Interventional Causal Representation Learning

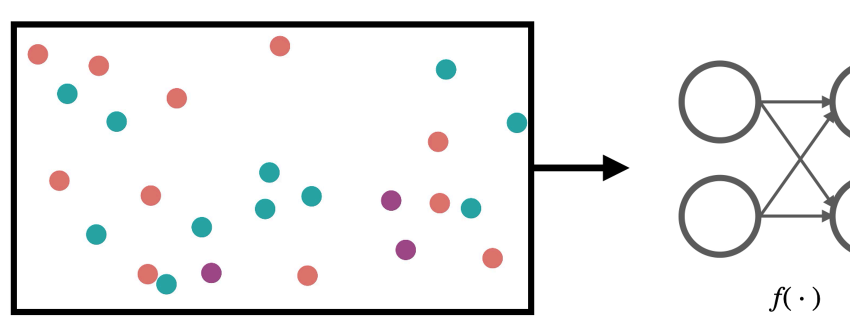
Kartik Ahuja, Divyat Mahajan, Yixin Wang, Yoshua Bengio

FAIR (Meta AI), Mila-Quebec AI Institute, University of Michigan

Motivation

- Current AI systems are still limited in terms of planning and reasoning
- Humans plan & reason using abstract concepts (e.g. objects & their properties)
- Causal models present a natural framework to represent such abstract concepts — latent causal variables and reason about interventions on them
- How to train representation learners that extract causal variables from high dimensional data (e.g., images) with minimal supervision?

Input



Desired output

Causal variables
- location
- shape
- color

Problem Statement

- True latent variables:
  Observational distribution: $z \sim \mathbb{P}_Z$ with support $\mathcal{Z}$
  Interventional distribution: $z \sim \mathbb{P}_Z^{(i)}$ with support $\mathcal{Z}^{(i)}$
- Mixing function: $g : \mathbb{R}^d \rightarrow \mathbb{R}^n$, which is injective
- Observations: $x \leftarrow g(z)$ with supports $\mathcal{X}, \mathcal{X}^{(i)}$ in observational and interventional distribution
- Learn an auto encoder:
  Reconstruction identity: $h \circ f(x) = x, \forall x \in \mathcal{X} \cup \mathcal{X}^{(i)}$
  $\hat{z} \triangleq f(x)$
- **Affine Identification**: $\hat{z} = Az + c$
- **Permutation and scaling Identification**: $\hat{z} = \Pi\Lambda z + c$

Identification under Hard do Interventions

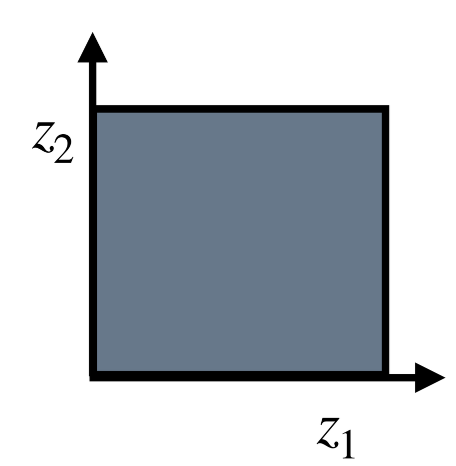
- **Assumption 1**: g is an injective polynomial, $\mathcal{Z}$ *has a non-empty interior*
- **Assumption 2:** $\mathbb{P}_Z^{(i)}$ hard do intervention on z_i
- **Do intervention constraint**: $f_k(x) = z^{\dagger}, \forall x \in \mathcal{X}^{(i)}$

Theorem (Informal): If Assumption 1 and 2 hold, then t*he solution to the reconstruction identity with* h *as a polynomial and do intervention constraint satisfies*

$$\hat{z}_k = ez_i + b, \ \forall z \in \mathcal{Z} \cup \mathcal{Z}^{(i)}$$

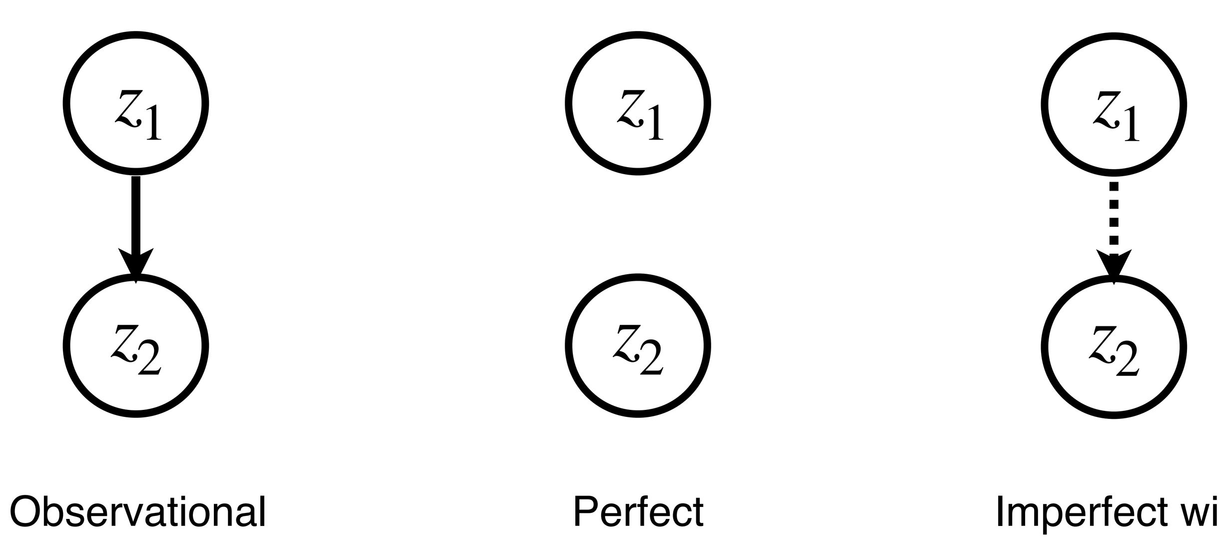
- For general diffeomorphisms, we show approximate component-wise identification under multiple do interventions per component

