Disentanglement via Mechanism Sparsity Regularization: A New Principle for Nonlinear ICA

Sébastien Lachapelle[1], Pau Rodríguez López[2], Yash Sharma[3], Katie Everett[4], Rémi Le Priol[1], Alexandre Lacoste[2], Simon Lacoste-Julien[1,5]

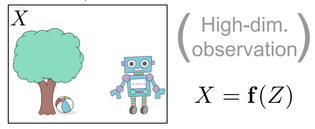
[1]Mila, Université de Montréal [2]ServiceNow Research [3]Tübingen AI Center, University of Tübingen [4]Google Reasearch [5]Canada CIFAR AI Chair

Contributions

- A new approach to achieve **nonlinear ICA/disentanglement**.
- Novel identifiability result based on the assumption that the **causal graph connecting the latent factors is typically sparse** [1, 2].
  - Objects usually interact sparsely with each others.
  - Actions typically affect only a few factors of variations.
- An estimation procedure which relies on the **variational autoencoder (VAE)** and a masking mechanism allowing to **regularize for sparsity**.
- As a special case of our framework, showing how **unknown-target interventions on the latent factors** can be leveraged to disentangle them, which is closely related to the **sparse mechanism shift hypothesis** introduced in [10].
- Illustrations of our theory with synthetic data.

1- Background

1.1 Motivating example



- $X \in \mathbb{R}^{d_x}$: High-dim. observation
- $Z \in \mathbb{R}^{d_z}$: Low-dim. latent representation
- $\mathbf{f} : \mathbb{R}^{d_z} \to \mathcal{X} \subset \mathbb{R}^{d_x}$: Invertible decoder
- $d_z \leq d_x$

1.2 Disentanglement

Definition (Disentanglement) Given a ground-truth model $\mathbf{f}$, a learned representation $\hat{\mathbf{f}}$ is *disentangled* if both are **permutation-equivalent**.

Definition (Permutation-equivalence)[a] [7]
Two models $\mathbf{f}$ and $\hat{\mathbf{f}}$ (representing the same data manifold $\mathcal{X}$) are *permutation-equivalent*, denoted by $\mathbf{f} \sim_p \hat{\mathbf{f}}$, when

$$\mathbf{f}^{-1}(x) = PD\tilde{\mathbf{f}}^{-1}(x) + b \;\; \forall x \in \mathcal{X}$$

where P is a permutation matrix and D an invertible diagonal matrix.

[a]Simplified definition, see paper for details.

1.3 Identifiability

- Unsupervised disentanglement is possible when $\mathbf{f}$ is identifiable up to $\sim_p$

$$\sim_p\textbf{-identifiability:} \;\; p(x) = \hat{p}(x)\; \forall x \implies \mathbf{f} \sim_p \hat{\mathbf{f}} \;\;,$$

where $p(x)$ and $\hat{p}(x)$ are the densities of X when the decoder is $\mathbf{f}$ and $\hat{\mathbf{f}}$, respectively.

- **Problem:** With the standard assumption of factor independence, the mixing function $\mathbf{f}$ is **not identifiable** from the distribution of X (without further assumptions) [3].
- **Recent development:** Identifiability of the latent factors is possible even with nonlinear mixing, as long as the latent variables are **conditionally independent given an observed auxiliary variable** [4, 7, 6].
- $A \in \mathbb{R}^{d_a}$: Observed auxiliary variable, e.g. an environment index or an action.

