Diverse Influence Component Analysis: A Geometric Approach to Nonlinear Mixture Identifiability

Hoang-Son Nguyen, Xiao Fu

School of Electrical Engineering and Computer Science, Oregon State University

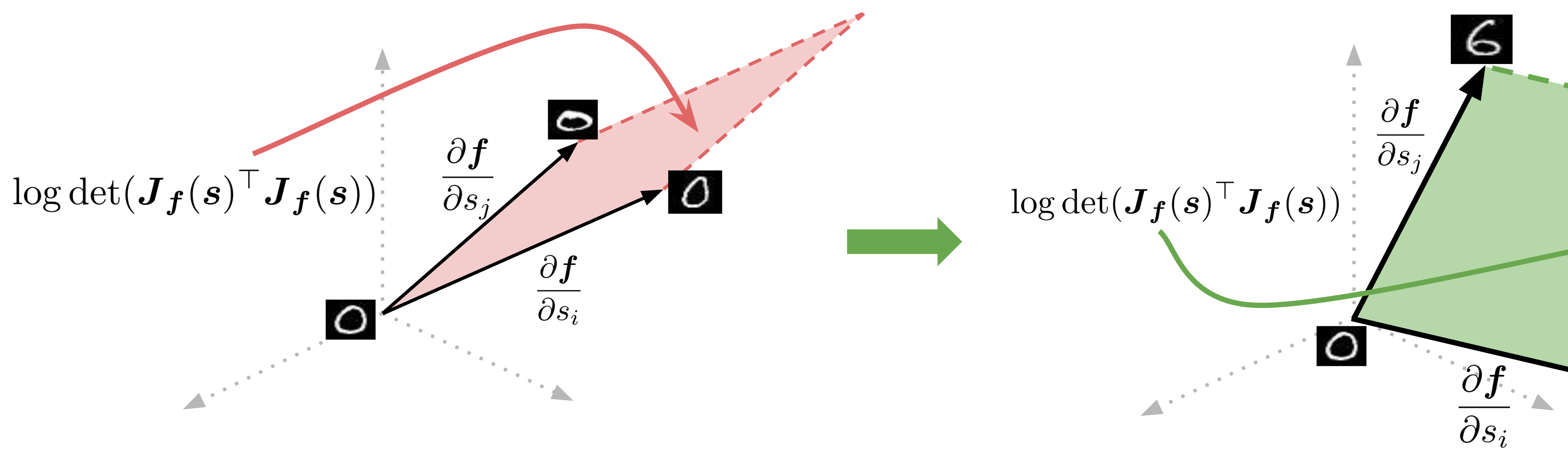


Illustration: DICA attempts to find representation $\boldsymbol{s}$ s.t. $\partial \boldsymbol{x}/\partial s_i$ and $\partial \boldsymbol{x}/\partial s_j$ spanning quite different directions, inducing larger convex hull volume (under constraints).

Nonlinear Mixture Model Identification (NMMI) 🤔

A diffeomorphism mapping latent $\boldsymbol{s}$ to a d-dim. data manifold embedded in $\mathbb{R}^m$:

$$\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{s}), \boldsymbol{s} \in \mathbb{R}^d, \boldsymbol{x} \in \mathbb{R}^m, d \leq m \tag{1}$$

- $\boldsymbol{s} = [s_1, s_2, ..., s_d] \sim p(\boldsymbol{s})$ are **latent** variables (object positions, lighting,...),
- $\boldsymbol{x} = [x_1, x_2, ..., x_m]$ are **observed** features (pixels).
- $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^m$ is nonlinear mixing function.

Goal: Recovering of $\boldsymbol{s}$ and $\boldsymbol{f}$ (up to acceptable ambiguities)
Learn an encoder $\boldsymbol{g}_{\boldsymbol{\phi}}(\boldsymbol{x}) = \hat{\boldsymbol{s}}$ such that $\hat{s}_i = \rho_i(s_{\boldsymbol{\pi}(i)}), \forall i \in \{1, ..., d\}$, for a permutation $\boldsymbol{\pi}(\cdot)$ & an element-wise invertible $\rho_i(\cdot)$.

Applications: disentanglement, causal representation learning, self-supervised learning, etc.

