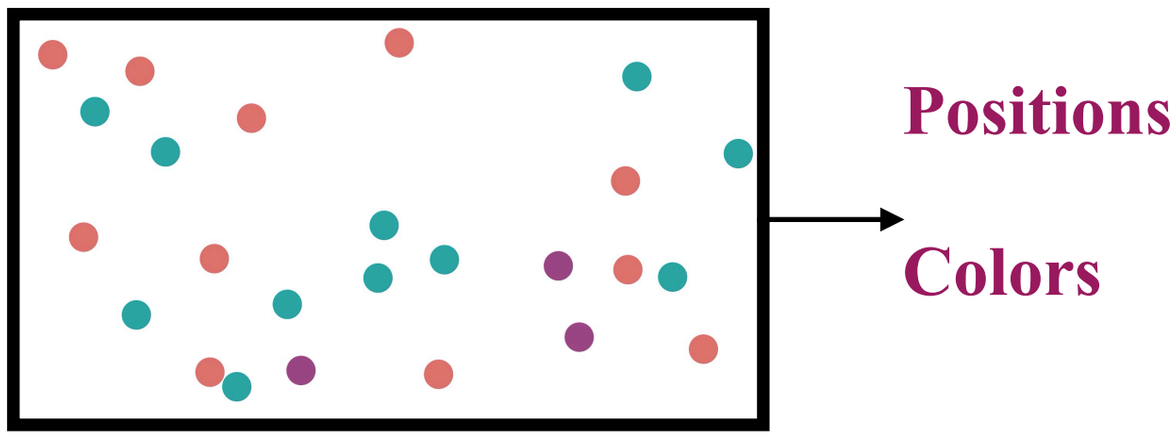
Multi-Task Causal Representation Learning

Motivation

Reasoning

Learn high-level latents from complex data

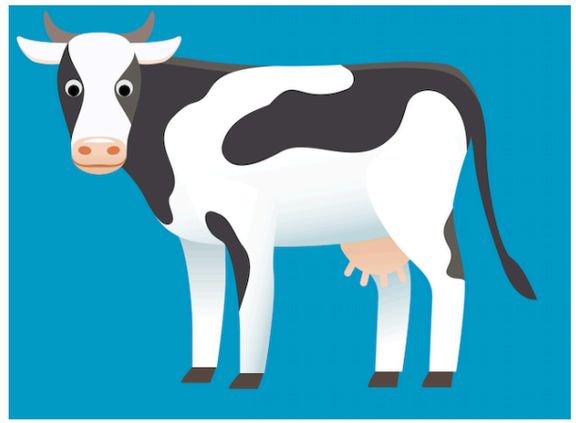


$$X \leftarrow g(Z)$$

$$Z' \leftarrow m(Z)$$

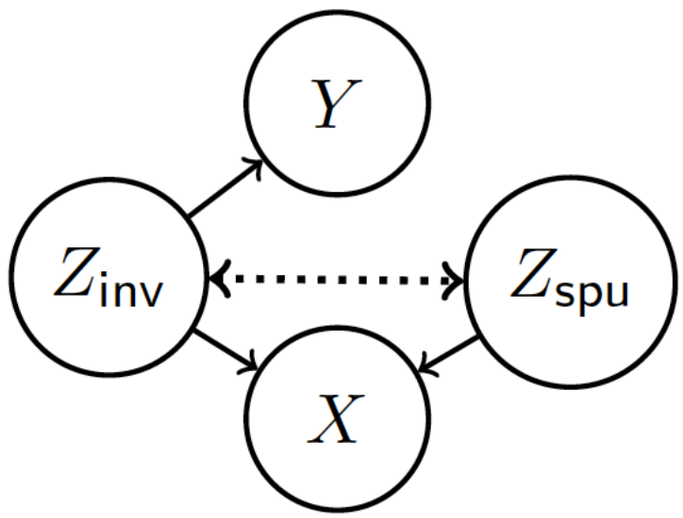
Distribution Shifts

Cow features



Background features

$$X \leftarrow g(Z)$$



Research Questions

i) How to invert ground-truth representations Z from observations X while making few but realistic assumptions on Z?

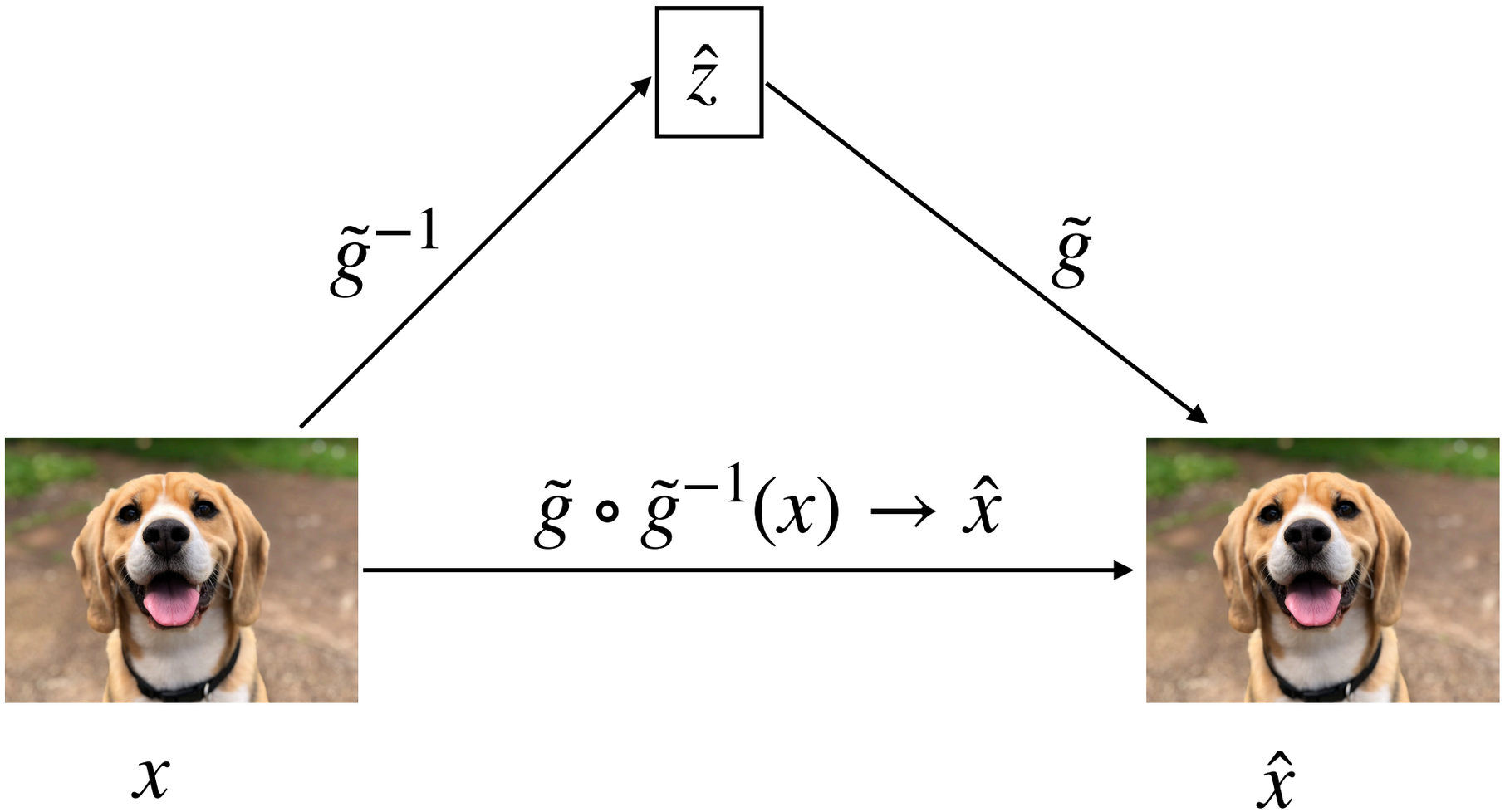
ii) How to leverage these representations for downstream prediction tasks?

