Learning Causally Disentangled Representations via the Principle of Independent Causal Mechanisms

Aneesh Komanduri[1], Yongkai Wu[2], Feng Chen[3], and Xintao Wu[1]

[1] University of Arkansas, Fayetteville, AR, USA

[2] Clemson University, Clemson, SC, USA

[3] University of Texas at Dallas, Richardson, TX, USA

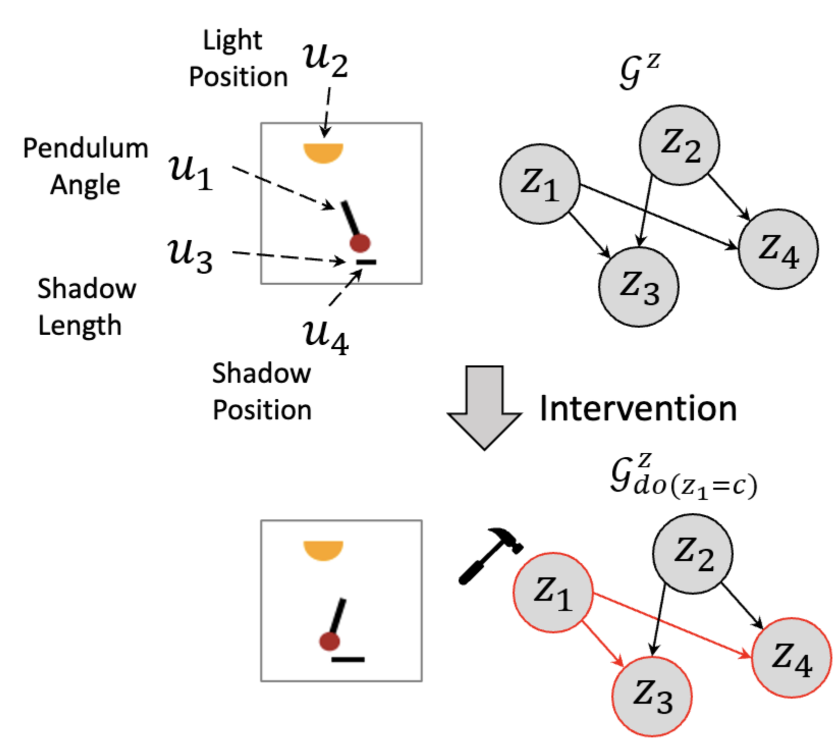
IJCAI JEJU 2024

Overview

❖ The goal of causal representation learning is to map low-level data to high-level concepts that are causally related.

❖ **Motivation**

- There is a lack of a unified definition for disentanglement from the perspective of independent causal mechanisms (i.e., recovering the true causal *mechanisms*).
- Latent causal models are often parameterized to be linear additive noise and are not general enough to accurately model causal mechanisms.



❖ **Our Contributions**

- We propose a new definition of causal disentanglement inspired by the principle of independent causal mechanisms.
- We propose ICM-VAE, a learning framework to flexibly learn causally disentangled representations with a causally factorized prior.
- We show identifiability of causal mechanisms up to permutation equivalence and empirically show disentanglement and counterfactual generation capability.

Preliminaries

Structural Causal Model (SCM)

❖ SCM $\mathcal{M} = \langle \mathcal{Z}, \mathcal{E}, F \rangle$

- Exogenous noise variables $\epsilon = \{\epsilon_1, \epsilon_2, \ldots, \epsilon_n\}$, and distribution $p(\epsilon)$
- Endogenous causal variables $z = \{z_1, z_2, \ldots, z_n\}$
- Functions $F = \{f_1, f_2, \ldots, f_n\} \rightarrow z_i = f_i(\epsilon_i, z_{\mathbf{pa}_i})$
- Independent Causal Mechanisms: $p(z_1, \ldots, z_n) = \prod_{i=1}^{n} p(z_i | z_{\mathbf{pa}_i})$

Generative Model Identifiability

❖ Identifiability: Can we recover true generative factors up to trivial transformation?