Challenge: Non-identifiability
An infinite number of $(\boldsymbol{f}, \boldsymbol{s})$ can satisfy $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{s})$ in (1) 😟

Related Works

Identifiability challenge is notorious in **nonlinear ICA** (nICA): even with statistically independent $s_1, ..., s_d$, the model $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{s})$ is non-identifiable [1].

nICA with Auxiliary Variables $\boldsymbol{u}$. Side information (time frame labels, observation group indices, view indices, etc.) can help underpin identifiability of NMMI via [2]

$$p(\boldsymbol{s}|\boldsymbol{u}) = \prod_{i=1}^{d} p(s_i|\boldsymbol{u}). \tag{2}$$

🤔 Diverse auxiliary $\boldsymbol{u}$ not always available

nICA with Structured $\boldsymbol{f}$. Conformal/local isometry/post-nonlinear/piecewise-affine $\boldsymbol{f}$.

🤔 Structured $\boldsymbol{f}$ holds in limited applications

Structured Jacobian. $[\boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{s})]_{i,j} = \partial x_i/\partial s_j$ describes how x_i is influenced by s_j.

1. *Independent Mechanism Analysis*: orthogonal columns of $\boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{s})$ [3].
   🤔 Lacking global identifiability
2. *Structural Sparsity*: a sparsity pattern on $\boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{s})$, proposes to minimize $||\boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{s})||_1$ [4].
3. *Object-centric Learning*: a compositional $\boldsymbol{f}$ — a sparsely structured $\boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{s})$ where non-zero blocks corresponds to an object in image [5].
   🤔 $\boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{s})$ is non-sparse in many settings

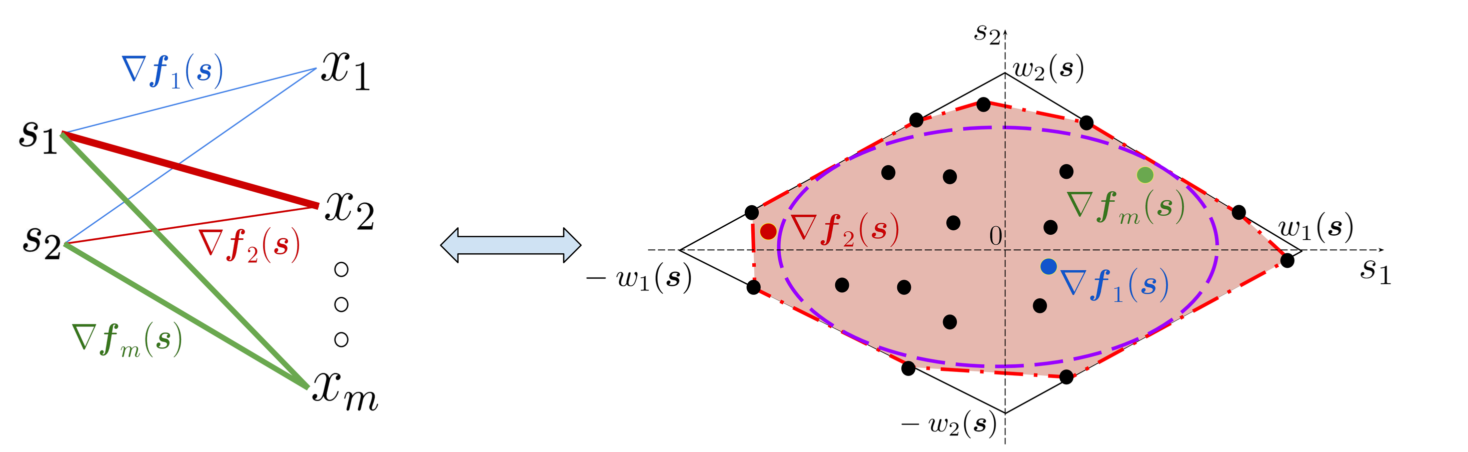
Sufficiently Diverse Influence (SDI) Condition

Assumption: Sufficiently Diverse Influence
At $\boldsymbol{s} \in \mathcal{S}$, there exists an $\boldsymbol{s}$-dependent weighted L_1-norm ball $\mathcal{B}_1^{\boldsymbol{w}(\boldsymbol{s})}$ such that $\nabla f_1(\boldsymbol{s}), ..., \nabla f_m(\boldsymbol{s}) \in \mathcal{B}_1^{\boldsymbol{w}(\boldsymbol{s})}$. In addition:

1. $\mathcal{E}(\mathcal{B}_1^{\boldsymbol{w}(\boldsymbol{s})}) \subseteq \mathrm{conv}\{\nabla f_1(\boldsymbol{s}), ..., \nabla f_m(\boldsymbol{s})\} \subseteq \mathcal{B}_1^{\boldsymbol{w}(\boldsymbol{s})}$,
2. $\mathrm{conv}\{\nabla f_1(\boldsymbol{s}), ..., \nabla f_m(\boldsymbol{s})\}^* \cap \mathrm{bd}(\mathcal{E}(\mathcal{B}_1^{\boldsymbol{w}(\boldsymbol{s})})^*) = \mathrm{extr}(\mathcal{B}_\infty^{\boldsymbol{w}(\boldsymbol{s})})$.

$\mathcal{S} \subset \mathbb{R}^d$: set of latent variables; $\mathcal{X} \subset \mathbb{R}^m$: set of observations; $\mathrm{conv}\{\cdot\}$: the convex hull of a set of vectors; $\mathcal{E}(\mathcal{P})$ is its MVIE of polytope $\mathcal{P}$; $\mathcal{P}^*$: polar set of $\mathcal{P}$; $\mathrm{extr}(\mathcal{P})$ extreme points of $\mathcal{P}$; $\mathrm{bd}(\mathcal{P})$: boundary of $\mathcal{P}$.

Illustration of Sufficiently Diverse Influence (SDI) Condition



Visualizing Sufficiently Diverse Influence: Row vectors $\nabla f_1(\boldsymbol{s}), ..., \nabla f_m(\boldsymbol{s})$ of $\boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{s})$ are sufficiently distinct — their convex hull contains MVIE of an L_1-norm ball $\mathcal{B}_1^{\boldsymbol{w}(\boldsymbol{s})}$.

Origin. SDI geometry originates from sufficiently-scattered condition (SSC) in NMF and PMF [6]; however, SSC characterizes the latent factors of a data matrix (e.g., $\boldsymbol{W}, \boldsymbol{H}$ in $\boldsymbol{X} = \boldsymbol{W}\boldsymbol{H}$), do not involve nonlinear functions or derivatives as in SDI.

Interpretation. SDI reflects how $\boldsymbol{s}$ diversely affects $x_1, ..., x_m$: some features are positively influenced by s_j ($\partial x_i/\partial s_j > 0$), others are negatively influenced ($\partial x_{i'}/\partial s_j < 0$).

- Statistically dependent $\boldsymbol{s}$ and dense $\boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{s})$ can satisfy SDI.
- SDI favors $m \gg d$ case (i.e., high-dim data with low-dim factors, say 📸/🎞️).
- SDI-satisfying geometric pattern of $\nabla f_1(\boldsymbol{s}), ..., \nabla f_m(\boldsymbol{s})$ can vary with $\boldsymbol{s}$.

Learning Criterion: Jacobian Volume Maximization

Using $\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}_{\boldsymbol{\phi}}$ as two neural networks for autoencoder architecture $\boldsymbol{x} = \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{g}_{\boldsymbol{\phi}}(\boldsymbol{x}))$.

$$\max_{\boldsymbol{\theta},\boldsymbol{\phi}} \quad \mathbb{E}[\log\det(\boldsymbol{J}_{\boldsymbol{f}_{\boldsymbol{\theta}}}(\boldsymbol{g}_{\boldsymbol{\phi}}(\boldsymbol{x}))^\top \boldsymbol{J}_{\boldsymbol{f}_{\boldsymbol{\theta}}}(\boldsymbol{g}_{\boldsymbol{\phi}}(\boldsymbol{x})))] \tag{3}$$

$$\text{s.t.} \quad ||\boldsymbol{J}_{\boldsymbol{f}_{\boldsymbol{\theta}}}(\boldsymbol{g}_{\boldsymbol{\phi}}(\boldsymbol{x}))_{i,:}||_1 \leq C, \quad \forall i = 1, ..., m, \tag{4}$$

$$\boldsymbol{x} = \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{g}_{\boldsymbol{\phi}}(\boldsymbol{x})), \quad \forall \boldsymbol{x} \in \mathcal{X} \tag{5}$$

- Objective (3) maximizes volume of convex hull of $\boldsymbol{J}_{\boldsymbol{f}_{\boldsymbol{\theta}}}(\hat{\boldsymbol{s}})$ spanned by its columns.
- Constraint (4) keeps the rows of $\boldsymbol{J}_{\boldsymbol{f}_{\boldsymbol{\theta}}}(\hat{\boldsymbol{s}})$ inside some L_1-norm ball, to comply with SDI.
- Constraint (5) keeps $\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}_{\boldsymbol{\phi}}$ invertible over d-dim manifold.

