Towards Efficient Identification in Supervised Learning

Kartik Ahuja*, Divyat Mahajan*, Vasilis Syrgkanis, Ioannis Mitliagkas
Proceedings of CleaR 2022

Background

Linear Independent Component Analysis

$X \leftarrow GZ$

Theorem [Darmois]:

Define $W_1 = \sum_{k=1}^{d} a_{1k} V_k,\ W_2 = \sum_{k=1}^{d} a_{2k} V_k.$

If W_1, W_2 are independent, all components of V are mutually independent, and $a_{1i} a_{2i} \neq 0$, then V_i is Gaussian.

Linear ICA

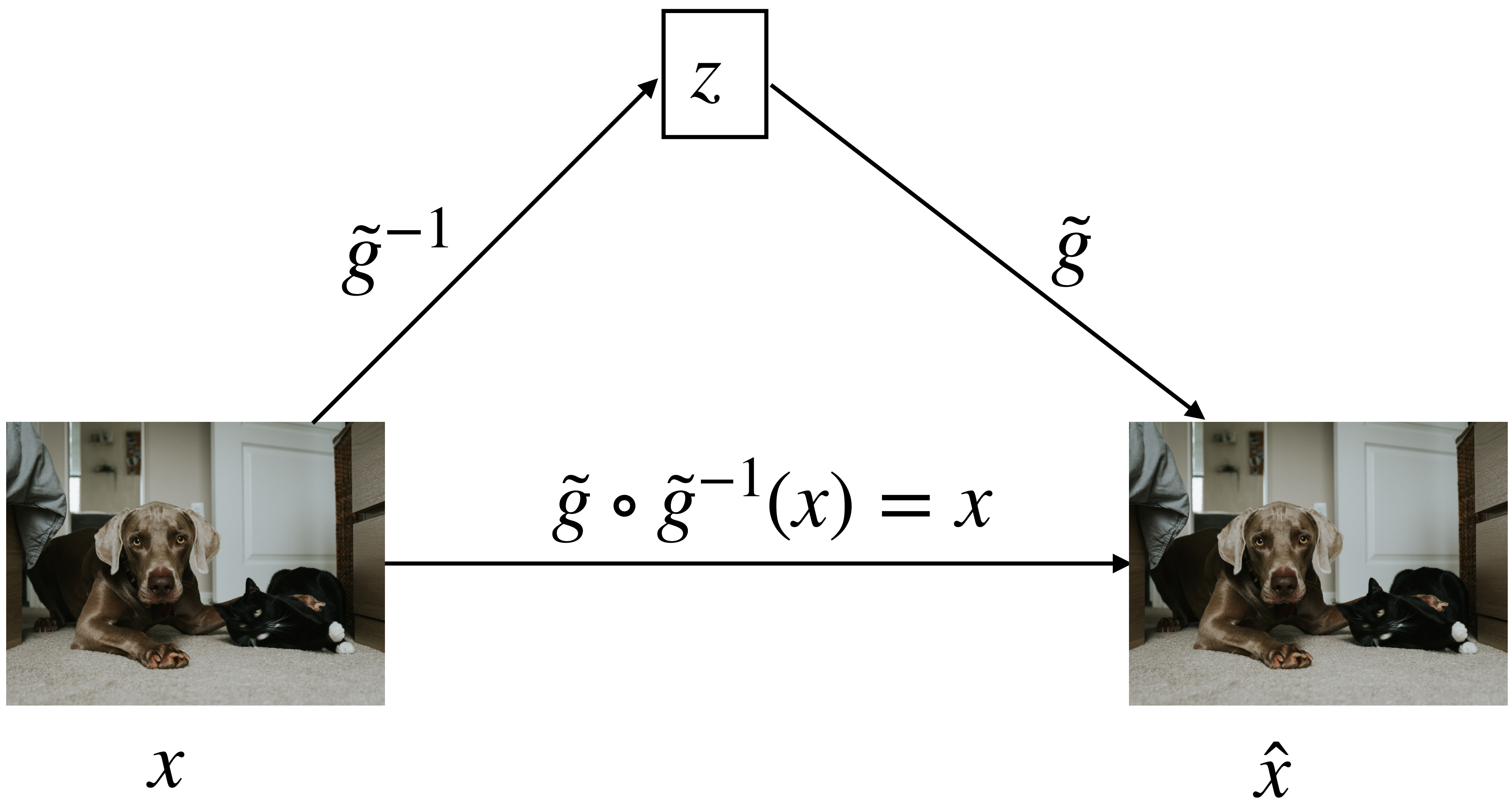
$X \leftarrow GZ$

Theorem [Comon]