Background

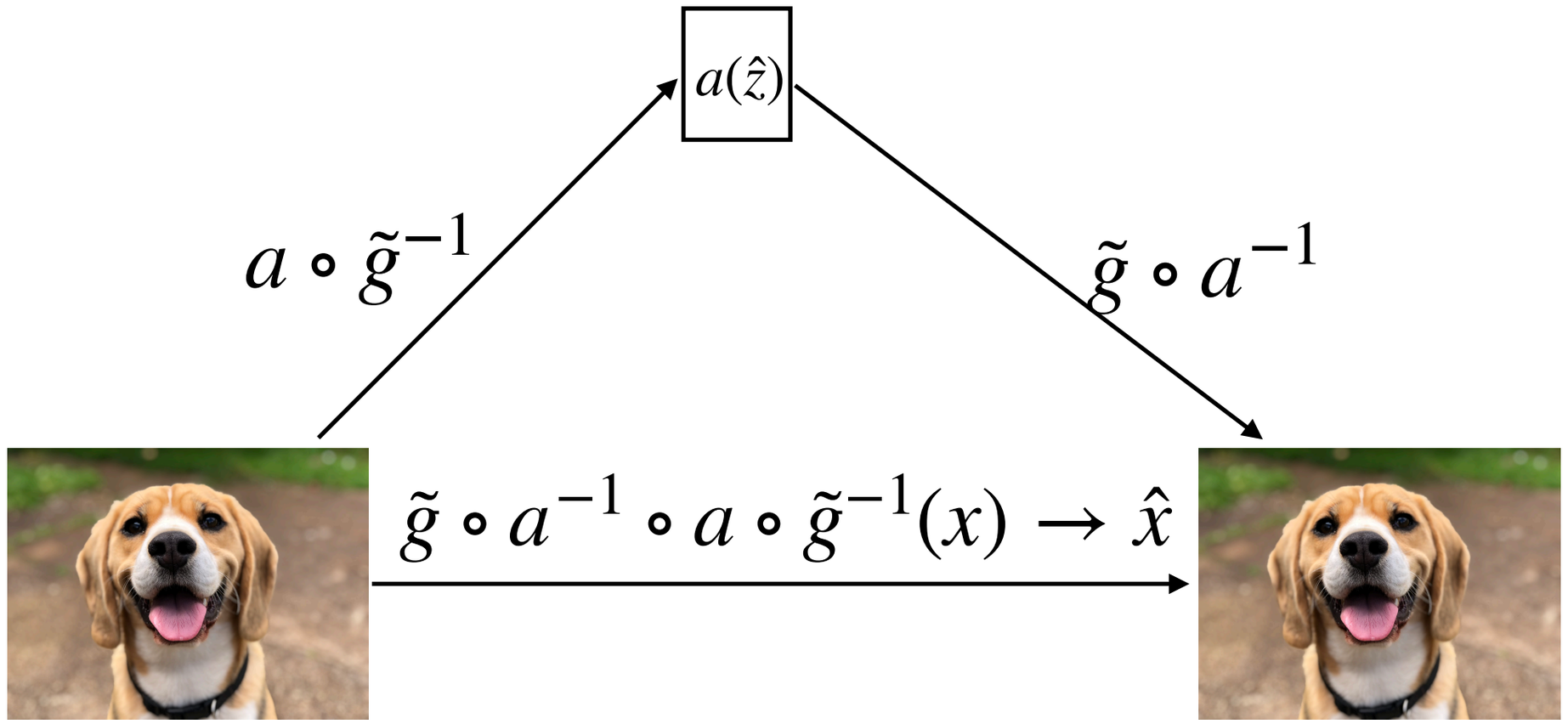
Non-identification in Autoencoders

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

Learned model: $\hat{X} \leftarrow \tilde{g} \circ \tilde{g}^{-1}(X)$

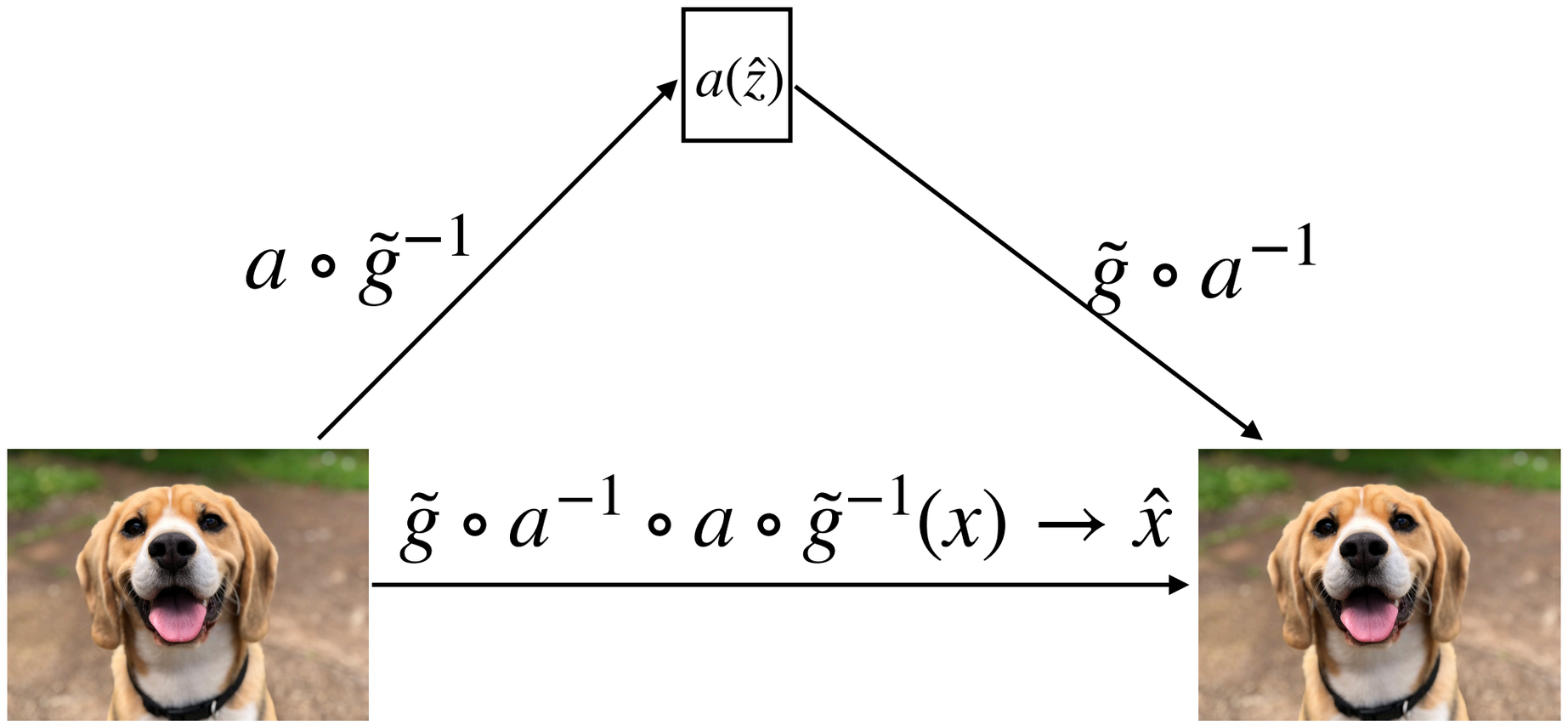


Non-identification in Autoencoders



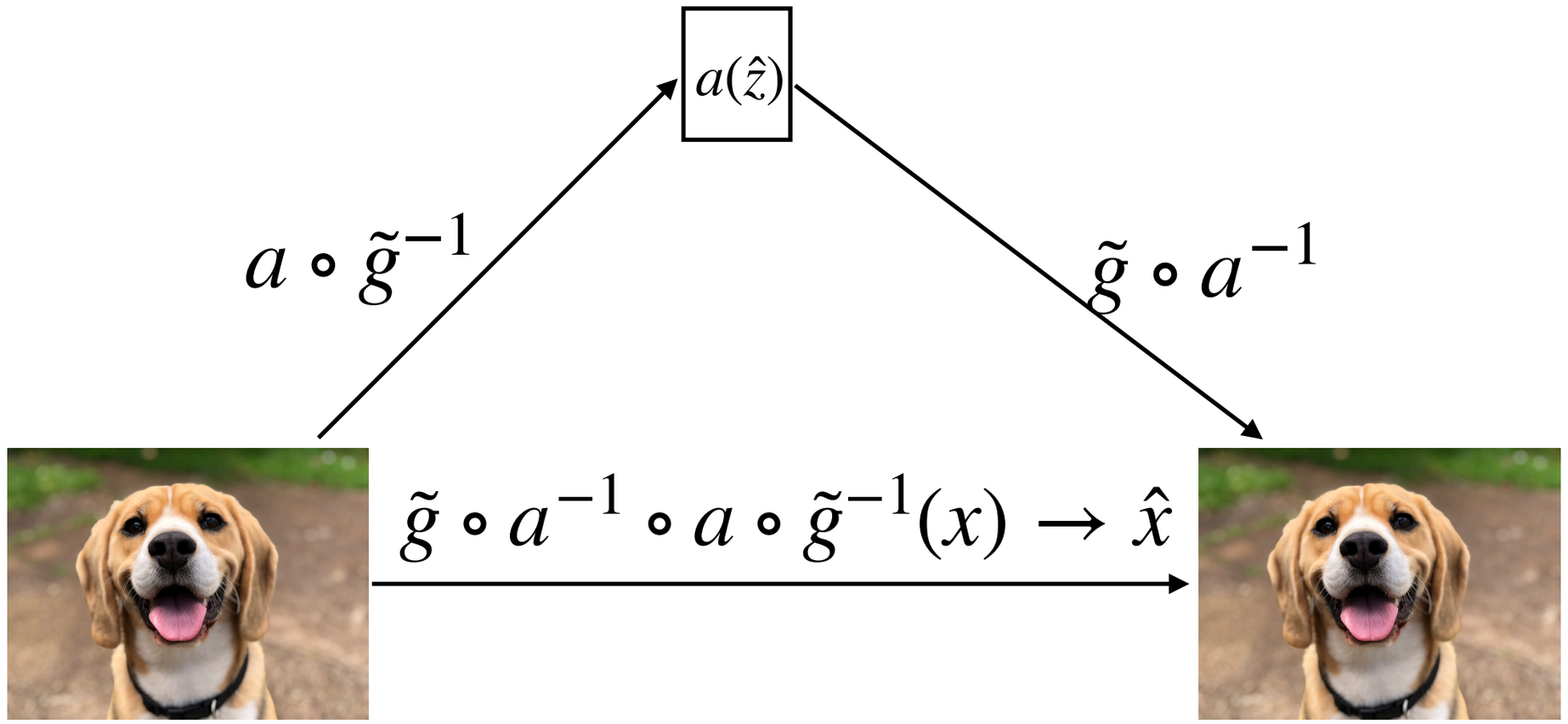
- Same reconstruction error can still lead to very different latents
- Can we simplify the bijective mapping a ?

