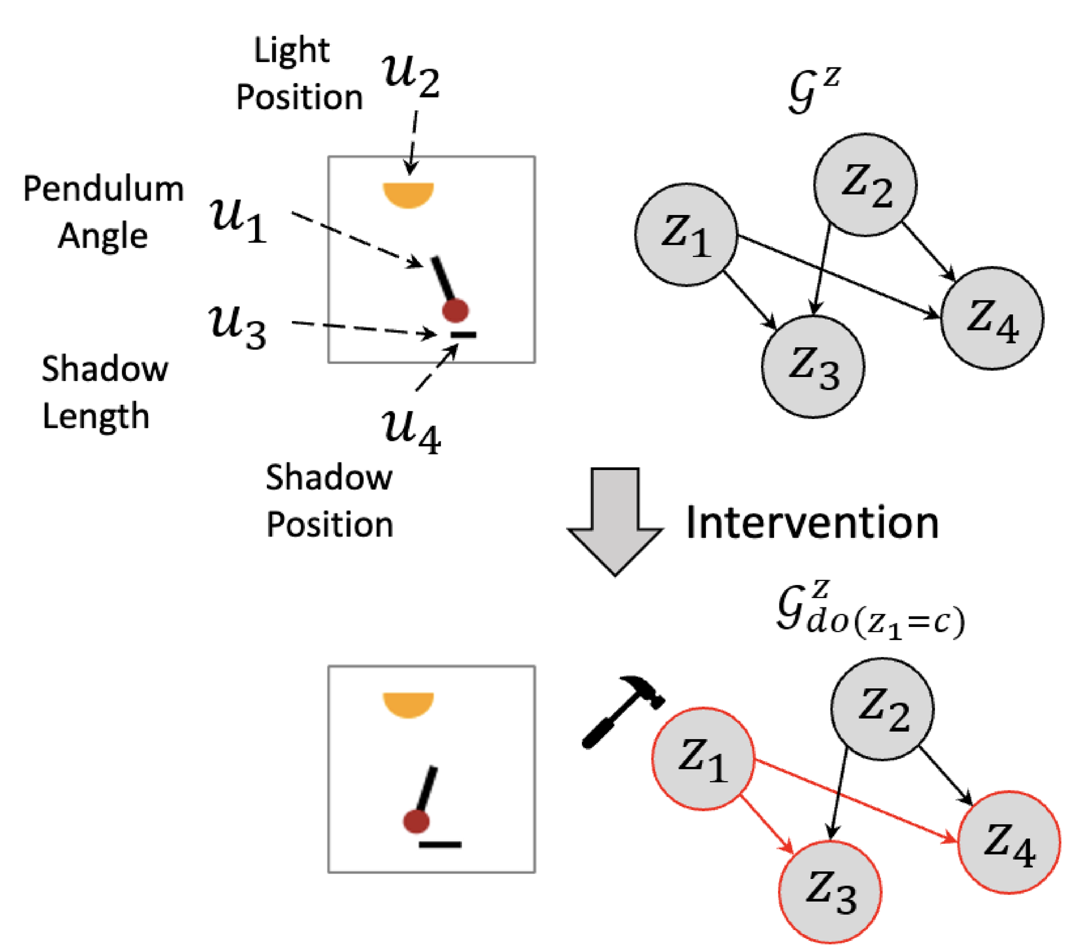
Learning Causally Disentangled Representations via the Principle of Independent Causal Mechanisms

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

[1]University of Arkansas, [2]Clemson University, [3]University of Texas at Dallas

Motivation

- We often observe high-dimensional data but desire to extract abstract causal variables and their structure.
- Disentangling causal factors is a challenging task and without any inductive bias, it is an impossible endeavor [1].
- Disentangled causal representations are useful for scheduling, planning, robustness to distribution shifts, and fairness in downstream applications.



Key Contributions

- We propose a reformulation of causal disentanglement from the perspective of independent causal mechanisms and generalize iVAE [2] to causally factorized distributions.
- We design a framework, ICM-VAE, for causal representation learning under supervision from labels.
- We theoretically show identifiability of causal mechanisms up to permutation and element-wise reparameterization.

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.

Acknowledgements

This work is supported in part by NSF 1910284, 1946391, 2147375 and NIH P20GM139768

Causal Mechanism Equivalence

<u>Violation of disentanglement of causal mechanisms from traditional disentanglement</u>

True SCM	Learned SCM	Marginals	
$z_1 = \epsilon_1 \sim \mathcal{N}(0,1)$	$\hat{z}_1 = \hat{\epsilon}_1 \sim \mathcal{N}(0,1)$	$z_1 \sim \mathcal{N}(0,1)$	$\hat{z}_1 \sim \mathcal{N}(0,1)$
$z_2 = \epsilon_2 \sim \mathcal{N}(0,1)$	$\hat{z}_2 = \hat{\epsilon}_2 \sim \mathcal{N}(0,1)$	$z_2 \sim \mathcal{N}(0,1)$	$\hat{z}_2 \sim \mathcal{N}(0,1)$
$z_3 = az_1 + bz_2 + \epsilon_3$	$\hat{z}_3 = b\hat{z}_1 + a\hat{z}_2 + \hat{\epsilon}_3$	$z_3 \sim \mathcal{N}(0, \sqrt{a^2+b^2+1^2})$	$\hat{z}_3 \sim \mathcal{N}(0, \sqrt{b^2+a^2+1^2})$
$z_4 = cz_1 + dz_2 + \epsilon_4$	$\hat{z}_4 = d\hat{z}_1 + c\hat{z}_2 + \hat{\epsilon}_4$	$z_4 \sim \mathcal{N}(0, \sqrt{c^2+d^2+1^2})$	$\hat{z}_4 \sim \mathcal{N}(0, \sqrt{d^2+c^2+1^2})$