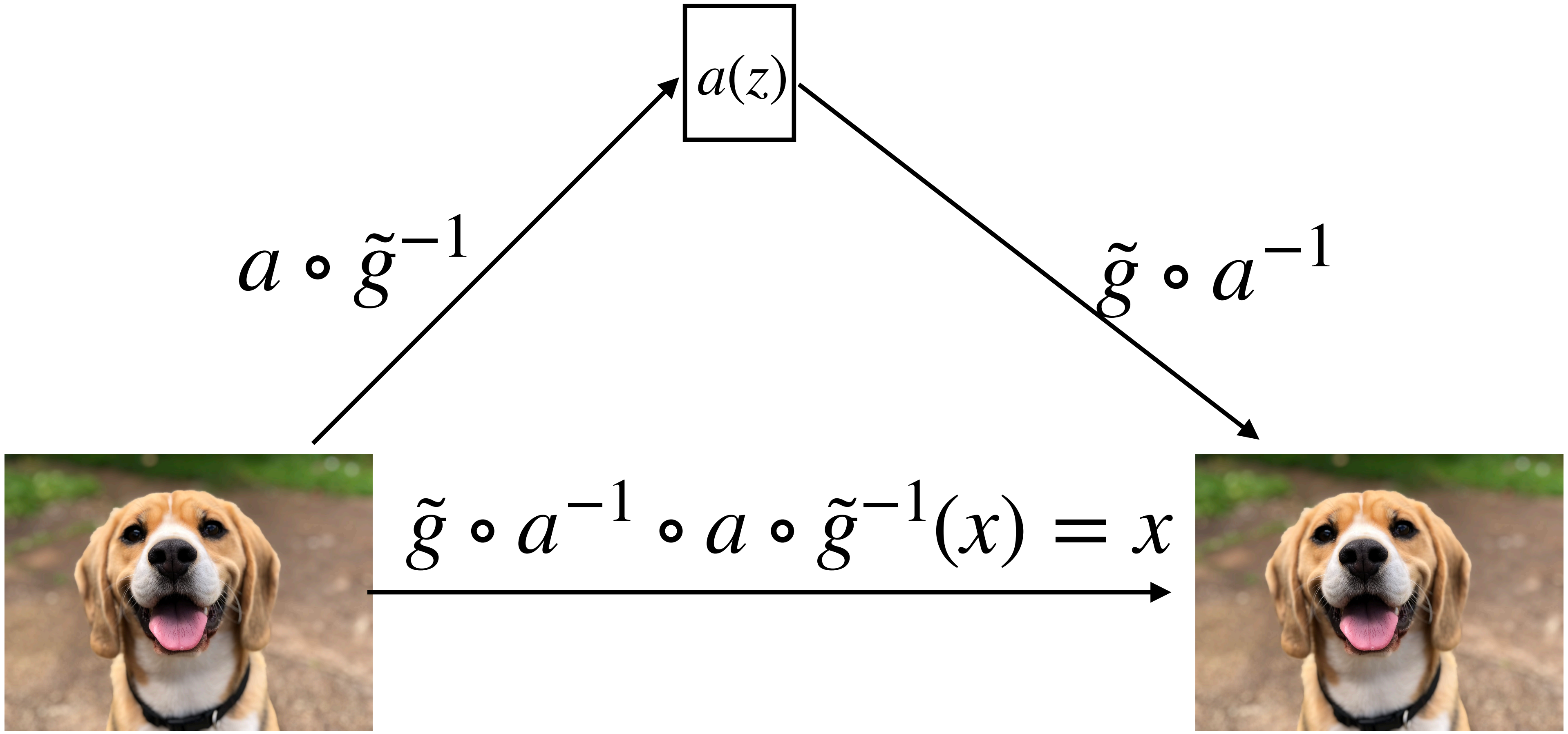
If at least one component of Z is non-Gaussian, then it is possible to recover Z up to permutation and scaling, i.e., $\hat{G}X = \hat{Z} = \Pi\Lambda Z$, where Π is permutation matrix and Λ is a diagonal matrix.