Independent Support & Imperfect Interventions



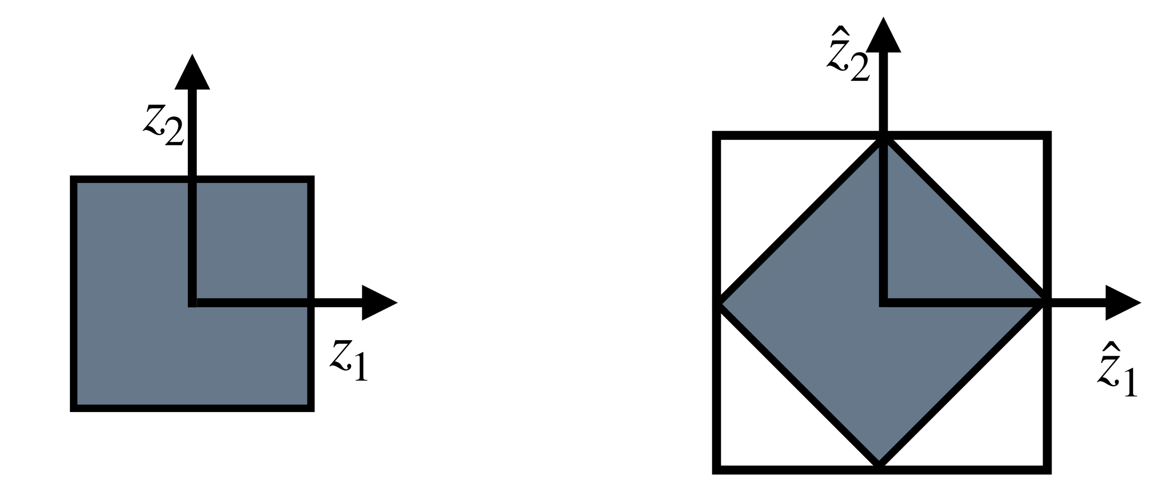
Independent Support (IS): $\mathcal{Z}_{12} = \mathcal{Z}_1 \times \mathcal{Z}_2$

Statistical Independence $\Longrightarrow$ IS



Identification under Imperfect Interventions

- **Assumption 3:** For $\mathbb{P}_Z^{(i)}$ $\exists$ $\mathcal{S}$, support of z_i is independent of latents in $\mathcal{S}$
- **IS constraint**: For a set $\mathcal{S}'$ support of $\hat{z}_k$ is independent of latents in $\mathcal{S}'$

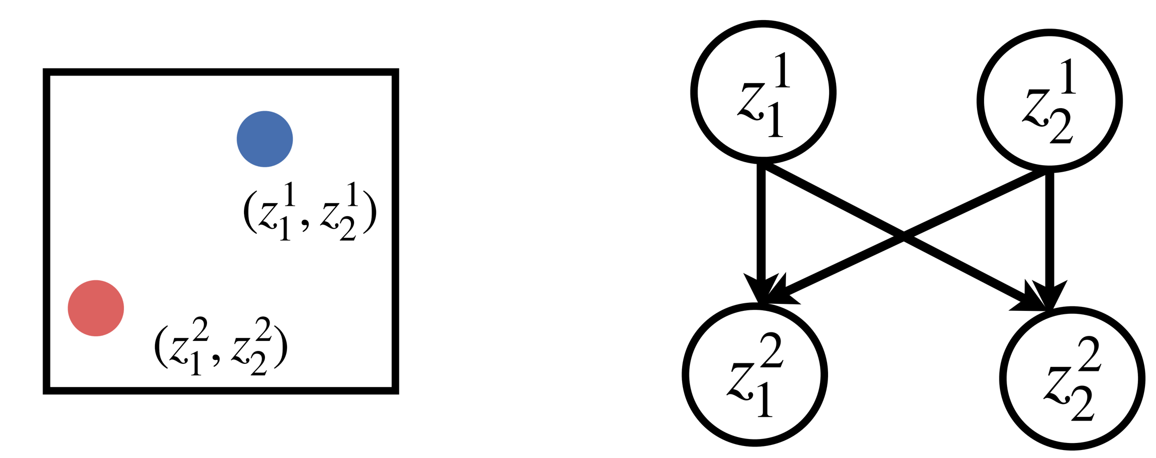


Theorem (Informal): If Assumption 1 and 3 hold, then t*he solution to the reconstruction identity with* h *as a polynomial and support independence constraint achieves block-affine identification*

$$\hat{z}_k = a_k^{\top} z + c_k, \ \hat{z}_m = a_m^{\top} z + c_m, \forall m \in \mathcal{S}'$$

a_k *and* a_m *do not share non-zero components.*

Experiments



# Interv Distbn	Uniform	SCM-Linear	SCM Non-Linear
1	33.2	42.7	34.9
3	72.2	73.9	65.2
5	88.3	83.6	77.2
7	88.1	85.5	81.9
9	87.5	84.8	81.1