	0.50 ± 0.00	0.62 ± 0.02	0.78 ± 0.01
d = 10, d' = 100	Linear baseline	28.4 ± 2.1	0.51 ± 0.04	0.17 ± 0.03	0.13 ± 0.03

- Metrics:
  - SHD - Structural Hamming Distance (a measure of distance between graphs)
  - MCC - Mean Correlation Coefficient (a measure of recovery of latent variables)
- Our contrastive method outperforms linear baseline as well as VAE based approaches.

Experiments - Image Data

- Sample DAG to generate coordinates of balls.
- f is an image rendering (non-linear) of balls

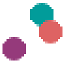


Figure: Sample image with 3 balls

Experiments - Image Data

- Sample DAG to generate coordinates of balls.
- f is an image rendering (non-linear) of balls

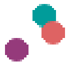


Figure: Sample image with 3 balls

Table: $d = 2 \cdot \#\text{balls}$ and $n_{\text{int}} = 25000$ (per environment), $n_{\text{obs}} = n_{\text{int}} \cdot d$.

# Balls	Method	SHD ↓	AUROC ↑	MCC ↑	R^2 ↑
2	Contrastive Learning	1.4 ± 0.4	0.95 ± 0.03	0.87 ± 0.03	0.84 ± 0.03
	VAE	6.0 ± 0.0	0.50 ± 0.00	0.19 ± 0.06	0.16 ± 0.08
5	Contrastive Learning	2.0 ± 0.3	1.00 ± 0.00	0.94 ± 0.01	0.91 ± 0.01
	VAE	18.6 ± 0.9	0.50 ± 0.00	0.31 ± 0.02	0.36 ± 0.03
10	Contrastive Learning	11.0 ± 3.3	0.98 ± 0.02	0.89 ± 0.01	0.83 ± 0.01
	VAE	37.2 ± 3.1	0.50 ± 0.00	0.22 ± 0.01	0.33 ± 0.02

Summary

- We saw interventional causal representation learning
- Identifiable for
  - Gaussian priors (common assumption)
  - Non-linear f (completely general)
  - Single-node intervention on all nodes
- Contrastive learning algorithm to learn the model

- **Future work**
  - Will contrastive algorithm scale?
  - Non-linear Z, multi-node interventions, etc.

Thank You