Non-identification in Autoencoders

Data generation process: $X \leftarrow g(Z)$



Non-identification in Autoencoders

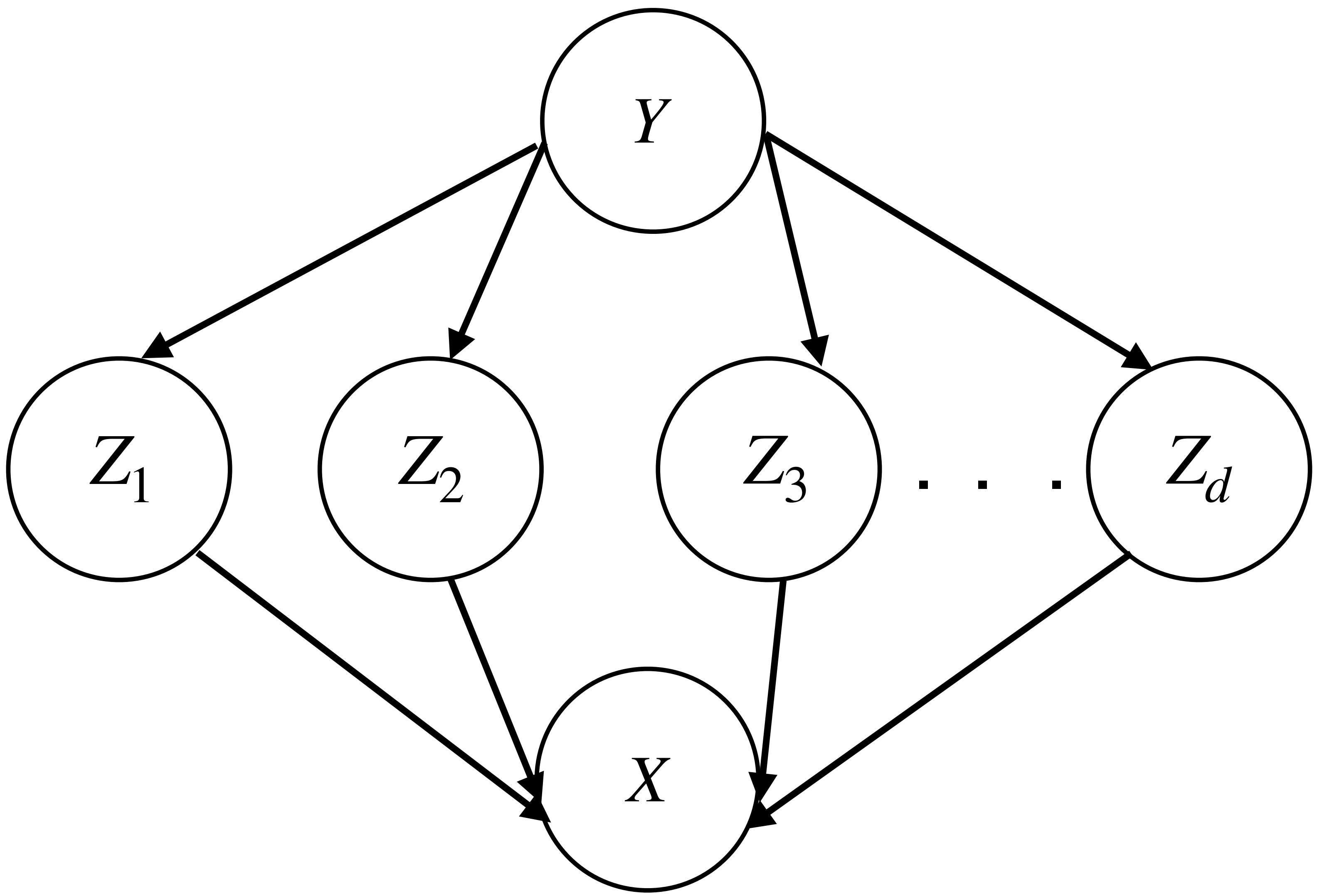


- Identification without assumptions on DGP impossible [Hyvarinen et al.]
- Existing works make assumptions on independence structure of latents