2- Disentanglement via Mechanism Sparsity

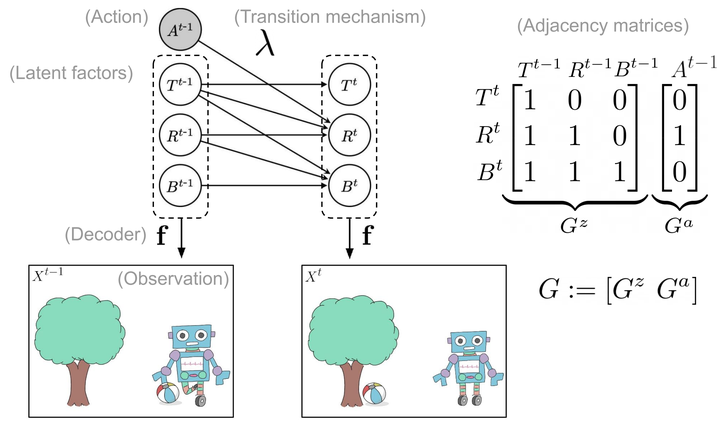


Figure 1: Model in the context of our motivating example

- We observe sequences $\{X^t\}_{t=1}^T$ and $\{A^t\}_{t=1}^T$ and have

$$X^t = \mathbf{f}(Z^t) + \texttt{noise}^t \,.$$

- We follow [7] and assume the Z_i^t **are independent given** $Z^{<t}$ **and** $A^{<t}$

$$p(z^t \mid z^{<t}, a^{<t}) = \textstyle\prod_{i=1}^{d_z} p(z_i^t \mid z^{<t}, a^{<t}) \,,$$

and assume each factor above lives in an **exponential family**, i.e.

$$p(z_i^t \mid z^{<t}, a^{<t}) = h_i(z_i^t) \exp\{\mathbf{T}_i(z_i^t)\boldsymbol{\lambda}_i(G_i^z \odot z^{<t}, G_i^a \odot a^{<t}) - \psi_i(z^{<t}, a^{<t})\} \,.$$

which, for example, includes Gaussian with fixed variance.

- $\mathbf{T}_i : \mathbb{R} \to \mathbb{R}$ is the *sufficient statistic* of Z_i^t
- $\boldsymbol{\lambda}_i$ outputs the *natural parameter*
- **Learnable parameters** are $\theta := (\mathbf{f}, \boldsymbol{\lambda}, G)$
- ψ_i is the normalizing constant

Theorem (Disentanglement via Mechanism Sparsity) Suppose we have two models with parameters $\theta = (\mathbf{f}, \boldsymbol{\lambda}, G)$ and $\hat{\theta} = (\hat{\mathbf{f}}, \hat{\boldsymbol{\lambda}}, \hat{G})$ representing the same distribution, i.e. $\mathbb{P}_{X^{\leq T} | a^{<T}; \theta} = \mathbb{P}_{X^{\leq T} | a^{<T}; \hat{\theta}}$ for all $a^{<T}$. Assume

1. The sufficient statistics $\mathbf{T}_i$ are strictly monotonic
2. **[Variability]** The mechanisms $\boldsymbol{\lambda}_i$ are "sufficiently complex" (see paper)
3. **[Sparsity]** $||\hat{G}||_0 \leq ||G||_0$
4. **[Graphical criterion]** For all $p \in \{1, ..., d_z\}$, there exist sets $\mathcal{I}, \mathcal{J} \subset \{1, ..., d_z\}$ and $\mathcal{L} \subset \{1, ..., d_a\}$ such that

$$\left(\bigcap_{i \in \mathcal{I}} \mathbf{Pa}_i^z\right) \cap \left(\bigcap_{j \in \mathcal{J}} \mathbf{Ch}_j^z\right) \cap \left(\bigcap_{\ell \in \mathcal{L}} \mathbf{Ch}_\ell^a\right) = \{p\} \,,$$

where $\mathbf{Pa}_i^z$ and $\mathbf{Ch}_i^z$ are the sets of parents and children of node z_i in G^z, respectively, while $\mathbf{Ch}_\ell^a$ is the set of children of a_ℓ in G^a.

Then, $\mathbf{f}$ and $\hat{\mathbf{f}}$ are **permutation-equivalent**, i.e. the model $\hat{\mathbf{f}}$ is **disentangled**.

3- Actions as Unknown-Target Interventions

- Suppose $A = \vec{0}$ corresponds to the **observational distribution**
- and $A = \vec{e}_\ell$ corresponds to the **ℓth intervention** (one-hot vector)
- Then, $G_{i,\ell}^a = 1 \iff z_i$ is targeted by the ℓth intervention. (G^a must be learned!)

