Towards efficient representation identification in supervised learning

Kartik Ahuja [1] Divyat Mahajan [1] Vasilis Syrgkanis [2] Ioannis Mitliagkas [1]

[1]Mila - Quebec AI Institute, University of Montreal [2]Microsoft Research, New England

Paper: arxiv/2204.04606 **Code:** github/divyat09/ood-identification

Representation Identification: Introduction

- **Non-Linear ICA:** Recover the latent variables Z given the observations $X = g(Z)$, where g in general is non-linear, invertible function.
- **Identifiability:** If the inferred latents $\tilde{Z} = \tilde{g}^{-1}(X)$ and the true latents $Z = g^{-1}(X)$ are related by some bijection $a \in \mathcal{A}$, such that $\tilde{Z}^{-1} = a \circ Z^{-1}$, then $\tilde{Z}^{-1} \sim_{\mathcal{A}} Z^{-1}$.
- **Unidentifiability of Non-Linear ICA:** Without further structural assumptions or access to auxiliary information, non-linear ICA will not be identifiable upto simple transformations. [1]
- **Contributions:**
  - Propose the independence-constrained ERM objective that guarantees solution to non-linear ICA upto permutation and scaling in supervised learning setup.
  - Practical implementation of the proposed objective with a two phase approach using ERM and Fast ICA.

Comparison of Data Generation Process

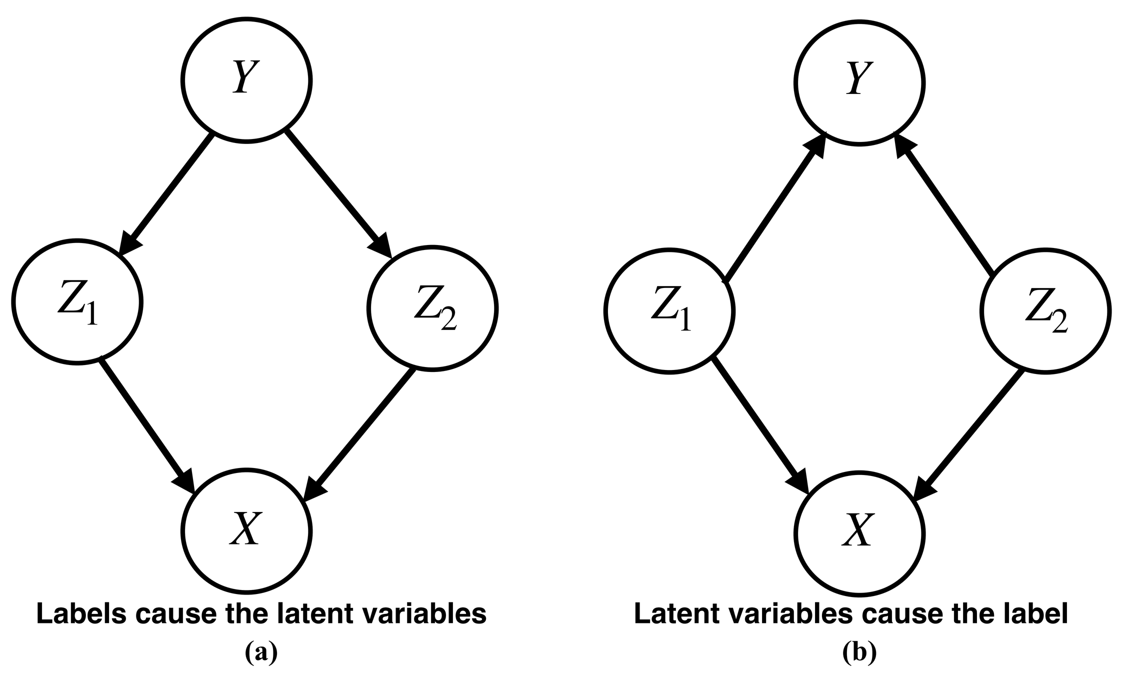

Figure: (a) Data generation process in [2]; (b) Data generation process studied in our work.