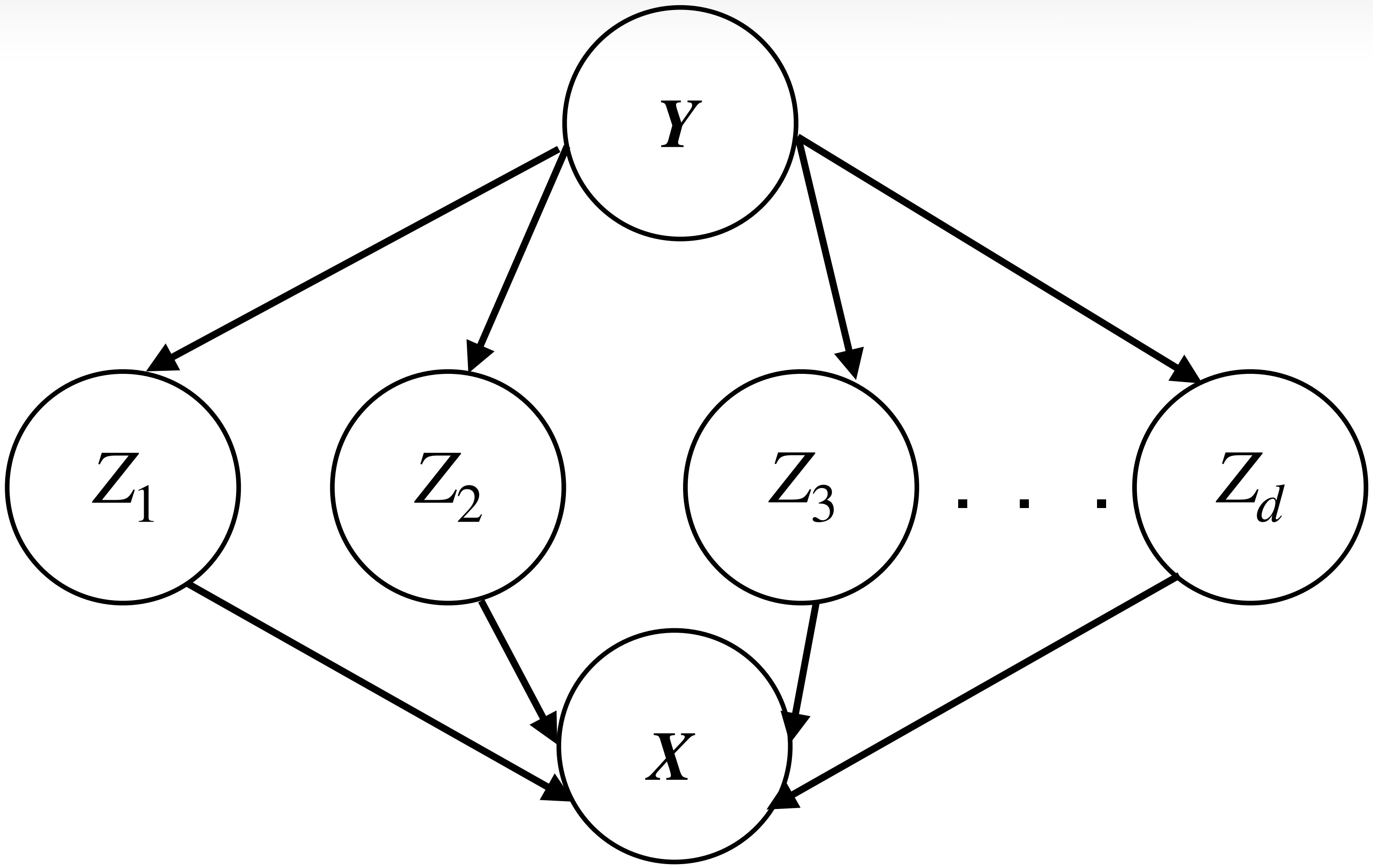
Non-Linear ICA

Auxiliary information (Labels) cause latents (e.g., Handwritten digits)

$$Z \leftarrow \mu_Y + N_Y$$
$$X \leftarrow g(Z)$$



Non-Linear ICA



Assumption: All components of Z are independent conditional on Y
Theorem [Khemakhem et al.]:

i) Number of label classes twice the latent dimension

ii) Mean and noise in latent generation satisfies sufficient variability implies

Permutation recovery of the latents

Limitations of existing works

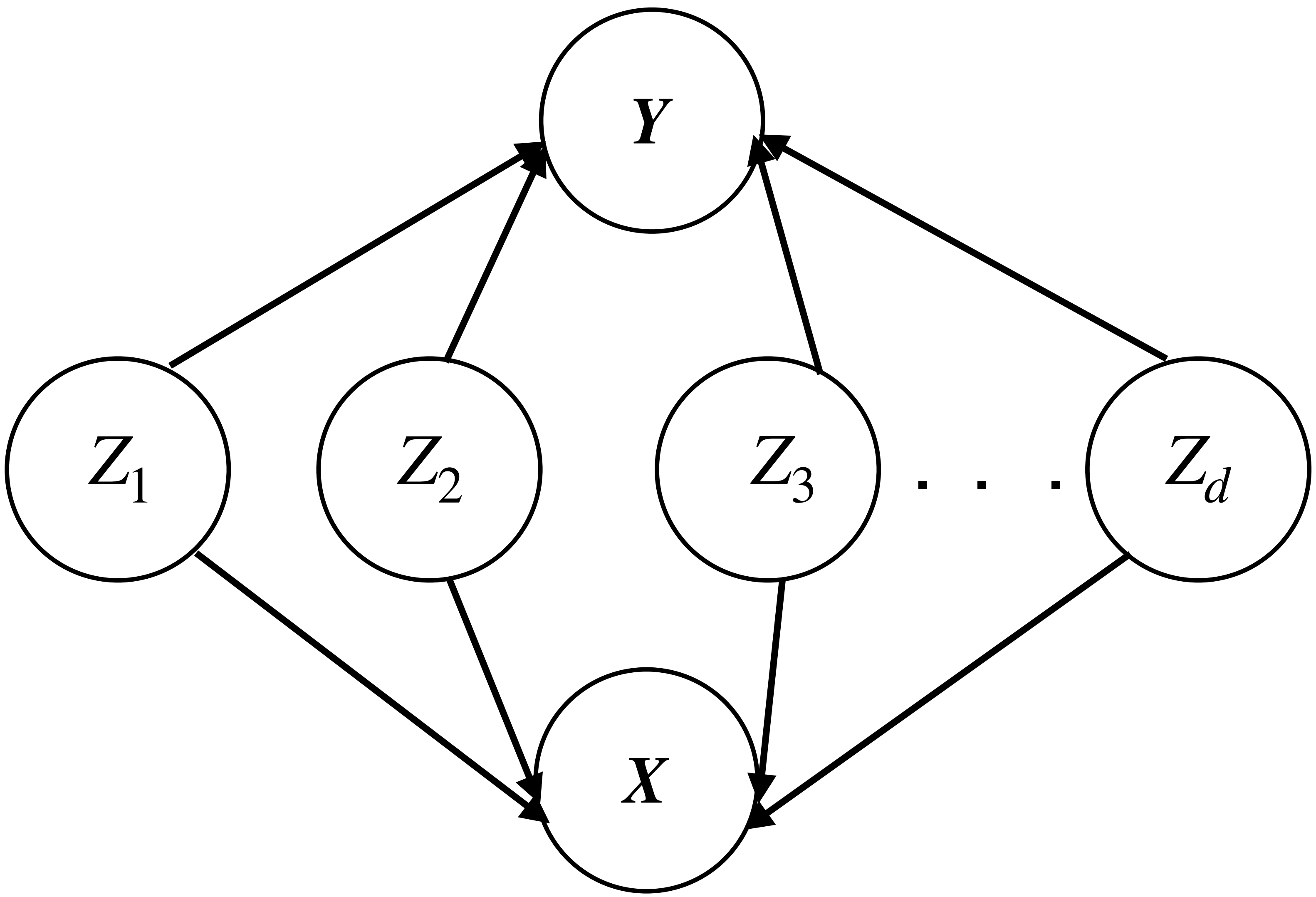
Existing works in non-linear ICA can rely on unrealistic assumptions

- Labels do not often cause latents (most human labelled datasets)
- Too much auxiliary information needed to recover the latent

Problem Setting

Data Generation: Latents Cause Labels

Latents cause labels (e.g., human labelled datasets)