4- Experiments

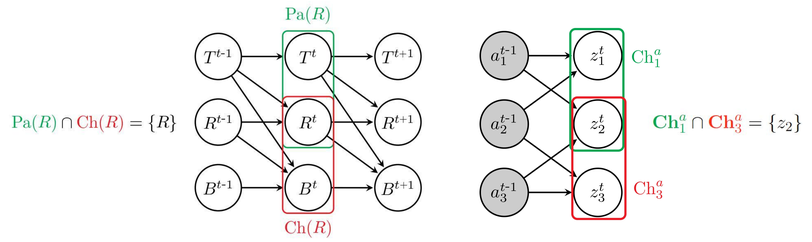


Figure 2: Two graphs that satisfies the **graphical criterion**.

4.1 Learning with regularized VAE

- We use the **VAE framework** [8] with a variational approximate posterior given by:

$$q(z^{\leq T} \mid x^{\leq T}, a^{<T}) := \textstyle\prod_{t=1}^T q(z^t \mid x^t) \,.$$

- This yields the following **evidence lower bound (ELBO)**: $\log p(x^{\leq T} | a^{<T}) \geq$

$$\textstyle\sum_{t=1}^T \mathop{\mathbb{E}}_{Z^t \sim q(\cdot|x^t)} [\log p(x^t \mid Z^t)] + \mathop{\mathbb{E}}_{Z^{t-1} \sim q(\cdot|x^{t-1})} KL(q(Z^t \mid x^t) || p(Z^t \mid Z^{t-1}, a^{t-1})) \,.$$

- We model the mechanisms $\boldsymbol{\lambda}_i$ with MLPs
- We add $-\alpha_z ||G^z||_0$ and $-\alpha_a ||G^a||_0$ to the objective to **regularize for sparsity**.
- The binary masks G^z and G^a are treated as random to allow for optimization via SGD using the **Gumbel-Softmax trick** [9, 5].

4.2 Experiments on synthetic data ($d_x = 20$, $d_z = 10$)

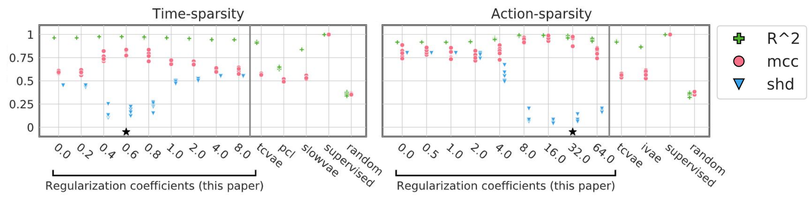


Figure 3: Positive effect of sparsity regularization on two datasets. **Left.** Data generated from a graph like the left one of Fig 2. **Right.** Data generated from a graph like the right one Fig 2.

- **MCC** = measure of disentanglement [4] (higher is better)
- **SHD** = distance to ground-truth graph (lower is better)

References

[1] Y. Bengio. "The Consciousness Prior". In: *arXiv preprint arXiv:1709.08568* (2019).
[2] A. Goyal et al. "Recurrent Independent Mechanisms". In: *ICLR*. 2021.
[3] A. Hyvärinen and P. Pajunen. "Nonlinear independent component analysis: Existence and uniqueness results". In: *Neural Networks* (1999).
[4] A. Hyvärinen, H. Sasaki, and R. E. Turner. "Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning". In: *AISTATS*. 2019.
[5] E. Jang, S. Gu, and B. Poole. "Categorical Reparameterization with Gumbel-Softmax". In: *ICML* (2017).
[6] I. Khemakhem et al. "ICE-BeeM: Identifiable Conditional Energy-Based Deep Models Based on Nonlinear ICA". In: *NeurIPS*. 2020.
[7] I. Khemakhem et al. "Variational Autoencoders and Nonlinear ICA: A Unifying Framework". In: *AISTATS*. 2020.
[8] D. P. Kingma and M. Welling. "Auto-Encoding Variational Bayes". In: *ICLR*. 2014.
[9] C. J. Maddison, A. Mnih, and Y. W. Teh. "The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables". In: *ICML* (2017).
[10] B. Schölkopf et al. "Toward Causal Representation Learning". In: *Proceedings of the IEEE - Advances in Machine Learning and Deep Neural Networks* (2021).