Non-identification in Autoencoders



- Linear Identification: $a = L$, where L is some invertible matrix
- Disentanglement: $a = \Pi\Lambda$, where Π is permutation and Λ is diagonal matrix
- Causal Representation: $\hat{Z} = \Pi\Lambda Z$ where $\hat{Z} = \tilde{g}^{-1}(X)$ and $Z = g^{-1}(X)$

Non-identification in Autoencoders



- Identification without assumptions on DGP impossible [Hyvarinen et al.]
- Make assumptions on the latents (Z) or the decoder (g)

Linear Independent Component Analysis

Data generation process:

$X \leftarrow GZ$

$Z = (Z_1, \cdots, Z_d)$ are mutually independent and non-gaussian

Learned Model:

$\hat{X} \leftarrow \hat{G}\hat{Z}$

$\hat{Z} = (\hat{Z}_1, \cdots, \hat{Z}_d)$ are mutually independent

Linear Independent Component Analysis

Data generation process:

$X \leftarrow GZ$

$Z = (Z_1, \cdots, Z_d)$ are mutually independent and non-gaussian

Learned Model:

$\hat{X} \leftarrow \hat{G}\hat{Z}$

$\hat{Z} = (\hat{Z}_1, \cdots, \hat{Z}_d)$ are mutually independent

Perfect Reconstruction: $\hat{X} = X \implies \hat{G}\hat{Z} = GZ$

$$\implies \hat{Z} = \hat{G}^{-1}GZ$$

$$\implies \hat{Z} = AZ$$

Linear Independent Component Analysis

Theorem [Darmois–Skitovich]:

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.

Perfect Reconstruction: $\hat{X} = X$

$$\Longrightarrow \hat{G}\hat{Z} = GZ$$

$$\Longrightarrow \hat{Z} = \hat{G}^{-1} GZ$$

(Darmois-Skitovich Theorem)

$$\Longrightarrow \hat{Z} = AZ$$

$$\Longrightarrow \hat{Z} = \Pi \Lambda Z$$

Non-Linear ICA

Data generation process:

$X \leftarrow g(Z)$

$Z = (Z_1, \cdots, Z_d)$ are mutually independent and non-gaussian

Learned Model:

$\hat{X} \leftarrow \hat{g}(\hat{Z})$

$\hat{Z} = (\hat{Z}_1, \cdots, \hat{Z}_d)$ are mutually independent

Theorem [Hyvärinen et al.]:

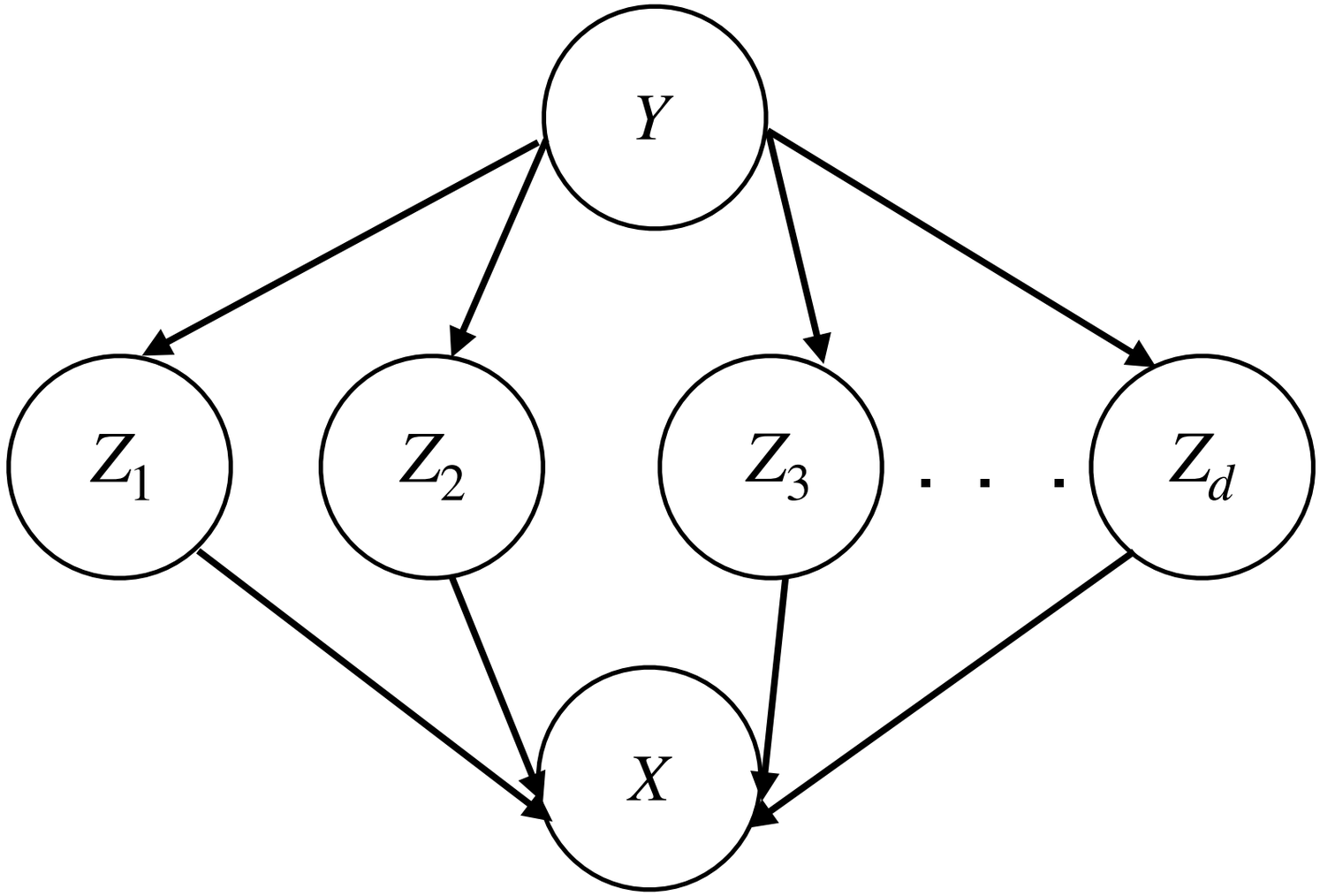
Non-Linear ICA with mutually independent latents is not identifiable

Non-Linear ICA

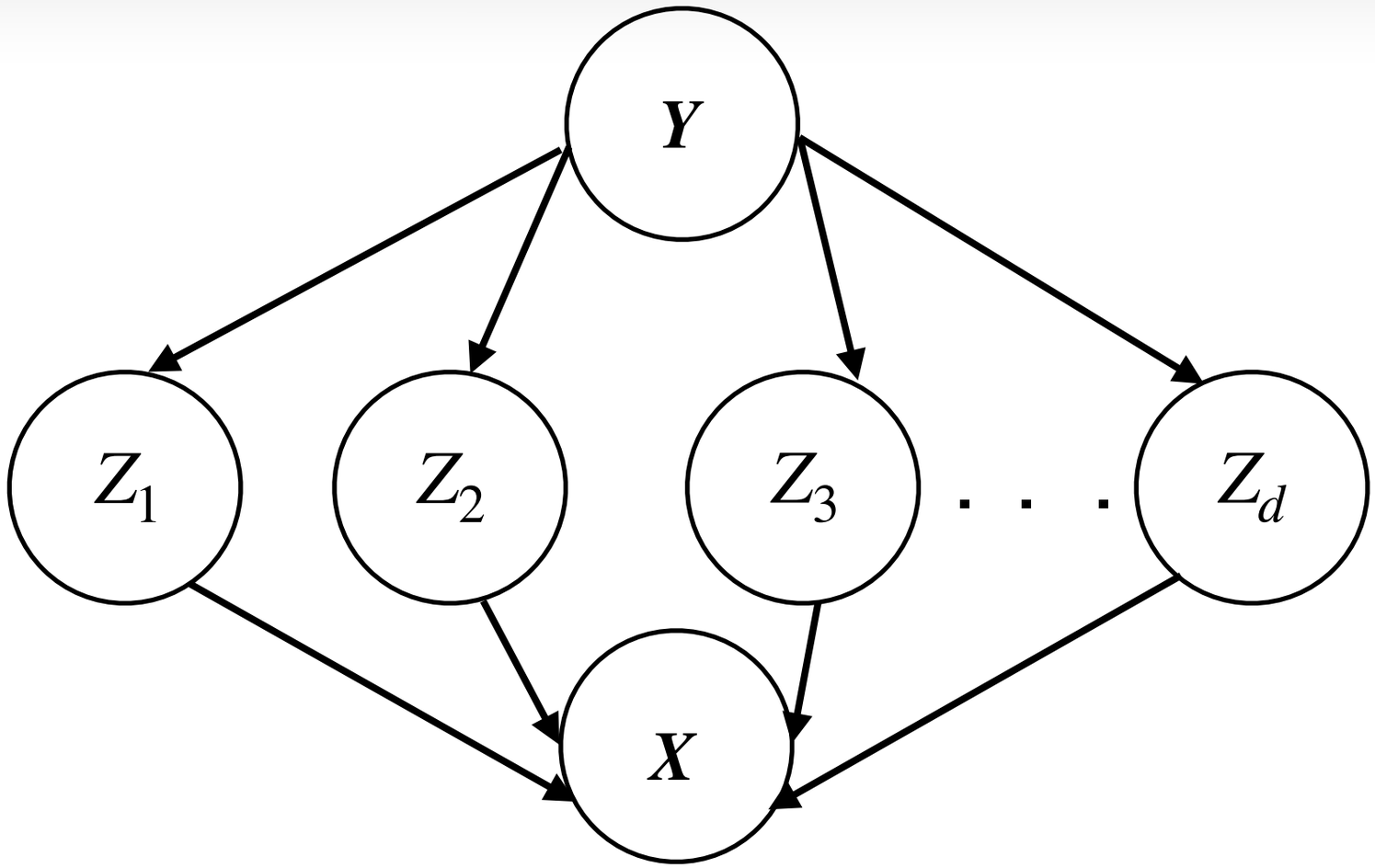
Auxiliary information (Labels) cause latents (e.g., Handwritten digits)

