A Sparsity Principle for Partially Observable Causal Representation Learning

Danru Xu[1] **Dingling Yao**[2,3] **Sébastien Lachapelle**[4] **Perouz Taslakian**[6] **Julius von Kügelgen**[7]
Francesco Locatello[2] **Sara Magliacane**[1]
[1] University of Amsterdam
[2] Institute of Science and Technology Austria
[3] Max Planck Institute for Intelligent Systems, Tübingen, Germany
[4] Mila, Université de Montréal, Samsung - SAIT AI Lab, Montreal
[6] ServiceNow Research, Montreal
[7] Seminar for Statistics, ETH Zürich

Abstract

Causal representation learning aims at identifying high-level causal variables from perceptual data. Most methods assume that *all* latent causal variables are captured in the high-dimensional observations. We instead consider a *partially observed* setting, in which each measurement only provides information about a subset of the underlying causal state. Prior work has studied this setting with multiple domains or views, each depending on a *fixed* subset of latents. Here we focus on learning from *unpaired* observations from a dataset with an *instance-dependent* partial observability pattern. Our main contribution is to establish two identifiability results for this setting: one for linear mixing functions without parametric assumptions on the underlying causal model, and one for piecewise linear mixing functions with Gaussian latent causal variables. Based on these insights, we estimate the underlying causal variables by enforcing sparsity in the inferred representation. Based on these insights, we propose two methods for estimating the underlying causal variables by enforcing sparsity in the inferred representation.

1 INTRODUCTION

Traditional causal inference methods assume that the causal variables are given a priori, but in many real-world settings, we only have unstructured, high-dimensional observations of a causal system. Motivated by this shortcoming, causal representation learning [CRL; Schölkopf et al., 2021] aims to infer high-level causal variables from low-level data such as images. A popular approach to identify (i.e., provably recover) high-level latent variables is (nonlinear) independent component analysis (ICA) [Hyvarinen and Morioka, 2016, 2017, Hyvarinen et al., 2019, Khemakhem et al., 2020], which aims to recover independent latent factors from entangled measurements. Several works generalize this setting to the case in which the latent variables can have causal relations [Yao et al., 2022, Brehmer et al., 2022, Lippe et al., 2022, 2023, Ahuja et al., 2023a,b, Lachapelle et al., 2022, 2023, 2024, von Kügelgen et al., 2021, 2023, Wendong et al., 2023, Squires et al., 2023, Buchholz et al., 2023, Zhang et al., 2023], establishing various identifiability results under different assumptions on the available data and the generative process. However, most existing works assume that *all* causal variables are captured in the high-dimensional observations. Notable exceptions include Sturma et al. [2023] and Yao et al. [2023] who study *partially observed* settings with multiple domains (datasets) or views (tuples of observations), respectively, each depending on a *fixed subset* of the latent variables.

In this work, we also focus on learning causal representations in such a *partially observed* setting, where not necessarily all causal variables are captured in any given observation. Our setting differs from prior work in two key aspects: (i) we consider learning from a dataset of *unpaired partial* observations; and (ii) we allow for *instance-dependent* partial observability patterns, meaning that each measurement depends on an unknown, varying (rather than fixed) subset of the underlying causal state.

This setting is motivated by real-world applications in which we cannot at all times observe the complete state of the environment, e.g., because some objects are moving in and out of frame, or are occluded. As a motivating example, consider a stationary camera that takes pictures of a parking lot on different days as shown in Fig. 1a. On different days, different cars are present in the parking lot, and the same car can be parked in different spots. Our task is to recover the position for each car that is present in a certain image. In this setting, we only have one observation for a given state of the system (i.e., one image per day), and the subsets of causal variables that are measured in the observation (the parked cars), change dynamically across images. In particular, we formalize the *Unpaired Partial Observations* setting for

