Interventional Causal Representation Learning
01 Motivation
Current AI systems are still limited in terms of planning and reasoning abilities

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 raw data (e.g., images) with minimal supervision?
02 Problem Setting
True latent variables:

Observational distribution: $z \sim P_Z$ with support $\mathcal{Z}$

Interventional distribution: $z \sim P_{Z}^{(i)}$ 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: \[ \hat{z} = \Pi \Lambda z + c \]
Prior Work:
Parametric assumptions of latent distribution

Independent Component Analysis (ICA):
Latent are independent and non-gaussian

Non-Linear ICA:
Latent are conditionally independent given auxiliary variables (Hyvärinen et al.)
Weak supervision with contrastive pairs \((x, \tilde{x})\)
(Brehmer et al.; Ahuja et al.)
03 Identification under Hard do Interventions
• **Assumption 1**: $g$ is an injective polynomial

• **Assumption 2**: $P^{(i)}_Z$ is hard do-intervention on $z_i$, $\mathcal{X} \cup \mathcal{X}^{(i)}$ has a non-empty interior

• Do intervention constraint: $f_k(x) = z^+, \forall x \in \mathcal{X}^{(i)}$

---

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

$$\hat{z}_k = e z_i + b, \forall z \in \mathcal{X} \cup \mathcal{X}^{(i)}$$
• **Assumption 1**: $g$ is an injective polynomial

• **Assumption 2**: $\mathbb{P}^{(i)}_{Z}$ is hard do-intervention on $z_i$ and multiple such interventions

• Approximate identification of the intervened component
04 Identification under Imperfect Interventions
Independent Support (IS): $\mathcal{X}_{12} = \mathcal{X}_1 \times \mathcal{X}_2$

Statistical Independence $\implies$ IS
Geometric Intuition:

\((\hat{z}_1, \hat{z}_2)\) is a transformation over \((z_1, z_2)\) such that we do not have identification upto permutation & scaling.

We lose IS property with such transformations; the only to preserve IS is to have transformations that recover latents unto permutation & scaling.
• **Assumption 3:** For $\mathbb{P}_{Z}^{(i)} \ni S$ s.t. support of $z_i$ is independent of other latents in $S$

• **IS constraint:** For a set $S'$ support of $\hat{z}_k$ is independent of other latents in $S'$

---

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

$$\hat{z}_k = a_k^T z + c_k, \ \hat{z}_m = a_m^T z + c_m, \ \forall m \in S'$$

$a_k$ and $a_m$ do not share non-zero components.
Thank you