Toward a more transparent causal representation learning

Mouad El Bouchattaoui[1,2] **Myriam Tami**[1] **Benoit Lepetit**[2] **Paul-Henry Cournède**[1]

[1]Paris-Saclay University, CentraleSupélec, MICS Lab, Gif-sur-Yvette, France
[2]Saint-Gobain, Paris, France

Abstract

This work addresses the challenge of causal representation learning (CRL) for complex, high-dimensional, time-varying data. We enhance transparency and confidence in learned causal abstractions by linking them to observational space. The existing literature rarely explores the association between latent causal variables and observed ones, with only one notable work imposing a simplistic single-latent-factor decoding constraint. Our approach, in contrast, allows for a flexible entangling of latent factors, reflecting the complexity of real-world datasets. We introduce a structural sparsity pattern over generative functions and leverage induced grouping structures over observed variables for better model understanding. Our regularization technique, based on sparse subspace clustering over the Jacobian matrix of the decoder, promotes the sparsity and readability of model results. We apply our model to real-world datasets, including Saint-Gobain purchase data and MIMIC III medical data.

Introduction: Complex high-dimensional (HD) data is abundant in many real-world disciplines and has garnered significant interest, particularly in learning high-level generative latent factors that are causally related and generate lower-level observable variables, i.e., causal representation learning (CRL) Schölkopf et al. [2021]. Efforts have been directed toward causally disentangling these latent variables, whether unconditionally or conditionally, on auxiliary variables like time indices and labels Yao et al. [2022], Komanduri et al. [2023], Song et al. [2024]. Text data, images, and videos have received extensive attention in interpreting the semantics behind changes in low-level data due to variations in high-level factors. However, it is a more difficult task with tabular, HD, and evolving data. For example, retail data on customer purchases forms an HD vector representing numerous products and their purchase states. In large retail corporations, this dimension can reach 10^5, far exceeding the complexity of datasets like 28x28 video data. Inferring causal structures in such HD settings is computationally demanding, making it difficult for experts to inspect causal edges among thousands of variables over time. Seeking causal structure on a less granular but more abstract level is a natural way to circumvent these bottlenecks. However, enhancing transparency and understanding of causal models is crucial for advancing CRL beyond simple academic validation routines (e.g., relatively simple semi-synthetic datasets), fostering adoption and confidence among non-experts and decision-makers.

This work addresses the "*so what?*" question after performing CRL. This involves effectively conveying CRL results for HD complex data and enhancing confidence in learned causal abstractions by transparently relating them to the highly granular observational space. Our approach is twofold: First, we apply a structural sparsity constraint over the decoding function, i.e., the mixing function of latent causal variables that generate the low-level observed variables corresponding to the input. Results on the identifiability of the causal variables and the mixing function will be presented in the workshop. Second, a byproduct of the enforced sparsity structure between latent and observed variables is that many observed features relate to one or a few latent factors, enabling observed features clustering. We leverage this induced grouping structure to enhance the model's transparency by analyzing the content and coherence of the clusters. This process helps relate the groups of observed variables to known labels unsolicited during modeling, such as product categories and sub-categories in purchase behavior.

Technically, we assume an anchor feature property, i.e., each latent factor has at least two exclusively related observed features, ensuring the possibility of latent factors mixing for other observed features. We design a regularization technique to impose this sparsity constraint over the generative model by explicitly constraining the Jacobian of the mixing

function. We leverage sparse subspace clustering Elhamifar and Vidal [2013] on the Jacobian matrix, enabling the expression of gradients with respect to anchor features as linear combinations of gradients from other anchor features related to the same latent factor. This approach decomposes the span of Jacobian columns into linear subspaces, allowing each gradient to be represented sparsely, thereby enforcing the sparsity constraint.