Multi-task regression

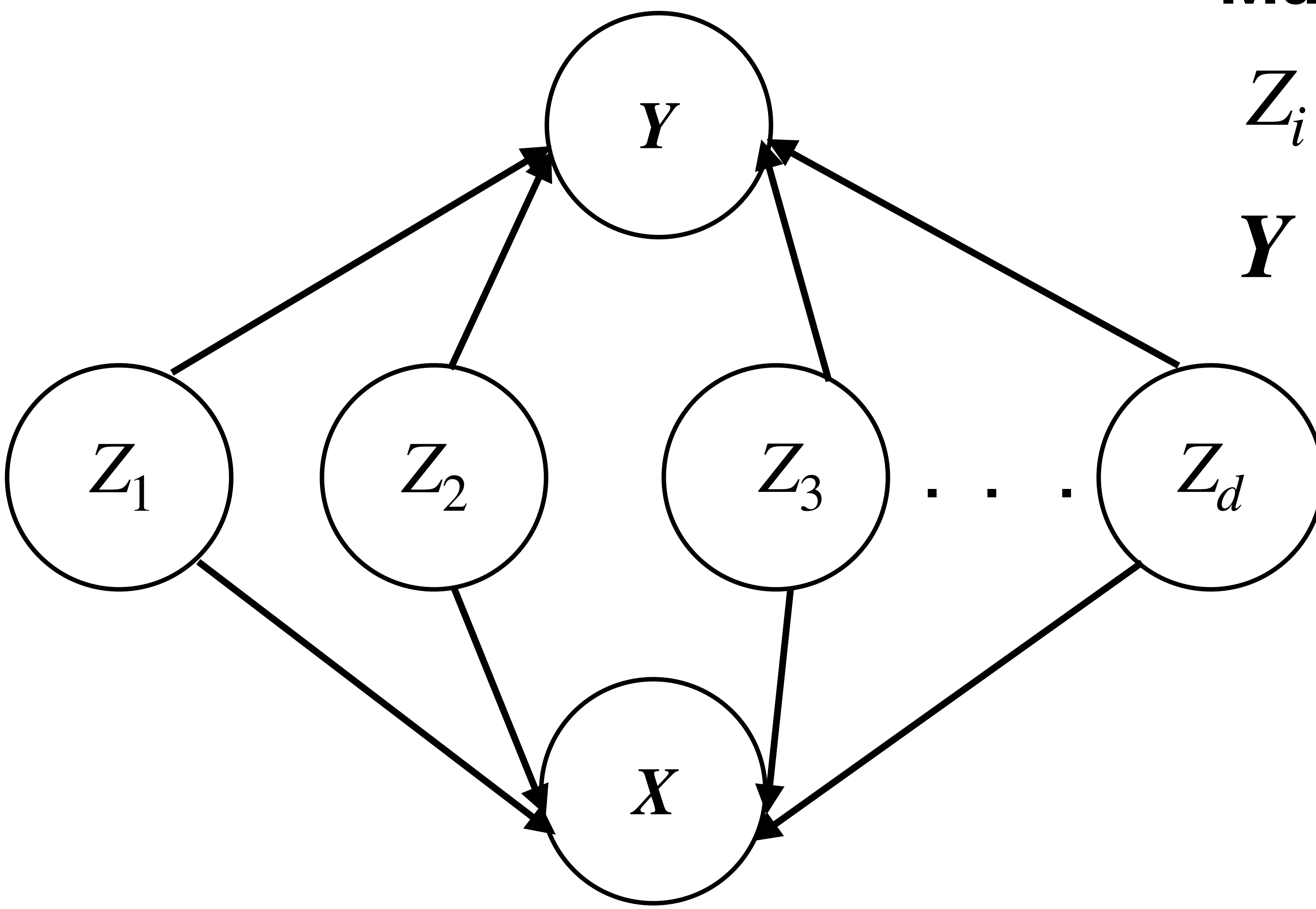
$$Z_i \leftarrow h_i(U_i), \forall i \in \{1, \cdots, d\}$$

$$\boldsymbol{Y} \leftarrow \boldsymbol{\Gamma}\mathbf{Z} + \boldsymbol{N}$$

$$\boldsymbol{X} \leftarrow g(\mathbf{Z})$$

Data Generation: Latents Cause Labels

Latents cause labels (e.g., human labelled datasets)



Multi-task classification

$$Z_i \leftarrow h_i(U_i), \forall i \in \{1, \cdots, d\}$$

$$\boldsymbol{Y} \leftarrow \text{Bernoulli}\Big(\sigma\Big(\boldsymbol{\Gamma}\mathbf{Z}\Big)\Big)$$

$$\boldsymbol{X} \leftarrow g(\mathbf{Z})$$

Key Identification Results

Empirical Risk Minimization

Model:

$W \circ \Phi - W$: Linear model, $\Phi \in \mathcal{H}_{\Phi}$: Non-linear representation

ERM: $$\min_{W \in \mathbb{R}^{d \times k}, \Phi \in \mathcal{H}_{\Phi}} \sum_{i=1}^{N} \ell\Big(W \circ \Phi(X_i), Y_i\Big)$$