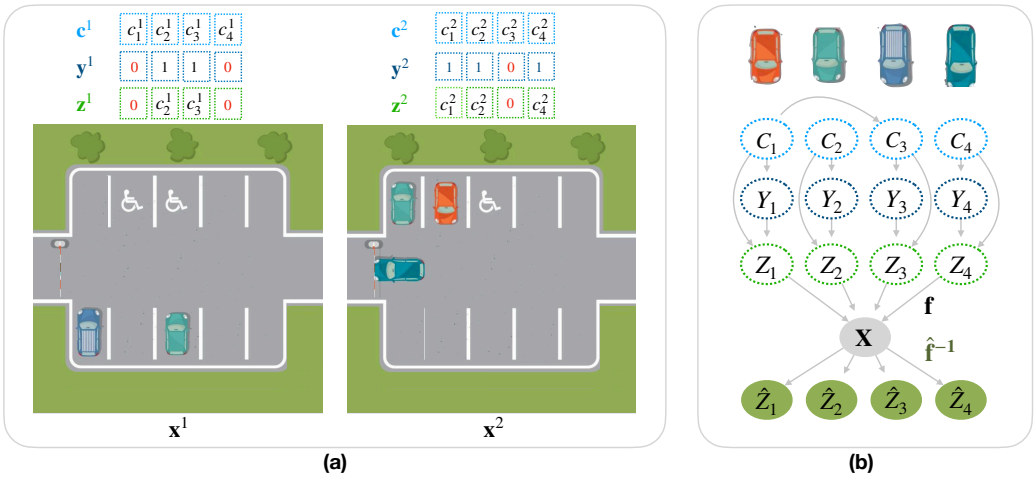

Figure 1: **(a)** Motivating example for the *Unpaired Partial Observation* setting: a stationary camera taking pictures of a car park. We consider $\mathbf{x}^1$ the image on day 1 and $\mathbf{x}^2$ the image on day 2. The latent causal variables $\mathbf{c}^1$ and $\mathbf{c}^2$ represent the positions of four cars on each day. In $\mathbf{x}^1$ only $Car2$ and $Car3$ are visible, while in $\mathbf{x}^2$ all cars except $Car3$ are visible. This is represented by the ones in the binary mask variables $\mathbf{y}^1$ and $\mathbf{y}^2$. The combination of the values of the latent causal variables $\mathbf{c}$ and the masks $\mathbf{y}$ are the *masked causal variables* $\mathbf{z}$, which used by the mixing function $\mathbf{f}$ to generate the images $\mathbf{x}$. **(b)** Causal model of the setting, the dotted line variables are not directly observed, but they are measured only through the observation $\mathbf{X}$. Our goal is to learn a representation $\hat{\mathbf{Z}}$ that identifies $\mathbf{Z}$ up to permutation and element-wise transformation.

CRL, where each *partial observation* captures only a subset of causal variables and the observations are unpaired, i.e., we do not have simultaneous partial observations of the same state of the system. We introduce two theoretical results for identifying causal variables up to permutation and element-wise transformation under partial observability. Both results leverage a sparsity constraint. The full version of this work [Xu et al., 2024] was accepted at ICML 2024.

2 IDENTIFIABILITY VIA SPARSITY

In this section, we briefly introduce the two theoretical results of how a simple sparsity constraint on the learned representations allows us to identify the ground truth variables *up to permutation and element-wise linear transformation*.

The first theorem proves identifiability for linear mixing function and without parametric assumptions on the underlying causal model. We show that, with some mild assumptions, for linear mixing functions under a perfect reconstruction, a simple *sparsity constraint* on the learned representation allows us to learn a disentangled representation of the ground truth latent variables.

Since linearity of mixing function is a strong assumption that may not hold in many applications, we consider exploring more for nonlinear cases. As a first step, we consider a piecewise linear mixing function and assume that masks are independent from the causal variables. Then the second theorem proves identifiability for the *piecewise linear mixing function* when the causal variables are Gaussian, and we can group observations by their partial observability patterns.