- **Prior Works:** Latent variables caused by labels, and rendered conditionally independent on labels.

$$Y \leftarrow \text{Bernoulli}\left(\frac{1}{2}\right) \qquad Z \leftarrow \mathcal{N}(Y\mathbf{1}, \mathrm{I}) \qquad X \leftarrow g(Z) \tag{1}$$

- **Our Work:** Labels are caused by the mutually independent latent variables

$$Z \leftarrow h(N_Z) \qquad X \leftarrow g(Z) \qquad Y \leftarrow \Gamma Z + N_Y \tag{2}$$

- **Notations:**
  - $N_Z \in \mathbb{R}^d$ is noise, $h : \mathbb{R}^d \to \mathbb{R}^d$ generates $Z \in \mathbb{R}^d$
  - $g : \mathbb{R}^d \to \mathbb{R}^d$ is a bijection that generates the observations X
  - $\Gamma \in \mathbb{R}^{k \times d}$ is a matrix that generates the label $Y \in \mathbb{R}^k$ and $N_Y \in \mathbb{R}^k$ is the noise vector (N_Y is independent of Z and $\mathbb{E}[N_Y] = 0$)

IC-ERM: Independence-constrained ERM

- **Definition of IC-ERM objective:**

$$\min_{\Theta \in \mathcal{H}_\Theta, \Phi \in \mathcal{H}_\Phi} R(\Theta \circ \Phi) \ \text{ s.t. } \Phi(X) \text{ is mutually independent} \tag{3}$$

- **Theorem 1:** If the assumptions on our data generation process hold and the number of tasks k is equal to the dimension of the latent d, then the solution $\Theta^\dagger \circ \Phi^\dagger$ to IC-ERM (3) with ℓ as loss function.
- **Case of single task k=1:** We consider a slightly modified data generation process.

$$Z \leftarrow h(N_Z) \qquad X \leftarrow g^{'}(U) \qquad Y \leftarrow \mathbf{1}^\mathsf{T} U + N_Y \tag{4}$$

- **Reparametrized IC-ERM objective:**

$$\min_{\Phi \in \mathcal{H}_\Phi} R(\mathbf{1} \circ \Phi) \ \text{ s.t. } \Phi(X) \text{ is i.i.d.} \tag{5}$$

- **Theorem 2:** If the assumptions on the modified data generation process hold and some extra assumptions hold, then the solution $\Phi^\dagger(X)$ of reparametrized IC-ERM objective recovers the true latent U up to permutations.

ERM-ICA: Practical Implementation of IC-ERM

- We propose a two step approximation method as ERM-ICA:
  - **ERM Phase:** Learn $\Theta^\dagger, \Phi^\dagger$ by solving the ERM objective.

$$\Theta^\dagger, \Phi^\dagger \in_{\Theta \in \mathcal{H}_\Theta, \Phi \in \mathcal{H}_\Phi} R(\Theta \circ \Phi) \tag{6}$$

  - **ICA Phase:** [3] Learn $\Omega^\dagger$ by linear ICA on the representation from ERM Phase ($\Phi^\dagger$).

$$\Omega^\dagger \in_{\Omega, \Omega \text{ is invertible}} I(\Omega \circ \Phi^*(X)) \tag{7}$$

- **Theorem 3:** If the assumptions on our data generation process hold and the number of tasks k is equal to the dimension of the latent d, then the solution $\Omega^\dagger \circ \Phi^\dagger$ to ERM-ICA with ℓ as loss function identifies true Z up to permutation and scaling.

References

[1] A. Hyvärinen and P. Pajunen, ``Nonlinear independent component analysis: Existence and uniqueness results,'' *Neural networks*, vol. 12, no. 3, pp. 429--439, 1999.