Independence Constrained ERM

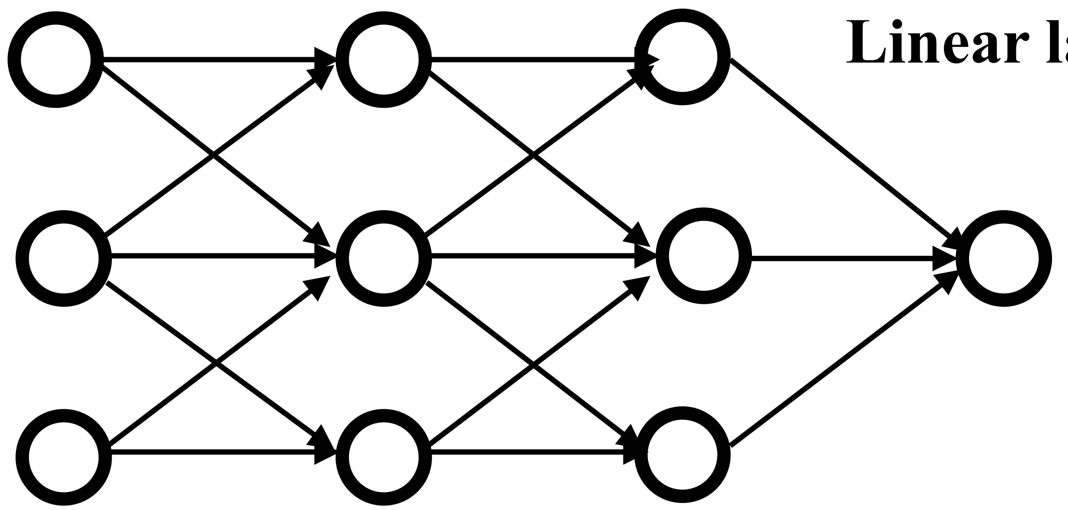
Independence-constrained ERM:

$$\min_{\Theta,\Phi} \sum_{i=1}^{N} \ell\Big(W \circ \Phi(X_i), Y_i\Big) \text{ **s.t. Components of** } \Phi(X) \text{ **are i.i.d.**}$$

ERM vs IC-ERM

Representation network

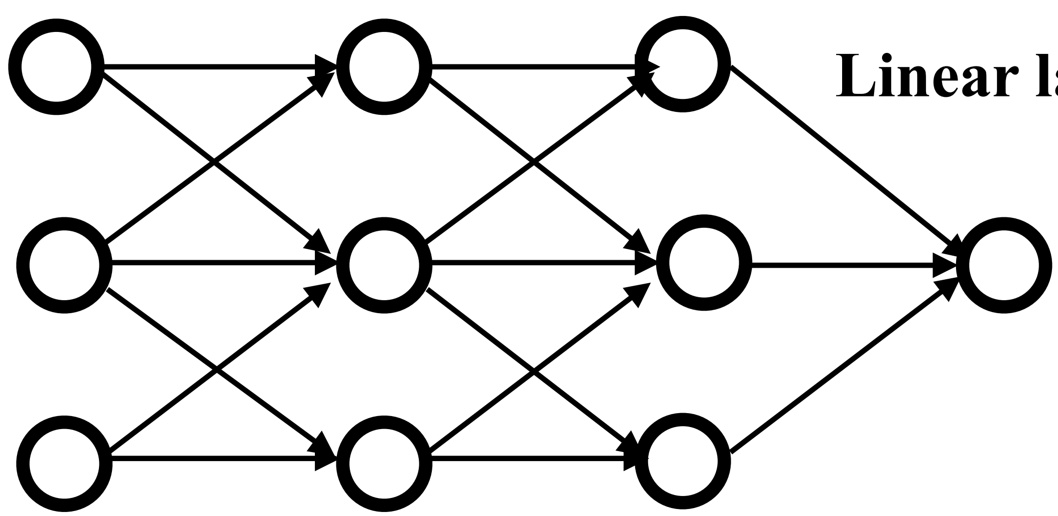
Linear layer



$$\min_{W \in \mathbb{R}^{d \times k}, \Phi \in \mathscr{H}_\Phi} \sum_{i=1}^{N} \ell\Big(W \circ \Phi(X_i), Y_i\Big)$$

ERM

Linear layer



$$\min_{W \in \mathbb{R}^{d \times k}, \Phi \in \mathscr{H}_\Phi} \sum_{i=1}^{N} \ell\Big(W \circ \Phi(X_i), Y_i\Big)$$

$\Phi(X)$ **is i.i.d.**

Representation network

IC-ERM

Inverting Latents Using IC-ERM

Assumption: Number of tasks is equal to the dimension of the latent

Theorem [Ahuja et al.]:
If number of tasks is equal to the latent dimension and $g^{-1} \in \mathscr{H}_{\Phi}$ then representation learned by

a) IC-ERM identifies true latent up to permutation & scaling
b) ERM identifies true latent up to linear transformation