❖ Definition. Let $\sim$ be an equivalence relation on θ. A generative model is $\sim$-identifiable if $p_\theta(x) = p_{\hat{\theta}}(x) \implies \theta \sim \hat{\theta}$

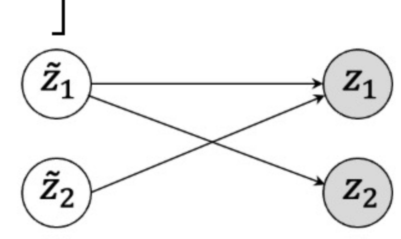
❖ Identifiable Variational Autoencoder (iVAE) – condition on auxiliary info u:

$$p_{T,\lambda}(z|u) = \prod_i h_i(z_i) \exp\left[\sum_{j=1}^{k} T_{i,j}(z_i)\lambda_{i,j}(u) - \psi_i(u)\right]$$

❖ Linear-equivalent (recovery up to linear transformation):

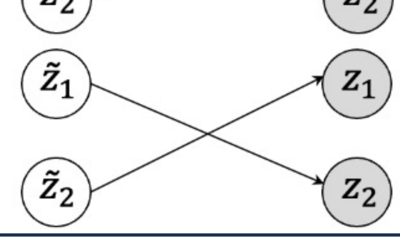
$$\mathbf{T}(g^{-1}(x)) = A\hat{\mathbf{T}}(\hat{g}^{-1}(x)) + b, \forall x \in \mathcal{X}$$

$$A^T \boldsymbol{\lambda}(u) + c = \hat{\boldsymbol{\lambda}}(u)$$

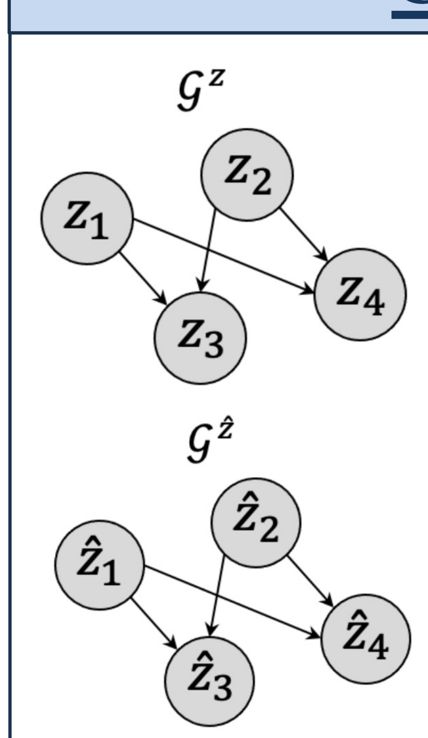


❖ Permutation-equivalent (recovery up to reordering):

$$P\hat{z} = [z_{\pi(1)}A_1^T, z_{\pi(2)}A_2^T, \ldots, z_{\pi(n)}A_n^T]^T$$



Causal Mechanism Equivalence



$$z_1 = \epsilon_1 \sim \mathcal{N}(0,1) \qquad z_2 = \epsilon_2 \sim \mathcal{N}(0,1) \qquad z_3 = 4z_1 + \sqrt{8}z_2 + \epsilon_3 \qquad z_4 = 2z_1 + 3z_2 + \epsilon_4$$

Marginalize $\rightarrow$

$$z_1 \sim \mathcal{N}(0,1) \qquad z_2 \sim \mathcal{N}(0,1) \qquad z_3 \sim \mathcal{N}(0, \sqrt{4^2 + \sqrt{8}^2 + 1^2} = 5) \qquad z_4 \sim \mathcal{N}(0, \sqrt{2^2 + 3^2 + 1^2} = \sqrt{14})$$

$$\hat{z}_1 = \hat{\epsilon}_1 \sim \mathcal{N}(0,1) \qquad \hat{z}_2 = \hat{\epsilon}_2 \sim \mathcal{N}(0,1) \qquad \hat{z}_3 = \sqrt{20}\hat{z}_1 + 2\hat{z}_2 + \hat{\epsilon}_3 \qquad \hat{z}_4 = \sqrt{5}\hat{z}_1 + \sqrt{8}\hat{z}_2 + \hat{\epsilon}_4$$