[2] I. Khemakhem, R. P. Monti, D. P. Kingma, and A. Hyvärinen, ``Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ica,'' *arXiv preprint arXiv:2002.11537*, 2020.

[3] P. Comon, ``Independent component analysis, a new concept?,'' *Signal processing*, vol. 36, no. 3, pp. 287--314, 1994.

Results: Regression Case

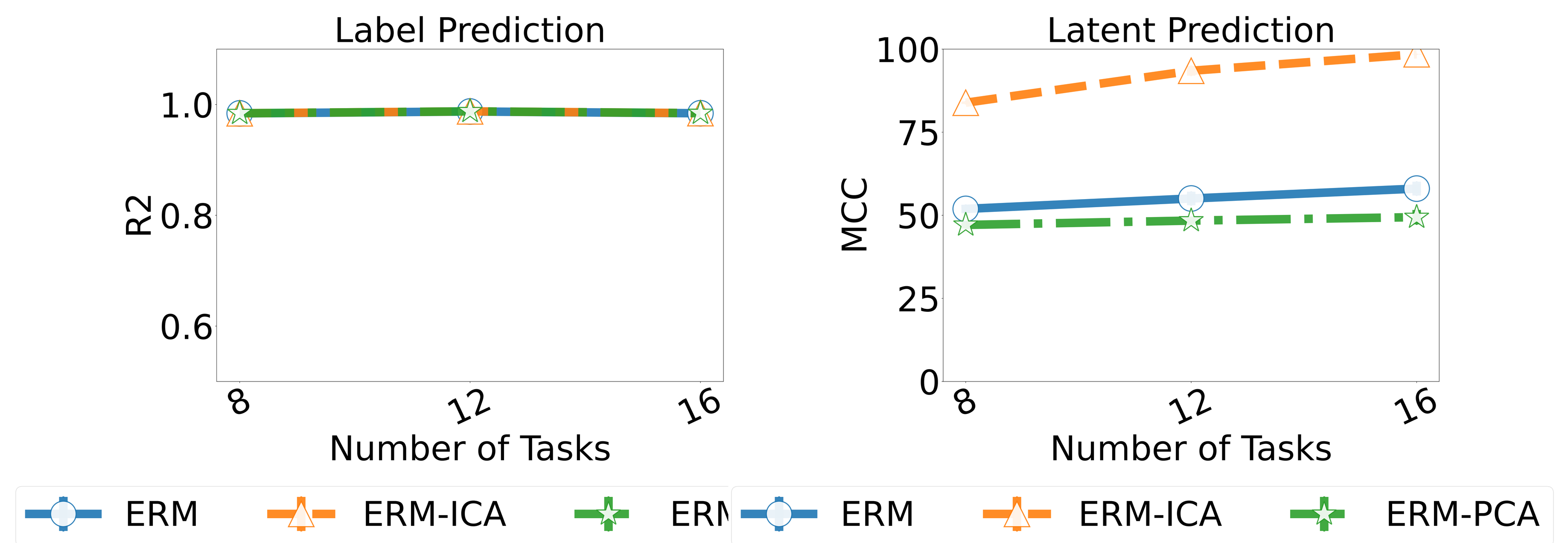

Figure: Comparison of label and latent prediction performance (regression, $d = 16$).