Other Implications

- Recover the ground-truth latent variable values up to permutations and scaling
- If two neural nets (with same architecture and trained ERM on same data) output the same logits, then their representations are linearly related
- If two neural nets (with same architecture and trained with IC-ERM on same data) output the same logits, then their representations are permutations and scaling of each other

Relaxing Assumption on Number of Tasks

Inverting Latents For Single Task

Assumption:

i) Number of tasks is equal to one

ii) **Exponential distribution that follows** $\log p(Z) = \sum_{i=1}^{p} a_i z^i$

Theorem (Informal) [Ahuja et al.]:

If the latent is from exponential family above and the degree of the polynomial p is sufficiently large, then the IC-ERM identifies the latents up to permutation

Proposed Approach

ERM + Linear ICA

- Extract the representation learned by ERM $\Phi(X)$
- Process $\Phi(X)$ using linear ICA

Theorem [Ahuja et al.]:
If number of tasks is equal to the dimension of the latents and $g^{-1} \in \mathscr{H}_{\Phi}$ then the representation learned by ERM + Linear ICA identifies true latent up to permutation and scaling

Experiments

Experiments

Data Generation

$$\boldsymbol{X} \leftarrow g(\mathbf{Z})$$

Multi-task regression

$$\boldsymbol{Y} \leftarrow \mathbf{\Gamma Z} + \boldsymbol{N}$$

Multi-task classification

$$\boldsymbol{Y} \leftarrow \text{Bernoulli}\left(\sigma\left(\mathbf{\Gamma Z}\right)\right)$$

Experiments

Methods

- ERM
- ERM-PCA
- ERM-ICA

Metrics

- **Prediction performance:** R^2, Accuracy
- **Representation quality**: Mean correlation coefficient

Experiments

Multi-task regression

Figure 3: Comparison of label and latent prediction performance (regression, $d = 50$).

Experiments

Multi-task classification

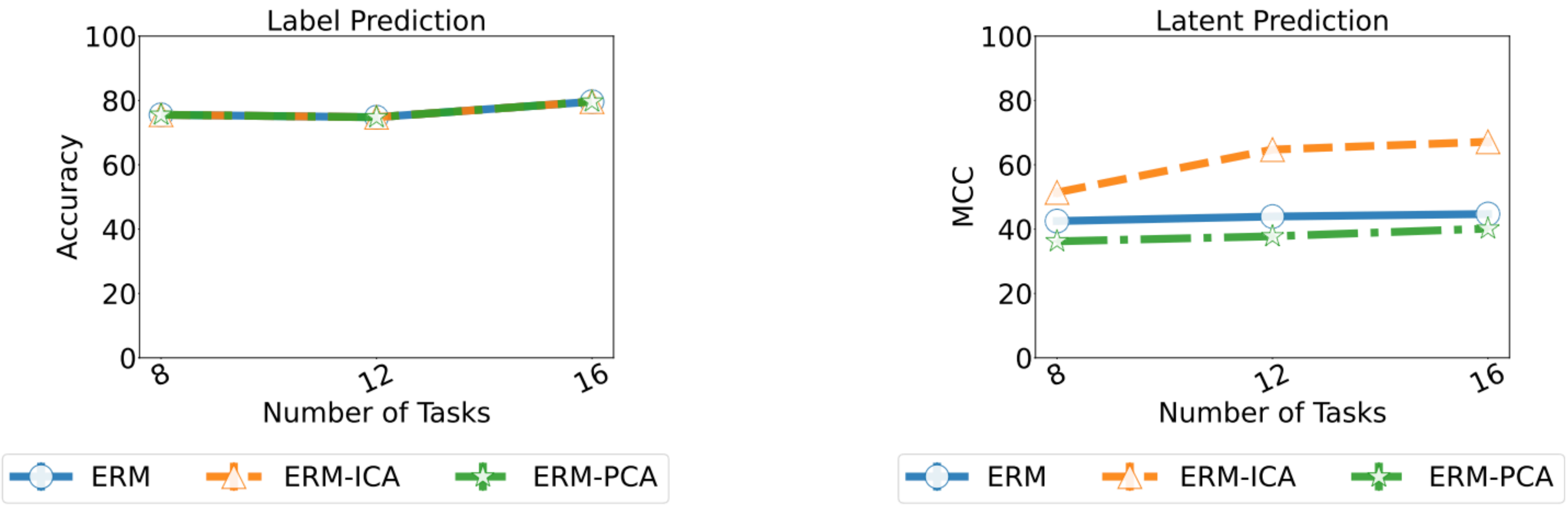


Figure 4: Comparison of label and latent prediction performance (classification, $d = 16$)

Thank You!