Marginalize $\rightarrow$

$$\hat{z}_1 \sim \mathcal{N}(0,1) \qquad \hat{z}_2 \sim \mathcal{N}(0,1) \qquad \hat{z}_3 \sim \mathcal{N}(0, \sqrt{\sqrt{20}^2 + 2^2 + 1^2} = 5) \qquad \hat{z}_4 \sim \mathcal{N}(0, \sqrt{\sqrt{5}^2 + \sqrt{8}^2 + 1^2} = \sqrt{14})$$

❖ **Issue**: Learned mechanisms may be different than true underlying mechanisms but produce same marginal distribution

❖ **Idea**: What if we consider disentanglement from a causal mechanism perspective?

$$p_\theta(z_i | z_{\mathbf{pa}_i}) = p_{\hat{\theta}}(z_i | z_{\mathbf{pa}_i})$$

❖ Three *sufficient* conditions for causal mechanism equivalence

1. z and $\hat{z}$ must be permutation equivalent
2. Equivalence of conditional sufficient statistics: $\mathbf{T}_i(z_i|z_{\mathbf{pa}_i}) = D_{ij}\hat{\mathbf{T}}_j(z_j|z_{\mathbf{pa}_j})$
3. Natural parameter mechanism equivalence: $\boldsymbol{\lambda}_i(z_{\mathbf{pa}_i}, u) = D_{ij}\hat{\boldsymbol{\lambda}}_j(z_{\mathbf{pa}_j}, u)$

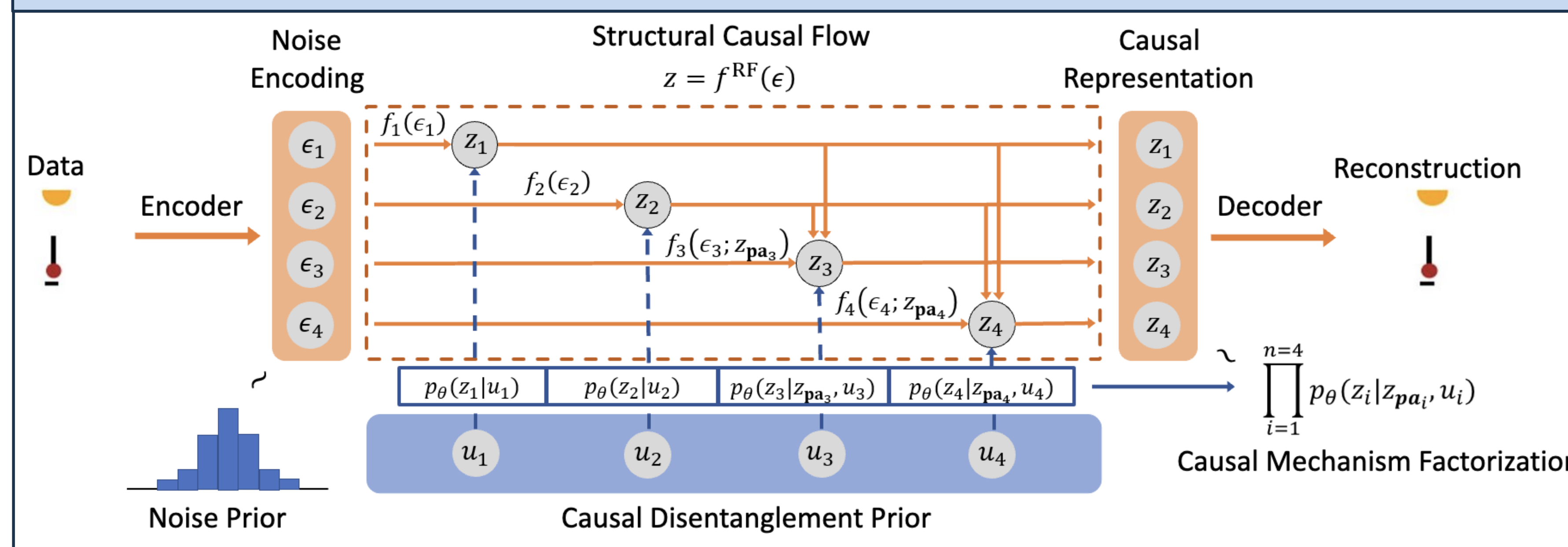
$\Rightarrow$ Causal Mechanism Permutation Equivalent and Causally Disentangled

References.