Related Work: CRL over time-varying data, aiming to understand the generation mechanism relating latent causal variables to observed ones, is scarce. One noteworthy paper is Moran et al. [2021], which implements a Sparse VAE with a sparsity constraint over the decoder, similar to our approach. However, this work focuses on static data and does not learn a causal structure in the latent space. To induce a sparse decoder, they applied a binary mask over the latent variables before mapping them to the observed space, assuming a sparse prior over the mask. It is unclear how such a sparse general prior truly reflects the specific sparsity constraint, making theoretical assumptions poorly encoded in the inference method. Other papers, such as Zheng et al. [2022], Zheng and Zhang [2023], used similar sparsity constraints but were mainly concerned with non-linear ICA in static settings, enforcing the sparsity constraint using either the L_1 norm or a hybrid penalty of L_0 and L_1. Two important papers that included CRL and sparsity are worth mentioning. First, Lachapelle et al. [2022] suggests learning a causal representation over time-varying data with a sparse mechanism shift, i.e., a few edges relating past latent representations to future ones. However, no sparse mechanism was applied to the mapping between latent and observed features. Secondly, the most similar and concurrent work to ours is Boussard et al. [2023], Brouillard et al. [2024], which enforced a sparsity constraint over the decoding function while carrying CRL for time series climate data. Yet, they supposed the *single parent encoding* as sparsity constraints, i.e., for each feature, *only one single parent* generates it. This single-parent assumption is very constraining as it doesn't allow a mixture of latent factors to generate observed variables, which could happen when causal relationships among features in the observed space lead to multiple latent parents. Our work stands out as a generalization since we allow for multiple parents in the latent space, which is a less restrictive and more plausible assumption for various datasets beyond climatology, typically purchase and medical data. We also prove the identifiability of the causal variables and the mixing function under this more general framework.

Modeling: Let $\mathbf{x}_t \in \mathbb{R}^{d_x}$ denote the observed vector at time t, generated from latent factors $\mathbf{z}_t$ via the function $\mathbf{f}$, i.e., $\mathbf{x}_t = \mathbf{f}(\mathbf{z}_t)$. We briefly describe the latent causal process $(\mathbf{z}_t)_{t\geq 1}$. For each time step t, the latent factors $\{z_t^j\}_{j=1}^{d_z}$ are assumed to be mutually independent given $\mathbf{z}_{<t}$. Each latent factor is expressed as:

$$z_t^j = \mathbf{f}_i(\{z_{k,t-\tau} | z_{k,t-\tau} \in Pa(z_{it}), \tau = 1, \dots, L\}, \epsilon_{it}) \quad \epsilon_{it} \sim p_{\epsilon_i}$$

$Pa(z_t^j)$ denotes the **latent factors** that are parents of the j-th dimension in $\mathbf{z}_t$, i.e., the direct causes, a subset of dimensions of lagged L factors $\mathbf{z}_{t-1}, \dots, \mathbf{z}_{t-L}$.

Sparse subspace clustering relies on the self-expressiveness property Elhamifar and Vidal [2013], enabling the representation of a gradient $\nabla_{\mathbf{z}} \mathbf{f}_j$ as a linear combination of other gradients for $j = 1, \dots, d_x$. Let there be n linear subspaces $(S_l)_{l=1}^n$ of $\mathbb{R}^{d_z}$ such that $\nabla_{\mathbf{z}} \mathbf{f}_j$ belongs to their union. Mathematically, we define the Jacobian matrix $Jac(\mathbf{f}) = [\nabla_{\mathbf{z}} \mathbf{f}_1, \dots, \nabla_{\mathbf{z}} \mathbf{f}_{d_x}] = [J_1, \dots, J_n]P$, where $P \in \mathbb{R}^{d_x \times d_x}$ is a permutation matrix. Each J_l represents a matrix of gradient vectors, with the span of its columns defining S_l. The self-expressiveness property is expressed by $\nabla_{\mathbf{z}} \mathbf{f}_j = Jac(\mathbf{f}) C_{:,j}$, with $C_{jj} = 0$ to eliminate the trivial identity solution. While this expression is generally not unique, our goal is to choose a solution that enforces sparsity in matrix C, with nonzero entries in $C_{:,j}$ s.t when $\nabla_{\mathbf{z}} \mathbf{f}_j$ belongs to an S_l; it can be expressed by gradients from the same subspace S_l. To achieve this, we can optimize for each j, $\min ||C_{:,j}||_1$ subject to $\nabla_{\mathbf{z}} \mathbf{f}_j = Jac(\mathbf{f}) C_{:,j}$ and $C_{jj} = 0$, or with a matrix formulation $\min ||C||_1$ subject to $Jac(\mathbf{f}) = Jac(\mathbf{f})C$ and $diag(C) = 0$.