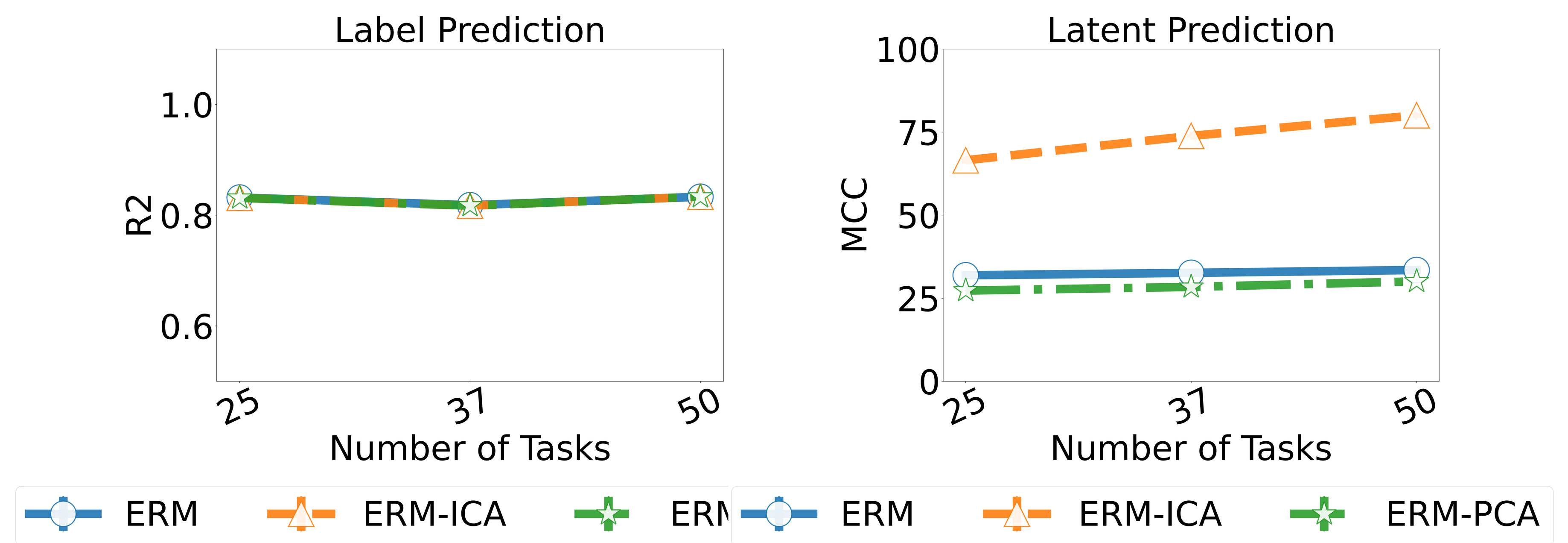

Figure: Comparison of label and latent prediction performance (regression, $d = 16$).