$\Rightarrow \partial \boldsymbol{x}/\partial s_1, ..., \partial \boldsymbol{x}/\partial s_d$ are encouraged to scatter in space (inside a certain L_1-norm ball).

Identifiability Results 🎯

Identifiability of DICA
Let an optimal solution be $(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\phi}})$. Assume $\bar{\boldsymbol{f}} = \boldsymbol{f}_{\bar{\boldsymbol{\theta}}}$ and $\bar{\boldsymbol{g}} = \boldsymbol{g}_{\bar{\boldsymbol{\phi}}}$ are universal function representers. Suppose the SDI condition holds for the NMMI model, for every $\boldsymbol{s} \in \mathcal{S}$. Then, $\bar{\boldsymbol{s}} = \bar{\boldsymbol{g}}(\boldsymbol{x}) = \bar{\boldsymbol{g}} \circ \boldsymbol{f}(\boldsymbol{s})$ where

$$[\bar{\boldsymbol{s}}]_i = [\bar{\boldsymbol{g}}(\boldsymbol{x})]_i = \rho_i(s_{\boldsymbol{\pi}(i)}), \ \forall i \in [d], \tag{6}$$

in which $\boldsymbol{\pi}$ is a permutation of $\{1, \ldots, d\}$ and $\rho_i(\cdot) : \mathbb{R} \to \mathbb{R}$ is an invertible function.

(Informal) Identifiability w/ Finite Number of SDI-Satisfying Points
Suppose each point in the finite set $\mathcal{S}_N := \{\boldsymbol{s}^{(1)}, ..., \boldsymbol{s}^{(N)}\}$ with $\mathcal{X}_N := \{\boldsymbol{x} \in \mathcal{X} : \boldsymbol{x} = \boldsymbol{f}(\boldsymbol{s}), \forall \boldsymbol{s} \in \mathcal{S}_N\}$ is SDI-satisfying. Under several regularity conditions, $\widehat{\boldsymbol{g}}(\boldsymbol{x}^{(n)}) = \widehat{\boldsymbol{\Pi}}\widehat{\boldsymbol{\rho}}(\boldsymbol{s}^{(n)}), \forall n \in [N]$ for a constant permutation $\widehat{\boldsymbol{\Pi}}$. With probability at least $1 - \delta$,

$$\mathbb{E}_{\boldsymbol{s}\sim p(\boldsymbol{s})}[||\widehat{\boldsymbol{g}}(\boldsymbol{x}) - \widehat{\boldsymbol{\Pi}}\widehat{\boldsymbol{\rho}}(\boldsymbol{s})||_2] = \mathcal{O}\left((L_{\boldsymbol{f}}L_{\widehat{\boldsymbol{g}}} + L_{\widehat{\boldsymbol{\rho}}})\mathcal{R}_N(\mathcal{G}) + \sqrt{\ln(1/\delta)/N}\right), \tag{7}$$

where $\mathcal{R}_N(\mathcal{G})$ is the empirical Rademacher complexity of the encoder class.

Implementation 🧑‍🍳

Given L realizations of $\boldsymbol{x}$, $\{\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(L)}\}$, use MLPs for $\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}_{\boldsymbol{\phi}}$. At t-th epoch, optimize