$$Z \leftarrow h(\mu_Y, \Sigma_Y)$$

$$X \leftarrow g(Z)$$



Non-Linear ICA



Therorem [Khemakhem et al.]:
Permutation recovery of latents under the following assumptions:

- All components of Z are independent conditional on Y
- $Z \mid Y$ follows exponential family distribution along with sufficient variability on the parameters

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

Towards Efficient Identification in Supervised Learning

Kartik Ahuja*, Divyat Mahajan*, Vasilis Syrgkanis, Ioannis Mitliagkas

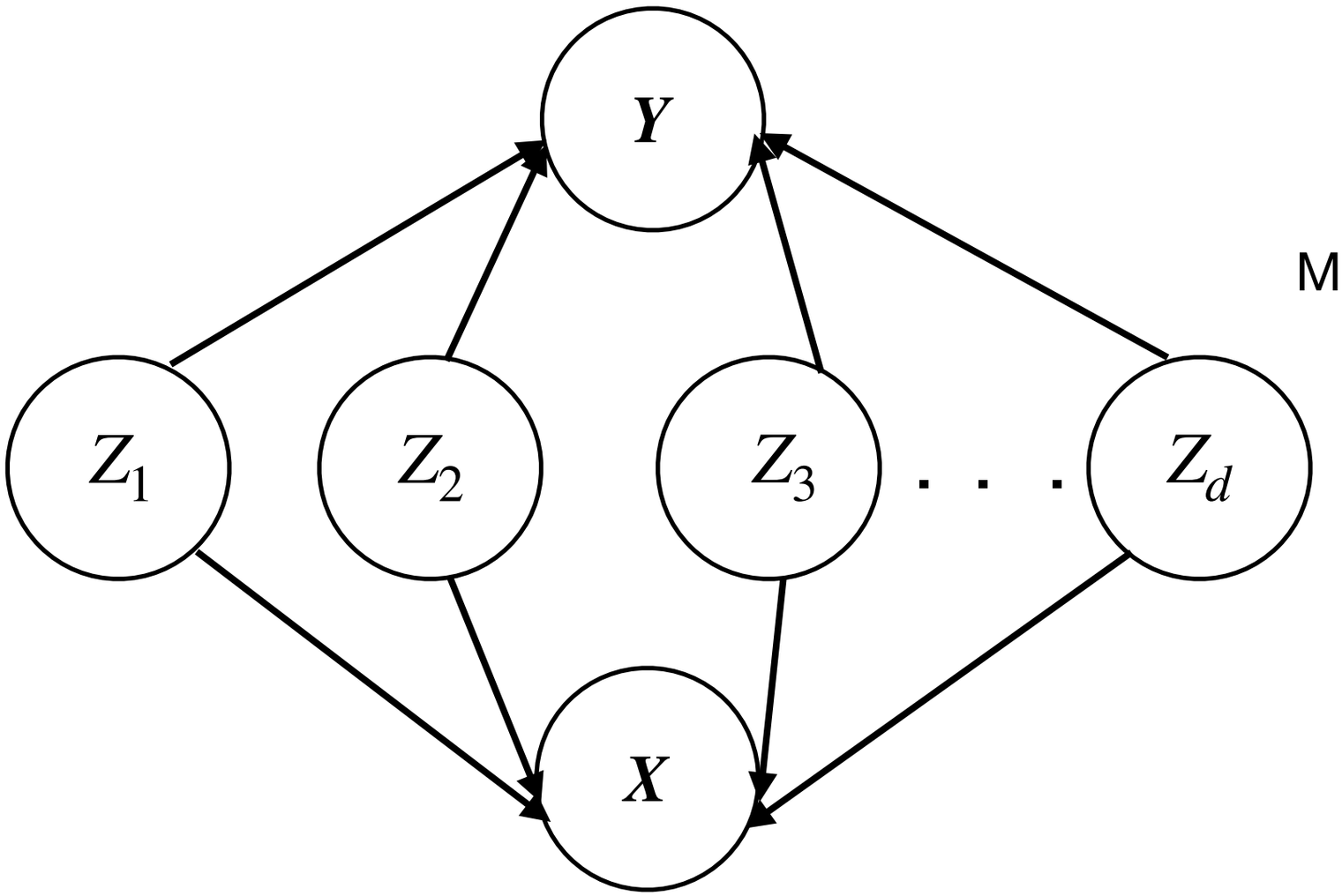
Proceedings of CleaR 2022

Mila
Université de Montréal

Problem Setting

Data Generation: Latents Cause Labels

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



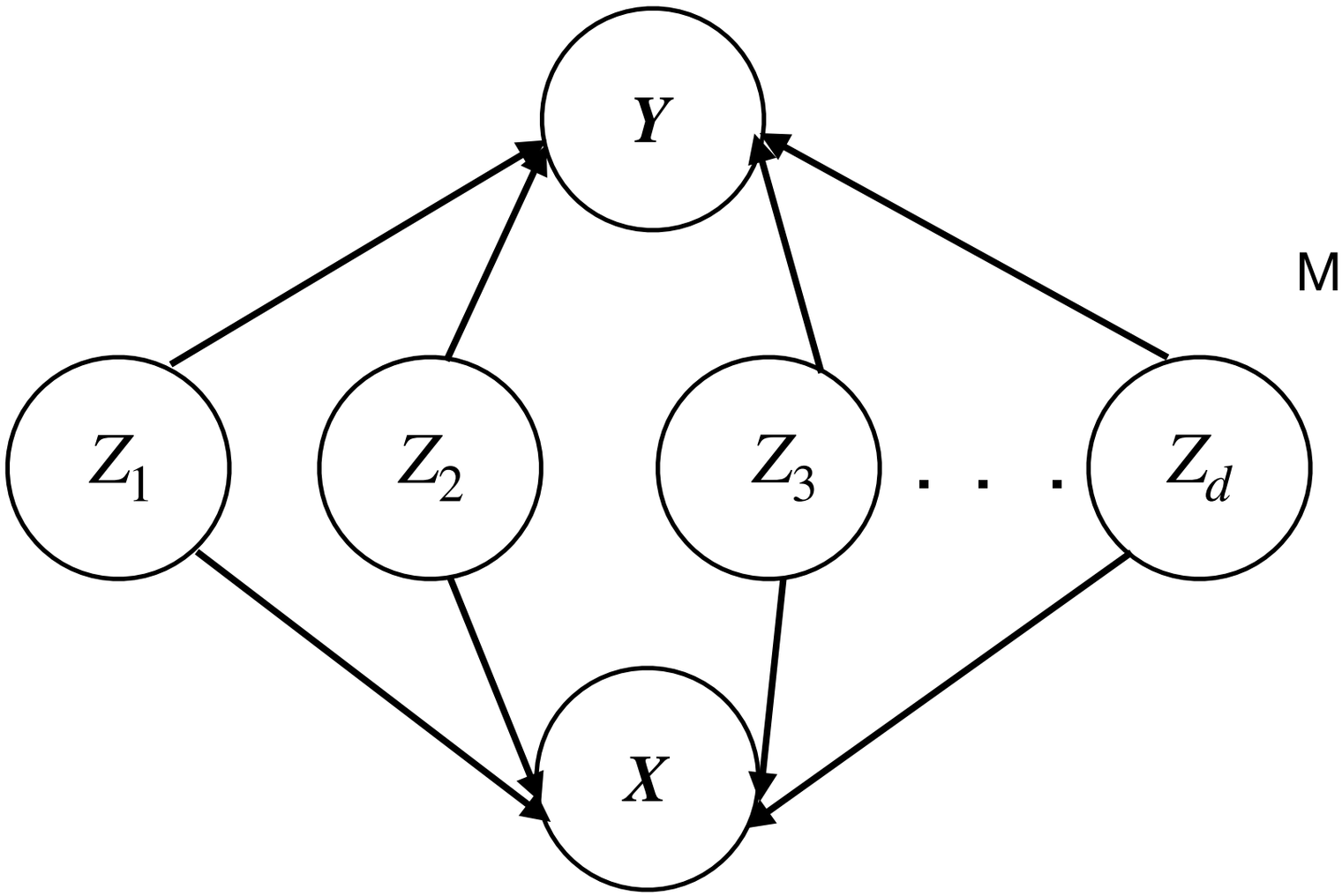