Alternatively, we can minimize $\min_C ||C||_1 + \lambda ||Jac(\mathbf{f}) - Jac(\mathbf{f})C||_F^2$. In the context of CRL with structural constraints, the objective function becomes:

$$\min_{C,\Theta} \mathcal{L}(\Theta) + ||C||_1 + \lambda ||Jac(\mathbf{f}) - Jac(\mathbf{f})C||_F^2$$

where $\mathcal{L}(\Theta)$ refers to the main loss of the CRL problem, typically an ELBO and $f \in \Theta$.

We finally link latent factors $\mathbf{z}_t$ to $\mathbf{x}_t$ by estimating the matrix C, which encodes sparsity over $Jac(\mathbf{f})$. Details about the clustering method and its consequences will be presented in the workshop.

Experiments and results: Our experiments cover diverse datasets: real-world data from Saint-Gobain and the MIMIC III medical data Johnson et al. [2016], alongside semi-synthetic data. We aim to unveil latent factors driving client purchases in Saint Gobain and understand vitals in MIMIC III. We validate our methodology across varied data settings through synthetic data experiments, ensuring robustness and generalizability. Results will be presented at the workshop.

Conclusion: Our study delves into CRL and the challenges of HD data. By analyzing datasets from Saint-Gobain and MIMIC medical records, we aim to bridge theory with practical application, fostering transparency and confidence in causal models by leveraging structural sparsity constraints faithfully encoded in our work. We aim to further consolidate our theory by showing the statistical guarantees for convergence and consistency.

References

Julien Boussard, Chandni Nagda, Julia Kaltenborn, Charlotte Emilie Elektra Lange, Philippe Brouillard, Yaniv Gurwicz, Peer Nowack, and David Rolnick. Towards causal representations of climate model data. *arXiv preprint arXiv:2312.02858*, 2023.

Philippe Brouillard, Sebastien Lachapelle, Julia Kaltenborn, Yaniv Gurwicz, Dhanya Sridhar, Alexandre Drouin, Peer Nowack, Jakob Runge, and David Rolnick. Causal representation learning in temporal data via single-parent decoding, 2024. URL https://openreview.net/forum?id=e2jDr8NdJm.

Ehsan Elhamifar and René Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence*, 35 (11):2765–2781, 2013.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Aneesh Komanduri, Yongkai Wu, Feng Chen, and Xintao Wu. Learning causally disentangled representations via the principle of independent causal mechanisms. *arXiv preprint arXiv:2306.01213*, 2023.

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.

Gemma E Moran, Dhanya Sridhar, Yixin Wang, and David M Blei. Identifiable deep generative models via sparse decoding. *arXiv preprint arXiv:2110.10804*, 2021.

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.

Xiangchen Song, Weiran Yao, Yewen Fan, Xinshuai Dong, Guangyi Chen, Juan Carlos Niebles, Eric Xing, and Kun Zhang. Temporally disentangled representation learning under unknown nonstationarity. *Advances in Neural Information Processing Systems*, 36, 2024.

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. URL https://openreview.net/forum?id=RDlLMjLJXdq.

Yujia Zheng and Kun Zhang. Generalizing nonlinear ica beyond structural sparsity. *Advances in Neural Information Processing Systems*, 36:13326–13355, 2023.

Yujia Zheng, Ignavier Ng, and Kun Zhang. On the identifiability of nonlinear ica: Sparsity and beyond. *Advances in Neural Information Processing Systems*, 35:16411–16422, 2022.