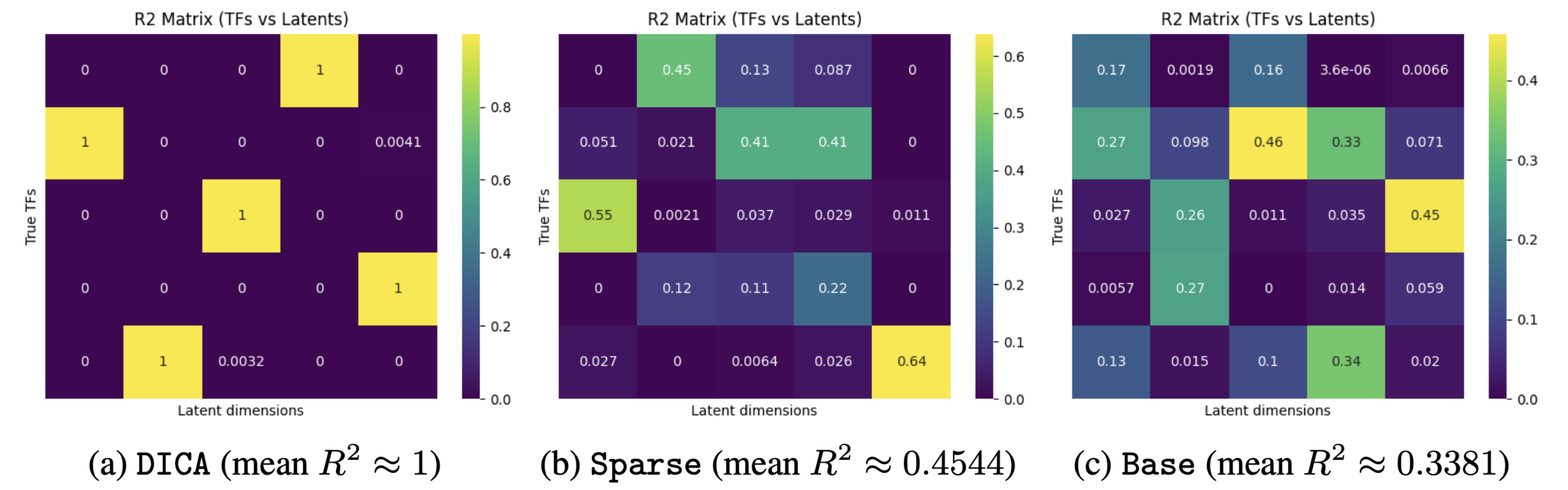
$$\min_{\boldsymbol{\theta},\boldsymbol{\phi}} \ \mathcal{L}_t := (1/L) \sum_{n=1}^{L} \left(||\boldsymbol{x}^{(n)} - \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{g}_{\boldsymbol{\phi}}(\boldsymbol{x}^{(n)}))||_2^2 - \lambda_{\mathrm{vol}}(t) \times c_{\mathrm{vol}} + \lambda_{\mathrm{norm}} \times c_{\mathrm{norm}}(t)\right) \tag{8}$$

using a warm-up heuristic with T_w warm-up epochs:

- $c_{\mathrm{vol}} := \log\det(\boldsymbol{J}_{\boldsymbol{f}_{\boldsymbol{\theta}}}(\boldsymbol{g}_{\boldsymbol{\phi}}(\boldsymbol{x}^{(n)}))^\top \boldsymbol{J}_{\boldsymbol{f}_{\boldsymbol{\theta}}}(\boldsymbol{g}_{\boldsymbol{\phi}}(\boldsymbol{x}^{(n)})))$ with $\lambda_{\mathrm{vol}}(t) := \frac{\lambda_{\mathrm{vol}}}{T_{\mathrm{w}}} \min\{t, T_{\mathrm{w}}\}$
  (a more computationally friendly trace-based surrogate of c_{vol} is available)
- $c_{\mathrm{norm}}(t) := \begin{cases} ||\boldsymbol{J}_{\boldsymbol{f}_{\boldsymbol{\theta}}}(\boldsymbol{g}_{\boldsymbol{\phi}}(\boldsymbol{x}^{(n)}))||_1 & \text{if } t \leq T_{\mathrm{w}} \\ \mathtt{Softplus}\{||\boldsymbol{J}_{\boldsymbol{f}_{\boldsymbol{\theta}}}(\boldsymbol{g}_{\boldsymbol{\phi}}(\boldsymbol{x}^{(n)}))||_1 - C\} & \text{if } t > T_{\mathrm{w}} \end{cases}$ with $\lambda_{\mathrm{norm}} > 0$,
  where C is average of last 10 epochs during warm-up.