Multi-task Regression

$$\mathbf{Z} = (Z_1, \cdots, Z_d)$$

Mutually independent & Non-Gaussian

Data Generation: Latents Cause Labels

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



Multi-task Regression

$$\mathbf{Z} = (Z_1, \cdots, Z_d)$$

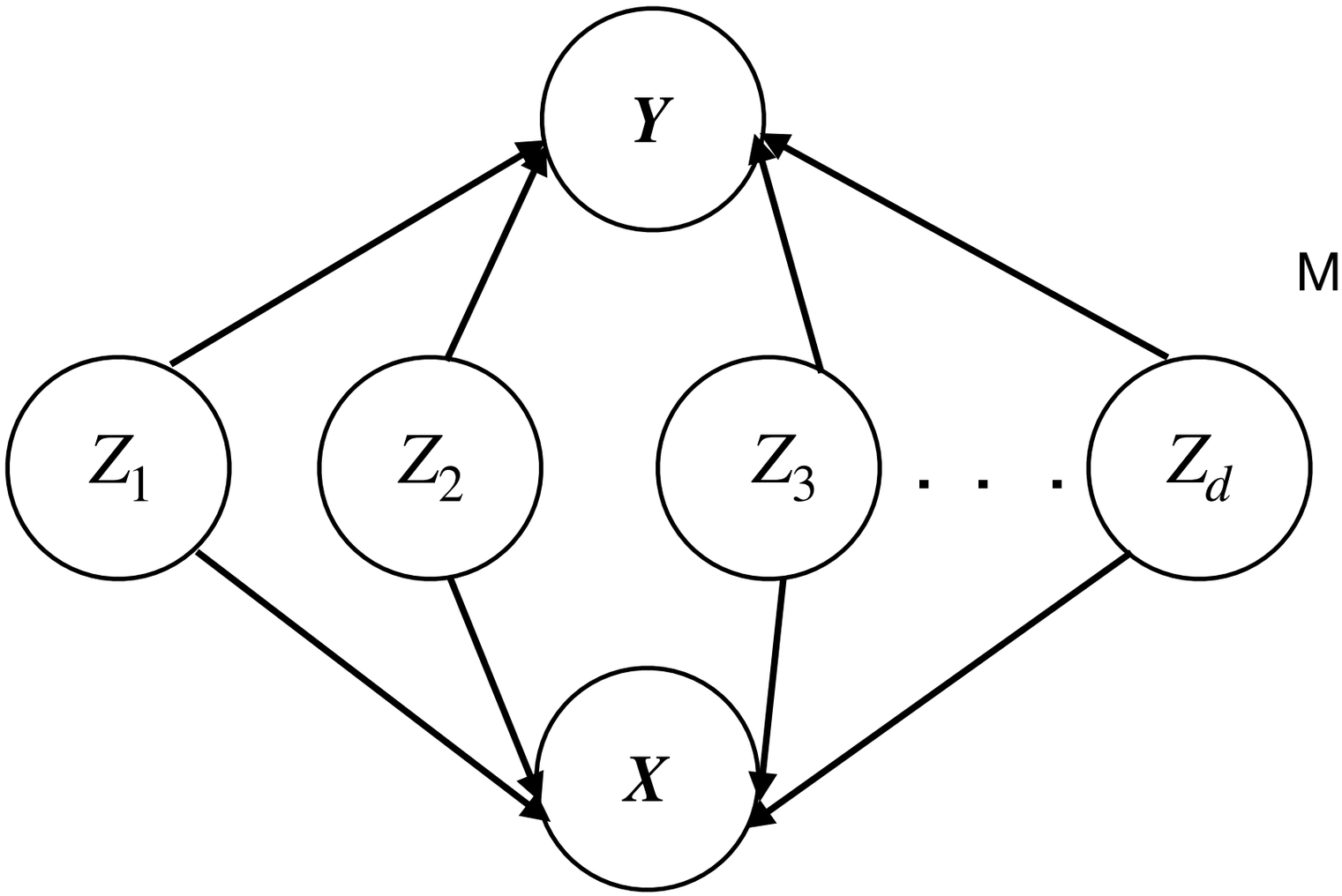
Mutually independent & Non-Gaussian

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

$$\mathbf{Y} \in R^k \,\&\, \mathbf{Z} \perp \mathbf{N}$$

Data Generation: Latents Cause Labels

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



Multi-task Regression

$$\mathbf{Z} = (Z_1, \cdots, Z_d)$$

Mutually independent & Non-Gaussian

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

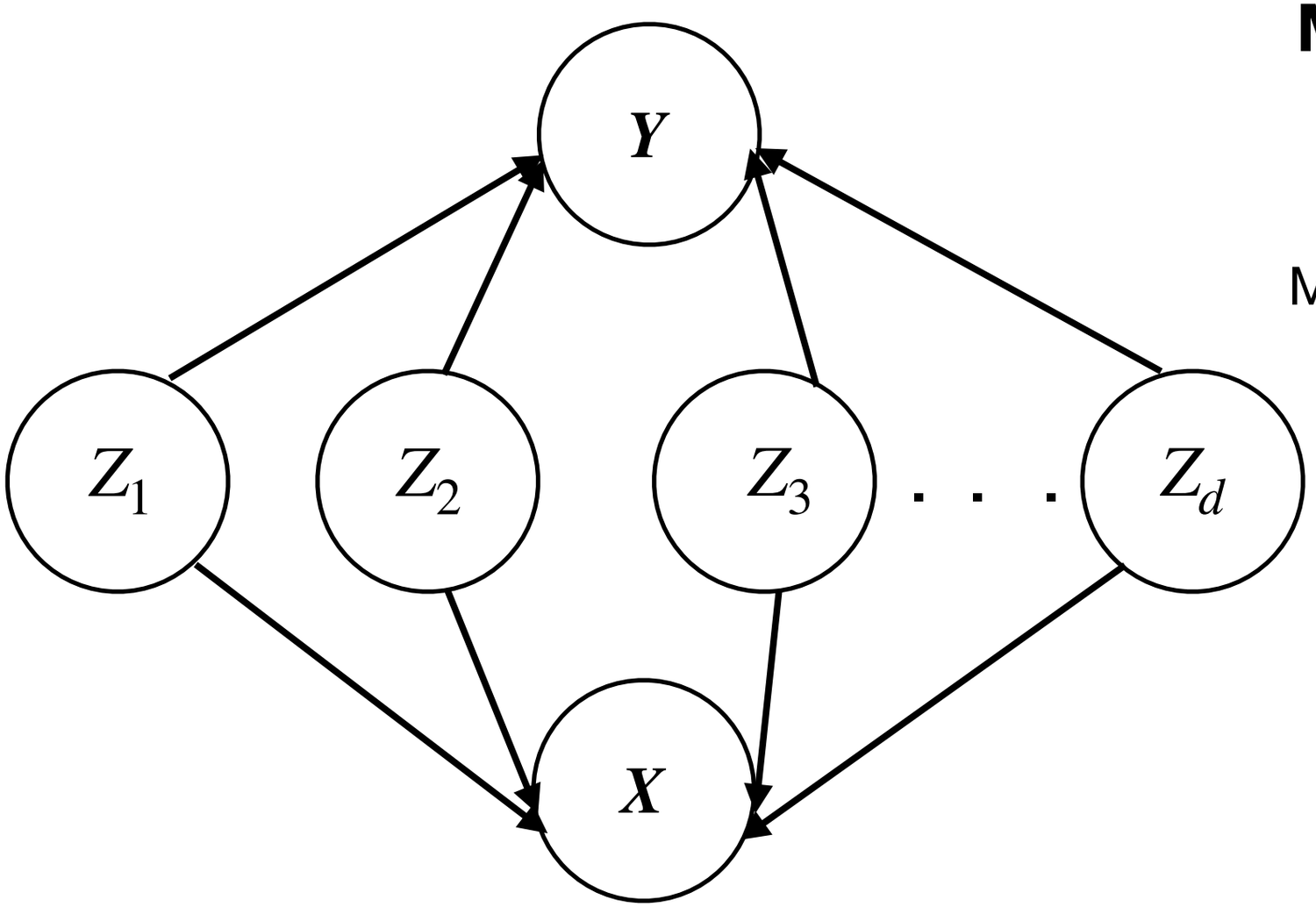
$\mathbf{Y} \in R^k$ & $\mathbf{Z} \perp \mathbf{N}$