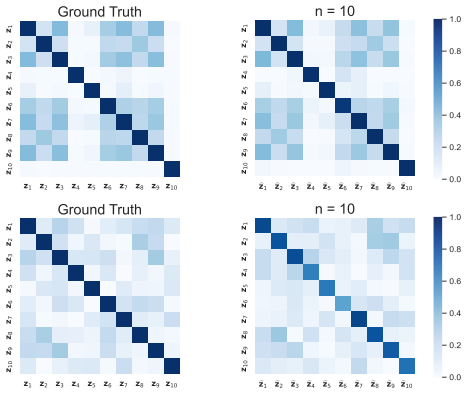

Figure 2: Linear Gaussian causal model with 10 latents. First row: linear $\mathbf{f}$, second row: piecewise linear $\mathbf{f}$.

3 EXPERIMENTAL RESULTS

We validate the theorems with experiments on simulated data. Here we provide a representative result for each theorem in Figure 2, where the first row is for linear $\mathbf{f}$ and second row is for piecewise linear $\mathbf{f}$. The heatmaps on the left are the correlation matrices of ground truth latents and the heatmaps on the right are the correlation matrices of learned representations and ground truth latents. Intuitively, the more similar the right heatmap is with the left one, the learned representations are closer to the ground truth.

References

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International Conference on Machine Learning*, pages 372–407. PMLR, 2023a.

Kartik Ahuja, Amin Mansouri, and Yixin Wang. Multi-domain causal representation learning via weak distributional invariances. In *Causal Representation Learning Workshop at NeurIPS 2023*, 2023b.

Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S Cohen. Weakly supervised causal representation learning. *Advances in Neural Information Processing Systems*, 35: 38319–38331, 2022.

Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. In *Advances in Neural Information Processing Systems*, 2023.

Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in neural information processing systems*, 29, 2016.

Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ica of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pages 460–469. PMLR, 2017.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. In *Conference on Causal Learning and Reasoning*, pages 428–484. PMLR, 2022.

Sébastien Lachapelle, Tristan Deleu, Divyat Mahajan, Ioannis Mitliagkas, Yoshua Bengio, Simon Lacoste-Julien, and Quentin Bertrand. Synergies between disentanglement and sparsity: Generalization and identifiability in multi-task learning. In *International Conference on Machine Learning*, pages 18171–18206. PMLR, 2023.

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Nonparametric partial disentanglement via mechanism sparsity: Sparse actions, interventions and sparse temporal dependencies, 2024.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Stratis Gavves. Citris: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning*, pages 13557–13603. PMLR, 2022.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. Biscuit: Causal representation learning from binary interactions. *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, 2023.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Chandler Squires, Anna Seigal, Salil Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In *40th International Conference on Machine Learning*, 2023.

Nils Sturma, Chandler Squires, Mathias Drton, and Caroline Uhler. Unpaired multi-domain causal representation learning. *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34: 16451–16467, 2021.

Julius von Kügelgen, Michel Besserve, Wendong Liang, Luigi Gresele, Armin Kekić, Elias Bareinboim, David M Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. In *Advances in Neural Information Processing 36*, 2023.

Liang Wendong, Armin Kekić, Julius von Kügelgen, Simon Buchholz, Michel Besserve, Luigi Gresele, and Bernhard Schölkopf. Causal component analysis. In *Advances in Neural Information Processing Systems*, 2023.

Danru Xu, Dingling Yao, Sébastien Lachapelle, Perouz Taslakian, Julius von Kügelgen, Francesco Locatello, and Sara Magliacane. A sparsity principle for partially observable causal representation learning. *International conference on machine learning*, 2024.

Dingling Yao, Danru Xu, Sébastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. *arXiv preprint arXiv:2311.04056*, 2023.

Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. Learning temporally causal latent processes from general temporal data. In *International Conference on Learning Representations*, 2022.

Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. In *Advances in Neural Information Processing Systems*, 2023.