[1] F. Locatello et al. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. ICML 2019.

[2] I. Khemakhem et al. Variational Autoencoders and Nonlinear ICA: A Unifying Framework. AISTATS 2020.

ICM-VAE Framework



Structural Causal Flow

❖ Causal mechanisms parameterized by affine-form autoregressive flow

$$z = f^{\text{RF}}(\epsilon)$$

$$z_i = f_i(\epsilon_i; z_{\mathbf{pa}_i}) = \exp(a_i) \cdot \epsilon_i + b_i$$

$$\log \prod_i \left|\frac{\partial \epsilon_i}{\partial z_i}\right| = \sum_i \log \left|\frac{\partial f_i^{\text{RF}}(\epsilon_i; \epsilon_{\mathbf{pa}_i})}{\partial \epsilon_i}\right|^{-1} = \sum_i a_i$$

$$\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{pmatrix} \mapsto \begin{pmatrix} f_1(\epsilon_1) \\ f_2(\epsilon_2) \\ f_3(\epsilon_3, z_1, z_2) \\ f_4(\epsilon_4, z_1, z_2) \end{pmatrix} = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{pmatrix}$$

Causal Disentanglement Prior

❖ Prior causally factorizes latent space and disentangles causal mechanisms

Bijective map $(z_{\mathbf{pa}_i}, u_i) \mapsto z_i$ *via autoregressive normalizing flow*

$$p_\theta(z|u) = \prod_{i=1}^{n} p_\theta(z_i | z_{\mathbf{pa}_i}, u_i) = \prod_{i=1}^{n} p(u_i) \left|\frac{\partial \boldsymbol{\lambda}_i(u_i; z_{\mathbf{pa}_i})}{\partial u_i}\right|^{-1}$$

$$p_\theta(z_i | z_{\mathbf{pa}_i}, u_i) = h_i(z_i) \exp(\mathbf{T}_i(z_i | z_{\mathbf{pa}_i}) \boldsymbol{\lambda}_i(\underbrace{G_i^z \odot}_{\text{causal mechanisms}} z, u_i) - \psi_i(z, u))$$

Learning Objective

❖ Maximize the following evidence lower bound (ELBO)

$$\log p_\theta(x, u) \geq \mathbb{E}_{\epsilon, z \sim q_\phi(\epsilon, z | x, u)} \Big[\log p_\theta(x|\epsilon) + \log p_\theta(x|z) - \beta\{\log q_\phi(\epsilon|x, u) + \log q_\phi(z|x, u) - \log p(\epsilon) - \log p_\theta(z|u)\} \Big]$$

Theorem (Identifiability of ICM-VAE)

1. The set $\{x \in X \mid \phi_\xi(x) = 0\}$ has measure zero
2. Decoder (mixing function) g is diffeomorphic onto its image
3. Sufficient statistics $\mathbf{T}_i$ are diffeomorphic
4. ***Sufficient Variability***: The conditional distribution depends sufficiently strongly on the derived parents $z_{\mathbf{pa}_i}$ and labels u_i

➔ θ and $\hat{\theta}$ are causal mechanism permutation equivalent (i.e., causal mechanisms are identified uniquely) and $\hat{\theta}$ is causally disentangled

Empirical Evaluation

❖ Experiments on Pendulum, Flow, and CausalCircuit image datasets with nonlinear ground-truth mechanisms and $n = 4$ continuous-valued causal factors