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

g is bijection

Data Generation: Latents Cause Labels

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



Multi-task Classification

$$\mathbf{Z} = (Z_1, \cdots, Z_d)$$

Mutually independent & Non-Gaussian

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

$$\mathbf{Y} \in \{0,1\}^k$$

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

g is bijection

Identification Results

Independence Constrained ERM

Model: $W \circ \Phi$

$W \in \mathbb{R}^{d \times k}$: Linear Classifier

$\Phi \in \mathscr{H}_{\Phi}$: Non-linear Representation

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

Empirical Risk Minimization

Model: $W \circ \Phi$

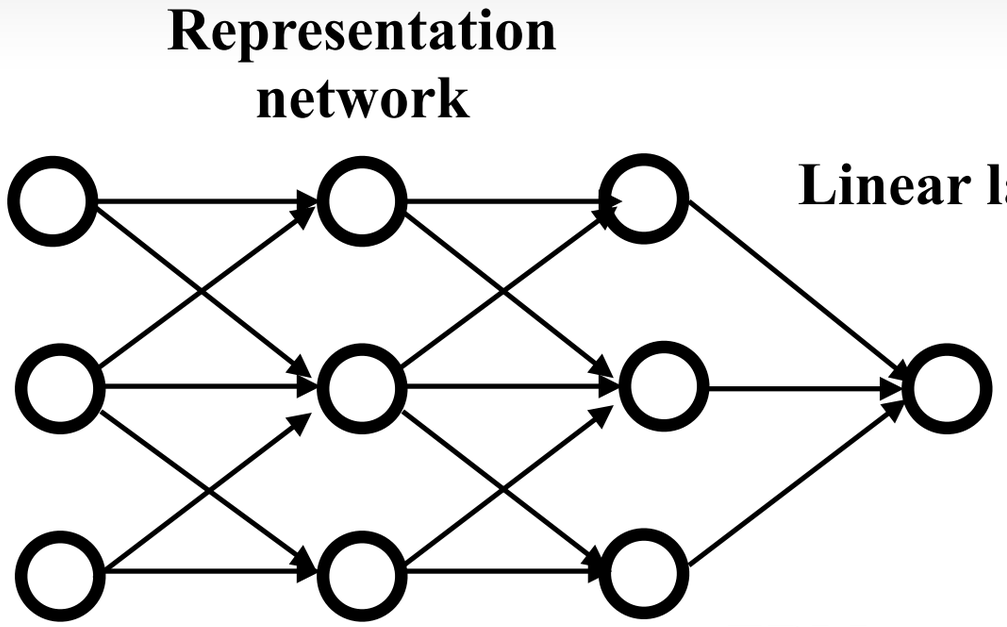
$W \in \mathbb{R}^{d \times k}$: Linear Classifier

$\Phi \in \mathscr{H}_{\Phi}$: Non-linear Representation

IC-ERM:

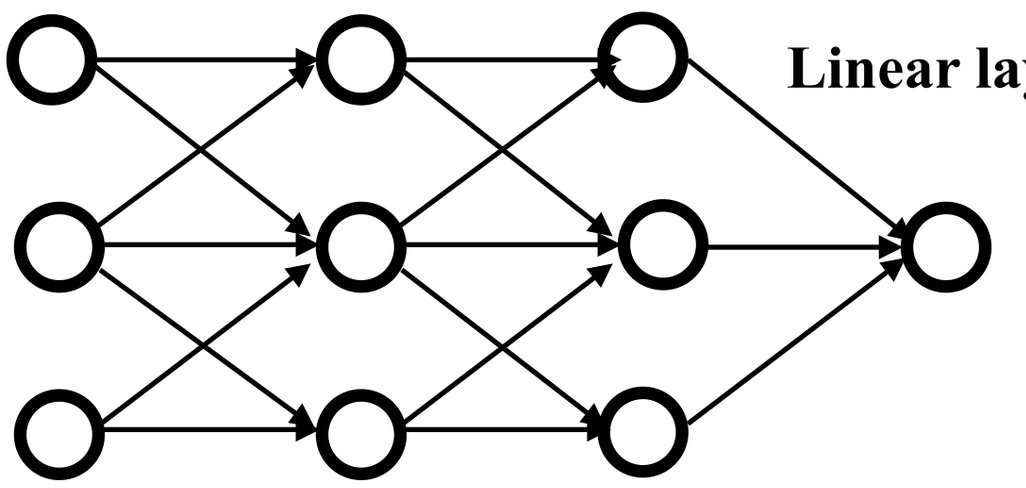
$$\min_{W \in \mathbb{R}^{d \times k}, \Phi} \sum_{i=1}^{N} \ell\Big(W \circ \Phi(X_i), Y_i\Big)$$ **s.t. Components of** $\Phi(X)$ **are i.i.d.**

ERM vs IC-ERM



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

ERM



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

IC-ERM

Inverting Latents Using IC-ERM

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

Theorem:

If number of tasks is equal to the latent dimension and $g^{-1} \in \mathscr{H}_{\Phi}$ then representation learned by **optimal**

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

Other Implications

- 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

Algorithm

ERM + Linear ICA

- Extract the representation learned by ERM:

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

- Process $\Phi^*(X)$ using Linear ICA to get independent latents!

ERM + Linear ICA

- Extract the representation learned by ERM:

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

- Process $\Phi^*(X)$ using Linear ICA to get independent latents!

Justification:

By the result proved before, ERM achieves linear identification

$\Phi^*(X) = \hat{Z} = AZ$ where Z has iid non-gaussian components

Same as the Linear ICA problem!

Experiments

Experiments

Data Generation

$$X \leftarrow g(Z)$$

Multi-task regression

$$Y \leftarrow \Gamma Z + N$$

Multi-task classification

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

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

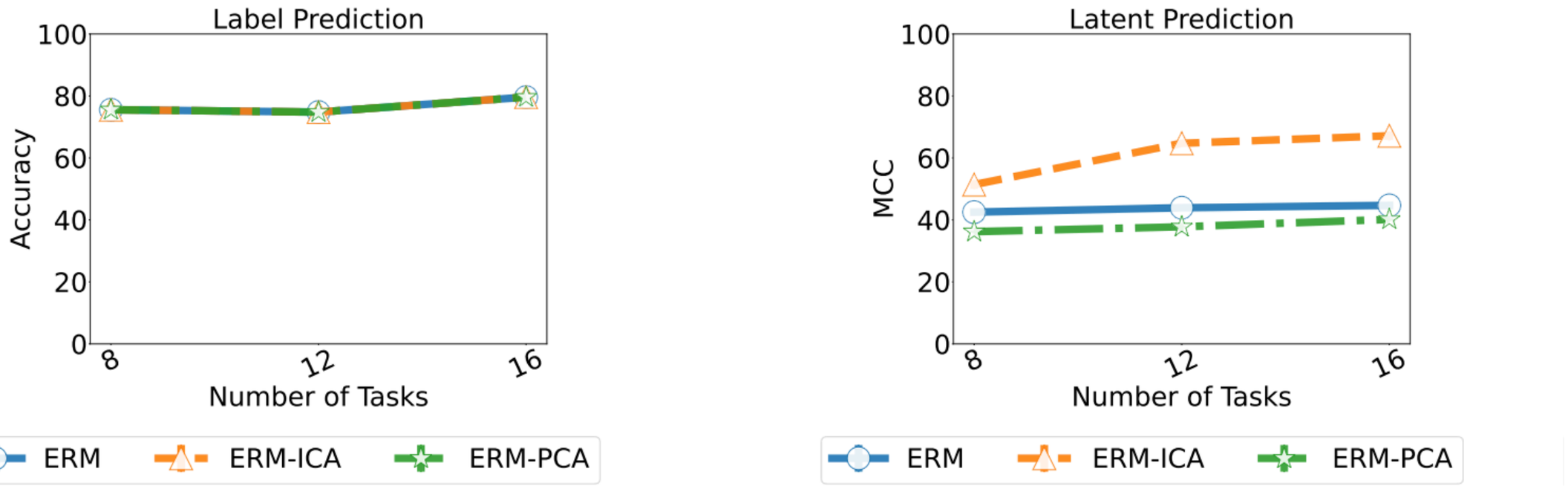


Figure 4: Comparison of label and latent prediction performance (classification, $d = 16$)

Assumptions in IC-ERM

i) Number of tasks is equal to the dimension of the latent

ii) Latent variables are all independent

Relaxing Assumption on Number of Tasks

Inverting Latents For Single Task

Assumption:

i) Number of tasks is equal to one

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

Theorem:

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

Relaxing Independence Assumption

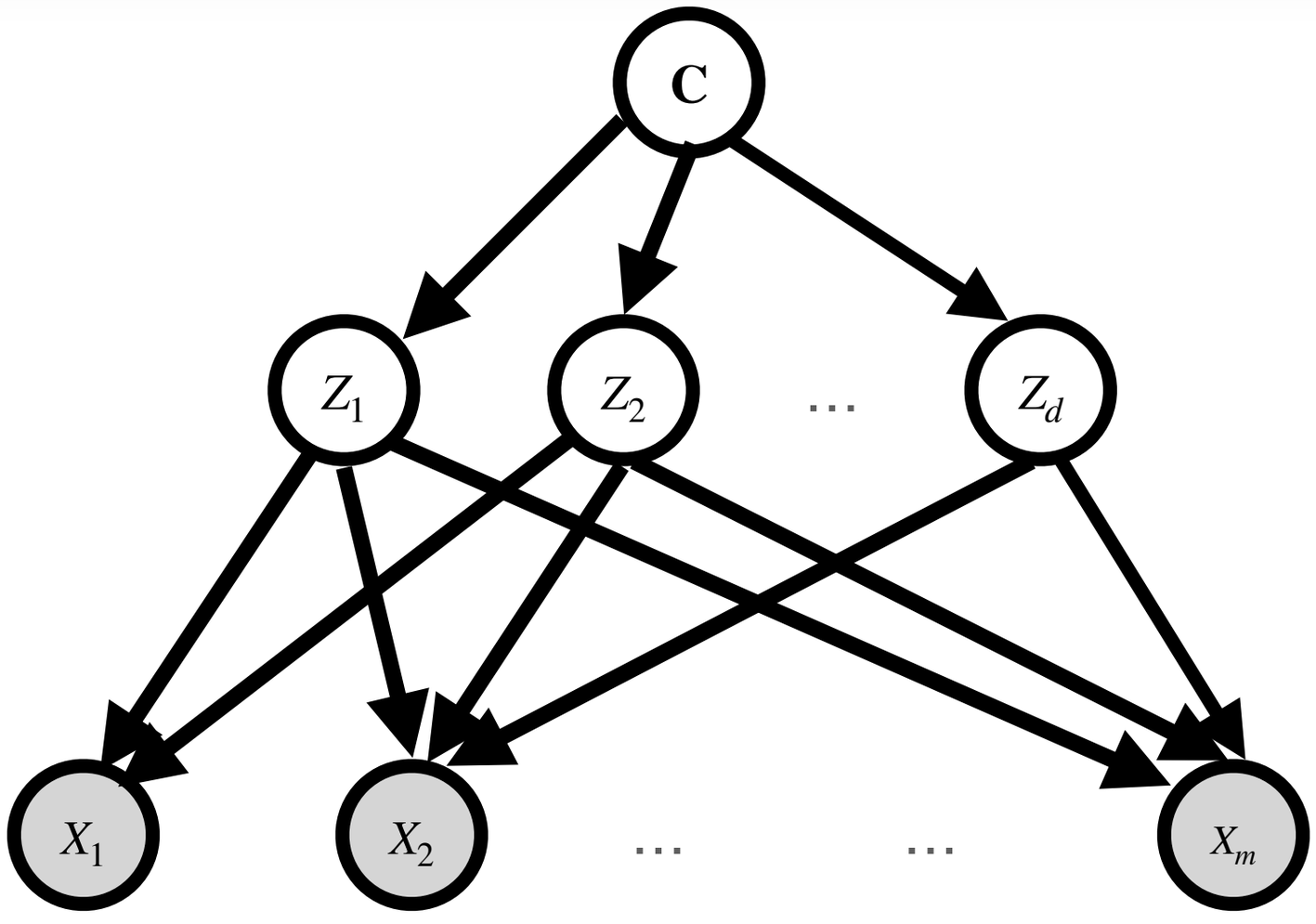
Interventional Causal Representation Learning

Kartik Ahuja, Divyat Mahajan, Yixin Wang, Yoshua Bengio

Proceedings of ICML 2023

Mila
Université de Montréal

Independent Support in Correlated Latents



$$Supp(Z_1, \cdots, Z_d) = Supp(Z_1) \times \cdots \times Supp(Z_d)$$

Linear Independent Support Analysis

Data generation process:

$X \leftarrow GZ$

$Z = (Z_1, \cdots, Z_d)$ have supports that are mutually independent

Learned Model:

$\hat{X} \leftarrow \hat{G}\hat{Z}$

$\hat{Z} = (\hat{Z}_1, \cdots, \hat{Z}_d)$ have supports that are mutually independent

Linear Independent Support Analysis

Data generation process:

$X \leftarrow GZ$

$Z = (Z_1, \cdots, Z_d)$ have supports that are mutually independent

Learned Model:

$\hat{X} \leftarrow \hat{G}\hat{Z}$

$\hat{Z} = (\hat{Z}_1, \cdots, \hat{Z}_d)$ have supports that are mutually independent

Theorem [Ahuja et al.]:

Linear ISA problem is identifiable upto permutation & scaling

ERM + Linear ISA

- Extract the representation learned by ERM:

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

- Linear ISA on $\Phi^*(X)$ to get independent support latents!

Synergies Between Disentanglement & Sparsity: Generalization & Identifiability in Multitask Learning

Sébastien Lachapelle*, Tristan Deleu*, Divyat Mahajan, Ioannis Mitliagkas, Yoshua Bengio, Simon Lacoste-Julien & Quentin Bertrand

Proceedings of ICML 2023

Mila
Université de Montréal

Sparse Task Predictors

Data generation process:

$Y \sim p(Y;\ \eta = W^{(t)} f_\theta(X))$ s.t. $W^{(t)}$ is sparse for all tasks t

$Z = (Z_1, \cdots, Z_d)$ s.t. Z_i may have statistical dependencies between them

Learned Model:

$$\min_{\hat{\theta}} -\log p(Y;\ \hat{W}^{(t)} f_{\hat{\theta}}(X))$$

s.t. $\hat{W}^{(t)} \in arg \min_{W} -\log p(Y;\ W f_{\hat{\theta}}(X))$ where W is restricted to be sparse

Sparse Task Predictors

Data generation process:

$Y \sim p(Y;\ \eta = W^{(t)} f_\theta(X))$ s.t. $W^{(t)}$ is sparse for all tasks t

$Z = (Z_1, \cdots, Z_d)$ s.t. Z_i may have statistical dependencies between them

Learned Model:

$$\min_{\hat{\theta}} -\log p(Y;\ \hat{W}^{(t)} f_{\hat{\theta}}(X))$$

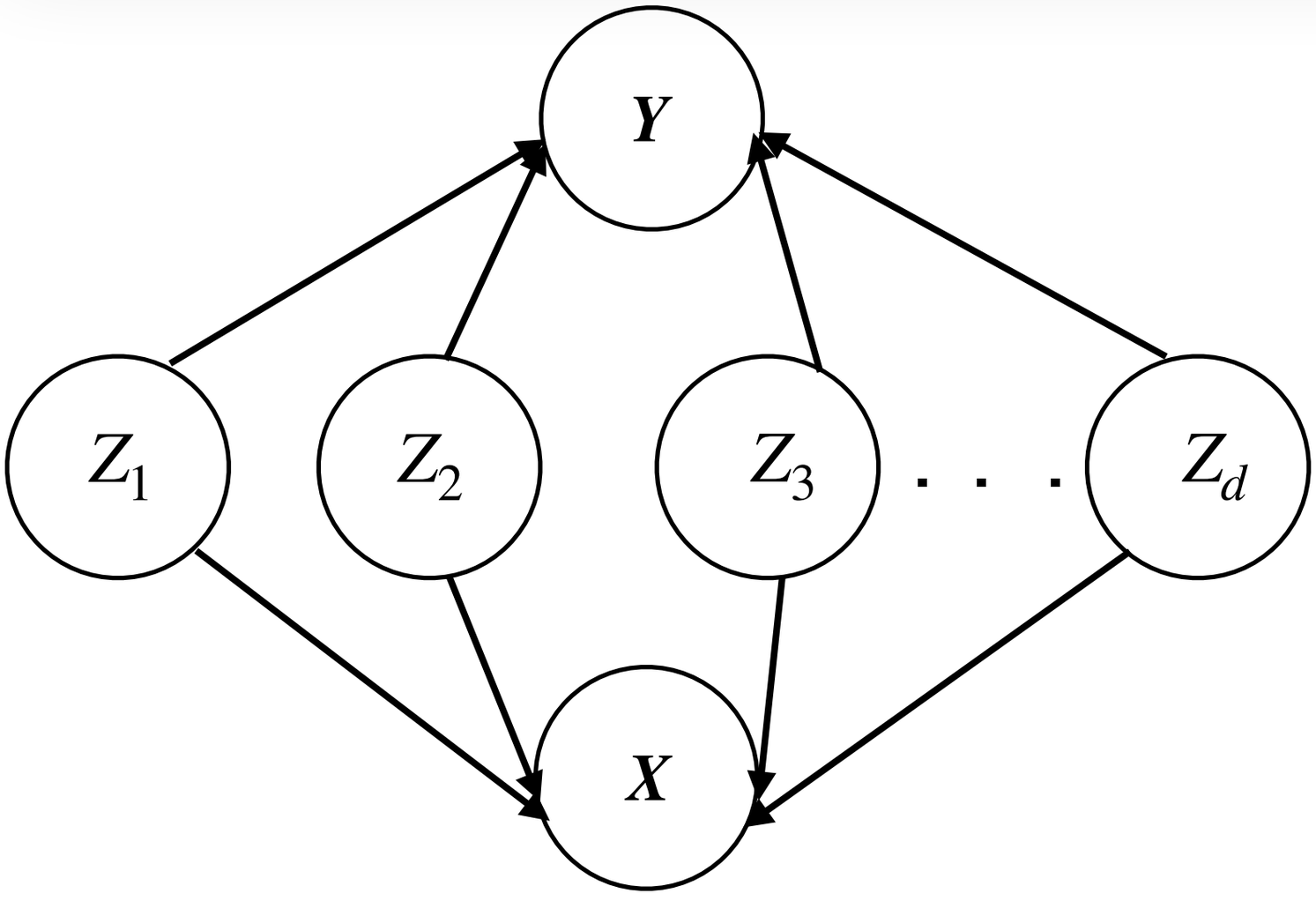
s.t. $\hat{W}^{(t)} \in arg \min_{W} -\log p(Y;\ W f_{\hat{\theta}}(X))$ where W is restricted to be sparse

Theorem [Lachapelle et al.]:

Learned model from the above optimisation objective identifies latents unto permutation & scaling under assumptions of sufficient sparsity on the task predictors ($W^{(t)}$)

How to use these learned representations?