Results: Classification Case

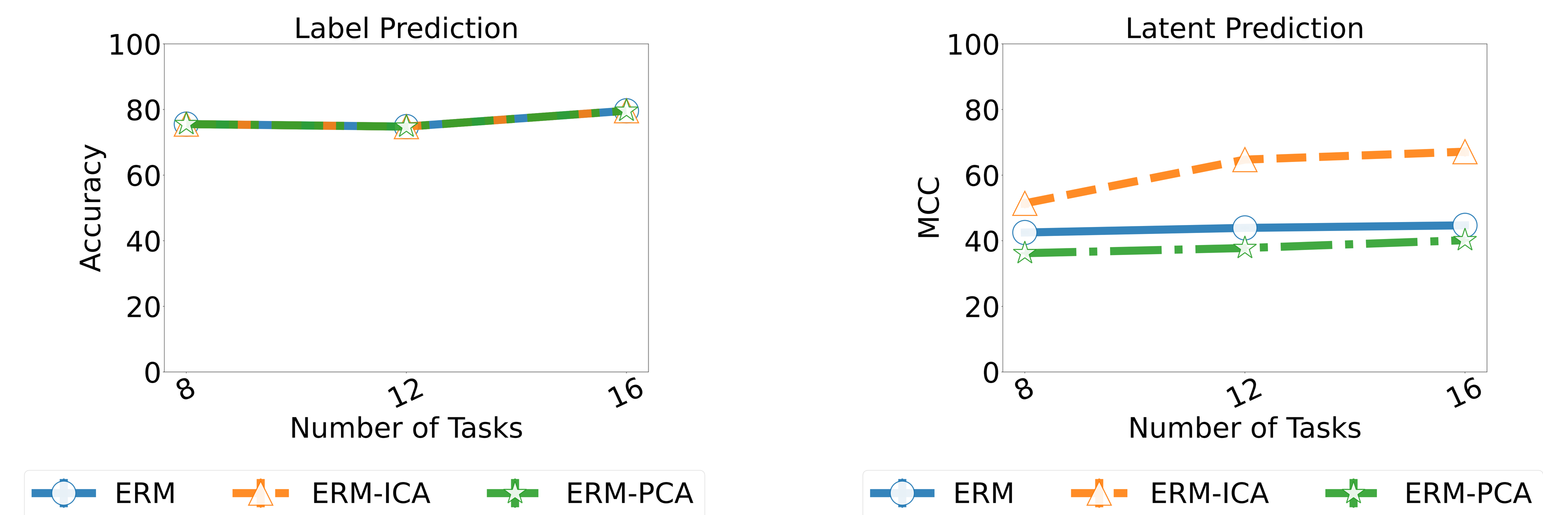

Figure: Comparison of label and latent prediction performance (regression, $d = 16$).

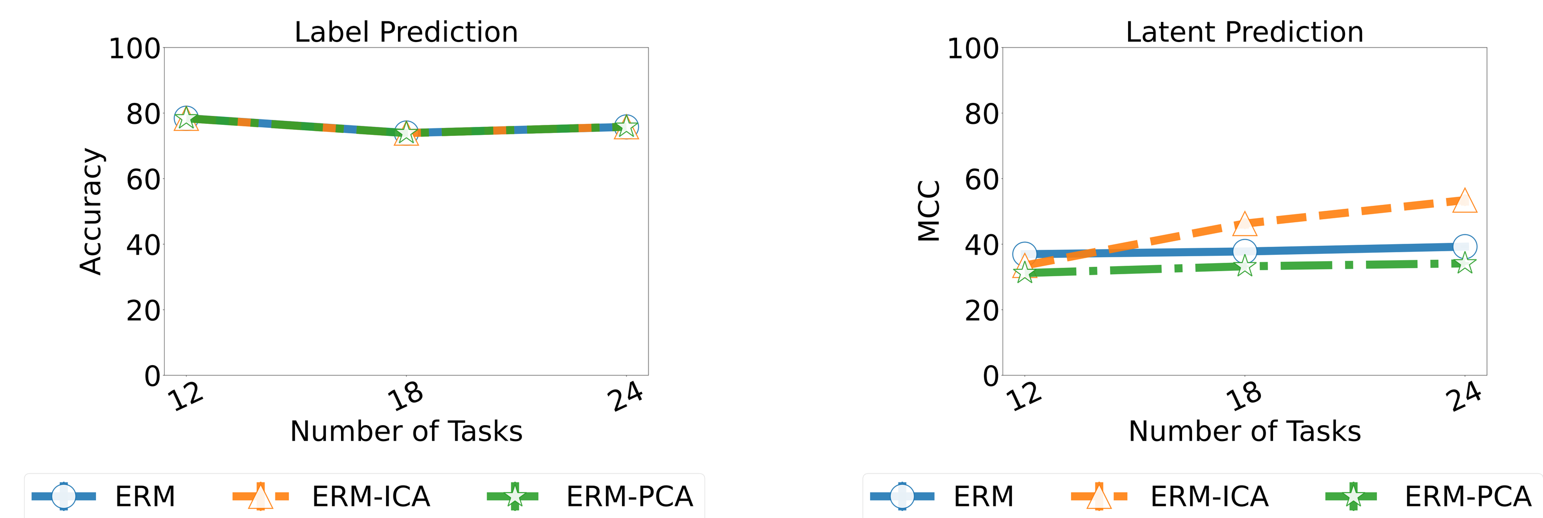

Figure: Comparison of label and latent prediction performance (regression, $d = 16$).