<u>Issue</u>: Learned mechanisms may be different than true underlying mechanisms but produce same marginal.

<u>Idea</u>: What if we consider disentanglement from a causal mechanism perspective? $p_\theta(z_i|z_{\mathbf{pa}_i}) = p_{\hat{\theta}}(z_i|z_{\mathbf{pa}_i})$

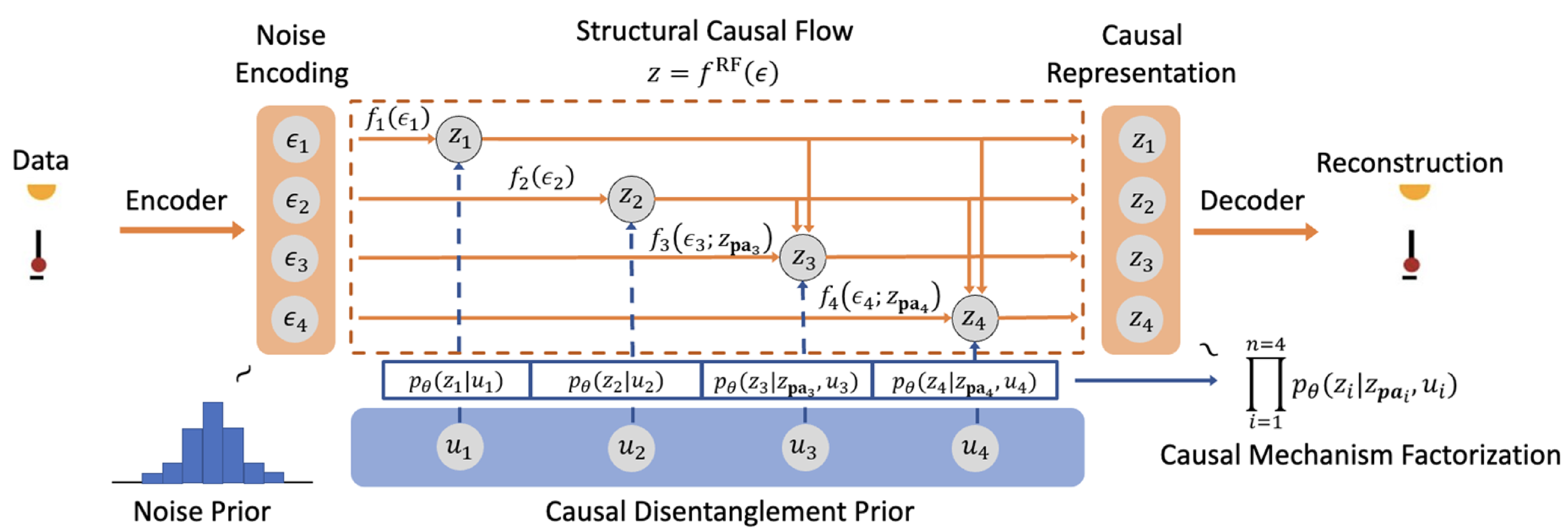
Three *sufficient* conditions for causal mechanism equivalence:

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

recover mechanisms up to permutation

$\Rightarrow$ Causally Disentangled and Causal Mechanism Permutation Equivalent

ICM-VAE Framework



<u>**Structural Causal Flow**</u>

- Parameterize causal mechanisms as nonlinear **diffeomorphic** functions via autoregressive normalizing flows

$$z_i = f_i(\epsilon_i; z_{\mathbf{pa}_i}) = \exp(a_i)\cdot\epsilon_i + b_i \qquad \begin{pmatrix}\epsilon_1\\ \epsilon_2\\ \epsilon_3\\ \epsilon_4\end{pmatrix} \mapsto \begin{pmatrix} f_1(\epsilon_1)\\ f_2(\epsilon_2)\\ f_3(\epsilon_3, z_1, z_2)\\ f_4(\epsilon_4, z_1, z_2)\end{pmatrix} = \begin{pmatrix} z_1\\ z_2\\ z_3\\ z_4\end{pmatrix}$$

<u>**Causal Disentanglement Prior**</u>

- Prior exponential family distribution to causally factorize the latent space and disentangle causal mechanisms

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

mechanisms

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

Empirical Evaluation

Experiments on Pendulum, Flow, and CausalCircuit image datasets with *nonlinear ground-truth mechanisms* and *four continuous-valued causal factors*

<u>**Causal Disentanglement**</u>

- 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**

<u>**Counterfactual Generation**</u>