Transfer to New Tasks



Multi-task Regression

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

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

New task

$$\tilde{Y} \leftarrow \tilde{\gamma}^{\mathsf{T}} \mathbf{Z} + N'$$

Transfer to Unseen Tasks

New Task: $\tilde{Y} \leftarrow \tilde{\gamma}^{\top} \boldsymbol{Z} + N'$

- New task is linear and sparse in the representation learned by ERM+ICA
- New task is linear but not sparse in the representation learned by ERM
- ERM+ICA should require fewer samples from the new task to adapt

Transfer to New Tasks

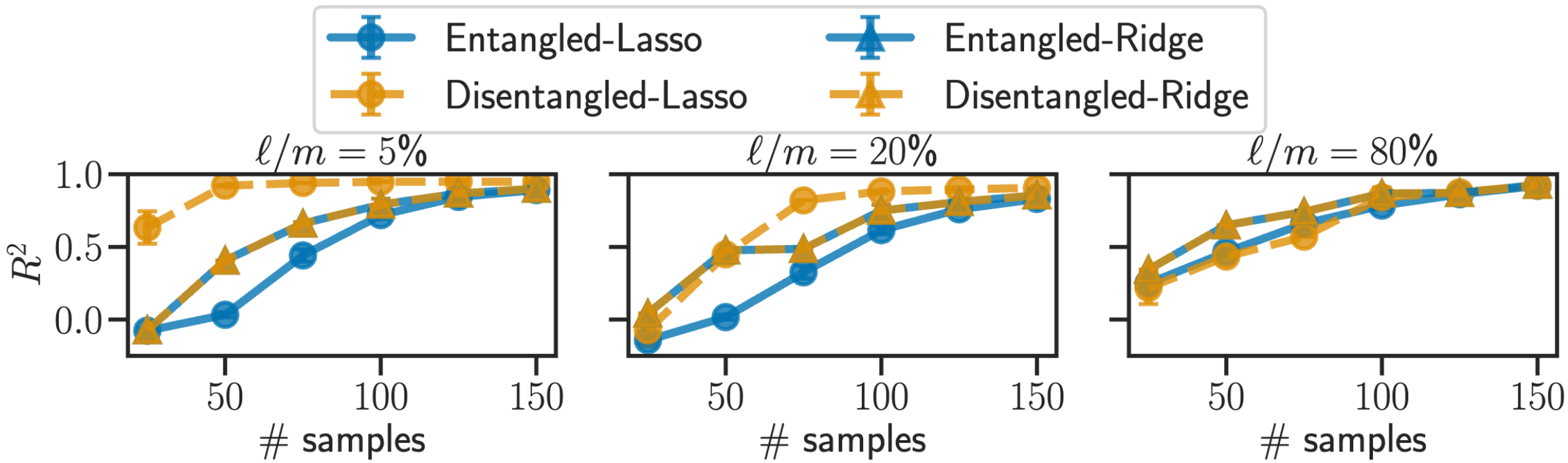


Figure 1: Test performance for the entangled and disentangled representation using Lasso and Ridge regression. All the results are averaged over 10 seeds, with standard error shown in error bars.

Lasso/Ridge: Choice for learning the final classifier

l/m: Sparsity Ratio

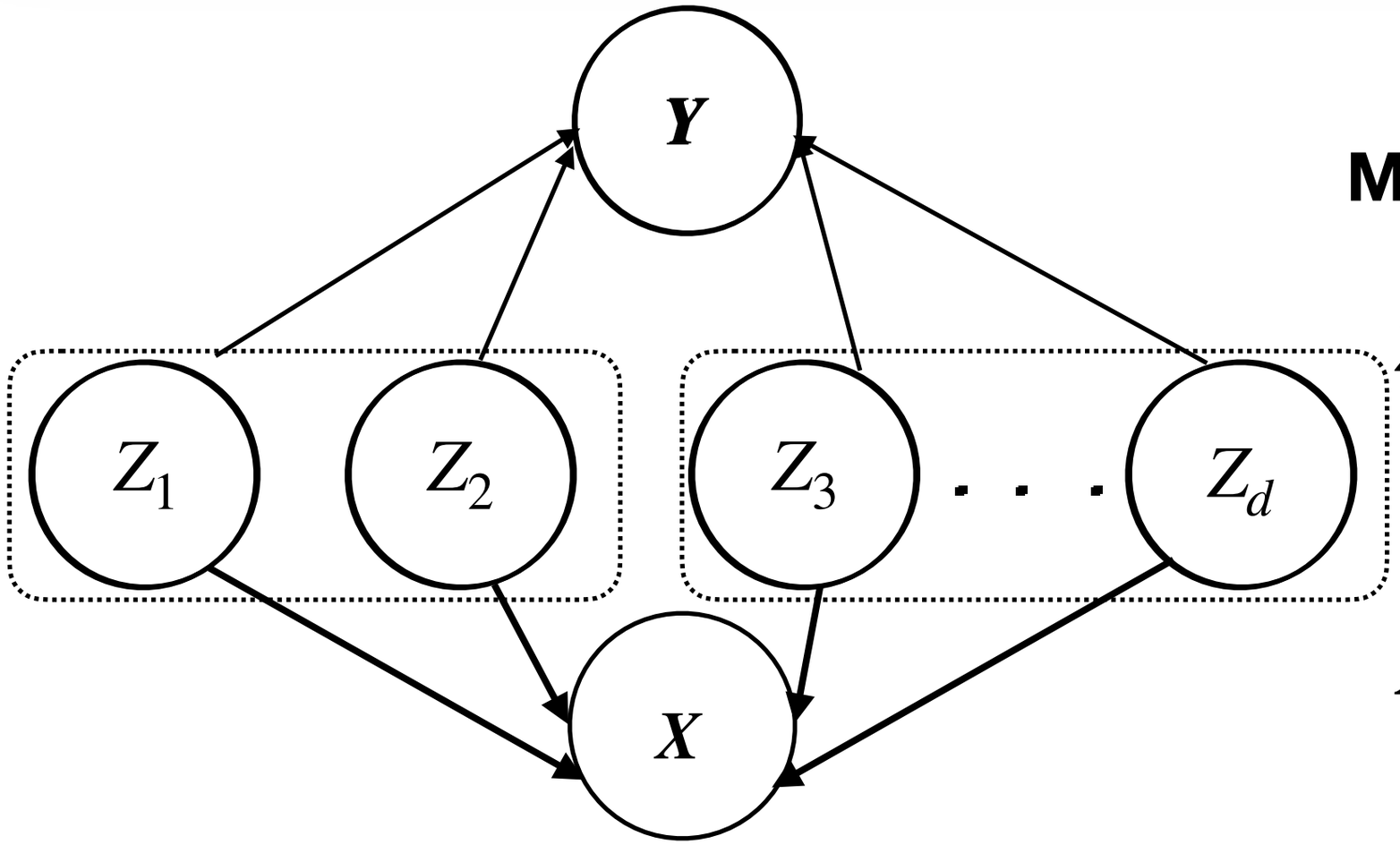
m: 100

Conclusion

- Bridge between multi-task learning and causal representation learning
- Identification guarantees even when we have few tasks
- Identification guarantees for statistically dependent latents
- Causal representations transfer better to novel sparse tasks

Thank You!

Block-wise independence



Multi-task regression

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

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

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

Linear Blockwise ICA

$$X \leftarrow GZ$$

Generalized Darmois Theorem:

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 Blockwise ICA

$$X \leftarrow GZ$$

Theis et al. result on linear ICA

If every k-block of Z is non-Gaussian, then it is possible to recover blocks of Z up to permutation and linear transformation, i.e., $\hat{G}X = \hat{Z} = \Pi\Lambda Z$, where Π is permutation matrix and Λ is a block-diagonal matrix.

Blockwise Independence Constrained ERM

Blockwise Independence-constrained ERM:

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

Inverting Latents Using BIC-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) BIC-ERM identifies true latent up to permutation & block diagonal scaling