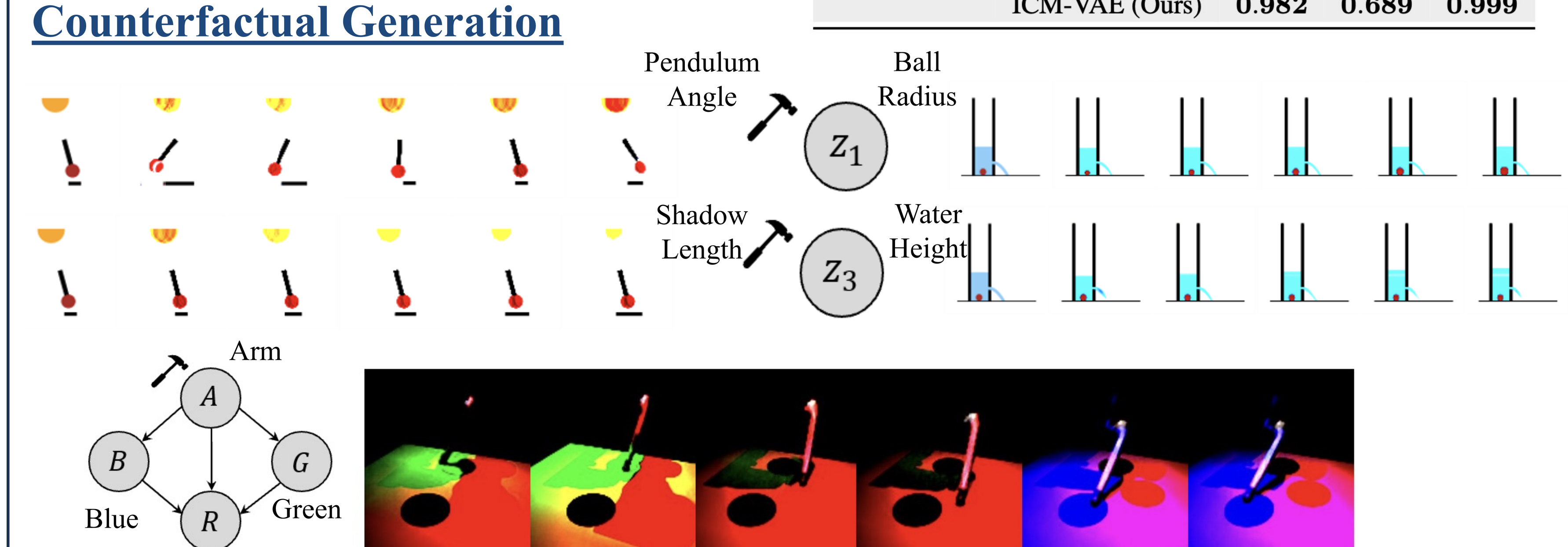
Causal Disentanglement

❖ High disentanglement (D), completeness (C), and interventional robustness (IRS) indicates causal mechanism disentanglement.

❖ ICM-VAE disentangles causal factors significantly better than other causal and acausal baselines.

Dataset	Model	D	C	IRS
Pendulum	β-VAE	0.182	0.285	0.449
	iVAE	0.483	0.385	0.670
	CausalVAE	0.885	0.539	0.817
	SCM-VAE	0.764	0.475	0.829
	ICM-VAE (Ours)	**0.997**	**0.882**	**0.869**
Flow	β-VAE	0.308	0.332	0.452
	iVAE	0.730	0.481	0.674
	CausalVAE	0.819	0.522	0.707
	SCM-VAE	0.854	0.483	0.811
	ICM-VAE (Ours)	**0.988**	**0.598**	**0.893**
CausalCircuit	β-VAE	0.692	0.442	0.982
	iVAE	0.745	0.541	0.992
	CausalVAE	0.886	0.625	0.994
	SCM-VAE	0.867	0.652	0.993
	ICM-VAE (Ours)	**0.982**	**0.689**	**0.999**

Counterfactual Generation



Acknowledgement: This work was supported in part by NSF 1910284, 1946391, 2147375, and NIH P20GM139768.

Contact Email: akomandu@uark.edu