Experiments (more in our paper) 📊

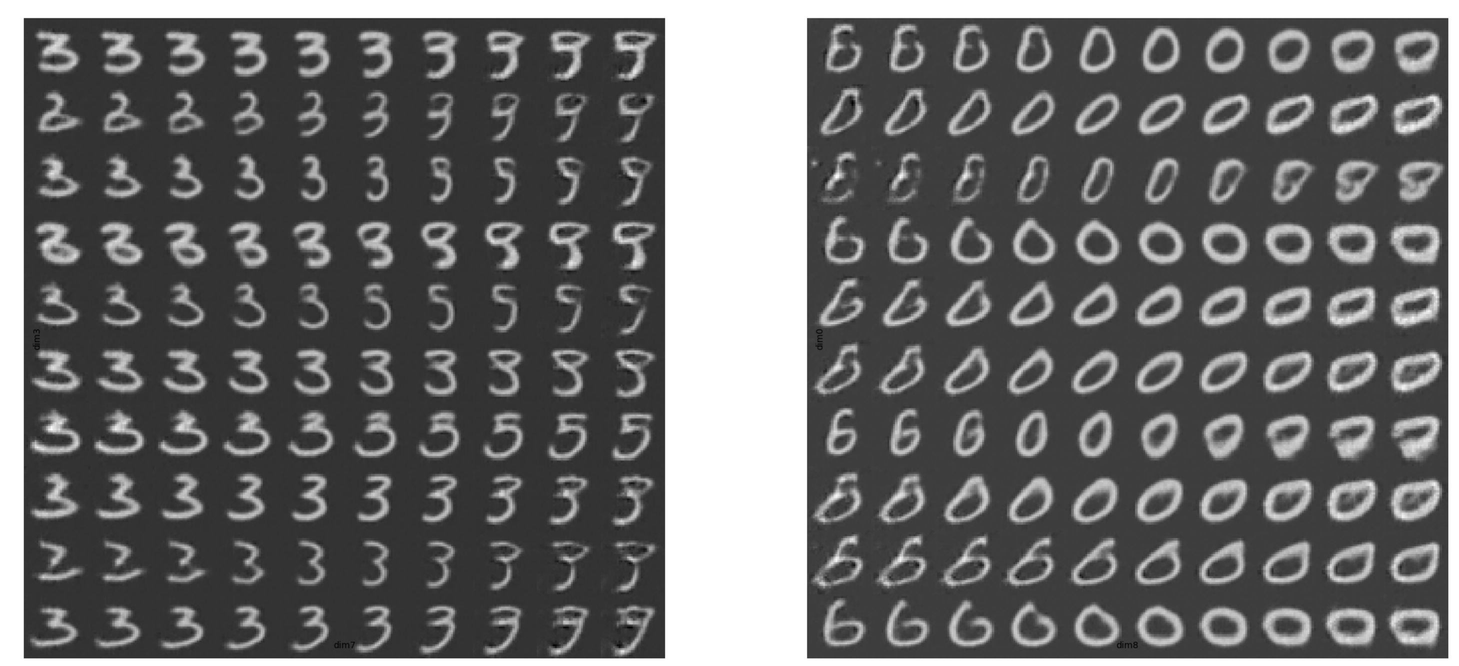
Single-cell Genomics 🧬. Inferring transcription factors' activities from gene expressions. Data generation follows SERGIO simulator, using TRRUST dataset + cross-talk noises.



(a) `DICA` (mean $R^2 \approx 1$) (b) `Sparse` (mean $R^2 \approx 0.4544$) (c) `Base` (mean $R^2 \approx 0.3381$)

Heatmap of R^2 scores between estimated components and ground-truth mRNA concentrations of TFs.

Image Disentanglement 📸. Applying autoencoder with DICA loss on MNIST dataset.



(a) [Anchor digit 3] As s_8 *increases*, digit 3 increasingly looks like digit 9. (b) [Anchor digit 0] As s_9 *decreases*, digit 0 increasingly looks like digit 6.

References:

[1] Hyvärinen & Pajunen, "Nonlinear independent component analysis: Existence and uniqueness results", Neural Networks, 1999.
[2] Hyvärinen et al., "Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning", AISTATS, 2019.
[3] Gresele et al., "Independent mechanism analysis, a new concept?", NeurIPS, 2021.
[4] Zheng & Zhang, "Generalizing Nonlinear ICA Beyond Structural Sparsity", NeurIPS, 2023.
[5] Brady et al., "Provably Learning Object-Centric Representations", ICML, 2023.
[6] Tatli & Erdogan, "Polytopic Matrix Factorization: Determinant Maximization Based Criterion and Identifiability", IEEE TSP, 2021.

Acknowledgements: This work is supported in part by the NSF CAREER Award ECCS-2144889.