Recovering Causal Features for Instrumental Variable Regression with Contrastive Learning

Gabin Agbalé[1,2] **Stefan Harmeling**[1,3] **Alexander Marx**[1,2]
[1]TU Dortmund University, Germany
[2]Research Center for Trustworthy Data Science and Security, UA Ruhr
[3]Lamarr Institute for Machine Learning and Artificial Intelligence
{gabin.agbale,stefan.harmeling,alexander.marx}@tu-dortmund.de

Abstract

Instrumental Variable (IV) regression is a standard technique for estimating causal effects in the presence of unobserved confounders. The classic IV setting assumes access to an external variable - called the instrument - which directly influences the treatment variable. In this work, we consider a more challenging yet realistic setting where the treatment is latent, and we can only observe a nonlinear transformation of it (e.g. an image). To overcome this problem, we leverage insights from the Independently Modulated Component Analysis (IMCA), which is a framework that relaxes the independence assumption in Independent Component Analysis (ICA). Specifically, we propose a general contrastive learning framework to recover the latent treatment up to an affine transformation which may be related to the instrument by a (non-)linear function. We prove that the recovered representation is compatible with standard IV techniques. Empirically, we demonstrate the effectiveness of our method using control function and two-stage least squares (2SLS) estimators and evaluate the robustness of the learned estimators in distribution shift setting.

1 Introduction

Conventional supervised learning techniques, such as ordinary least squares (OLS), are widely used to model relationships between features and outcomes. To correctly capture causal effects of the predictors, these methods rely on the assumption that the residuals of the target variable are independent of the features. This assumption, however, does not generally hold. Consider a setting where we observe a treatment X and an outcome Y which can be expressed as $Y = f_0(X) + \varepsilon$, with $\mathbb{E}[\varepsilon] = 0$ but $\mathbb{E}[\varepsilon|X] \neq 0$. Such a data generative mechanism violates the standard assumption that the noise is independent of the features, leading to $\mathbb{E}[Y|X] \neq f_0(X)$. Thus, classical supervised learning methods fail to recover the true causal effect. To address this, *Instrumental Variable* (IV) regression [Imbens and Angrist, 1994] assumes the observation of an *instrument* that affects the outcome only through the treatment variable and is thus independent from the residuals. While originally formulated for linear functions f_0, nonparametric approaches to IV regression [Newey and Powell, 2003, Ai and Chen, 2003, Darolles et al., 2011] have emerged.

Similarly, the field of causal representation learning (CRL) [Schölkopf et al., 2021] often relies on an observed auxiliary variable to learn representations that are suitable for performing causal downstream tasks. It is closely related to nonlinear independent component analysis (ICA), which aims to recover independent sources from nonlinearly mixed signals. A central question in ICA is that of identifiability—whether the sources can be recovered from observational data alone. This task is provably impossible without additional assumptions on the data-generating process [Hyvärinen and Pajunen, 1999]. Just as an instrumental variable enables identification of a causal effect, iden-

Accepted to the NeurIPS 2025 Workshop on CauScien: Uncovering Causality in Science.

tifiability of nonlinear ICA can be achieved by introducing an auxiliary variable [Hyvärinen et al., 2019, Khemakhem et al., 2020a], under the assumption that the latent variables are independent conditioned on the auxiliary. This setting provides a more natural theoretical foundation for CRL, where the underlying features are causally related and thus not independent. We show that these assumptions are fulfilled by standard assumptions in the IV setting, allowing us to recover latent features for IV that we can also leverage to perform extrapolation tasks.

We consider a *representation-based* IV setting in which the treatment variable is latent and causally related to an observed auxiliary variable. We show that, under assumptions compatible with the IV setting, the latent variables are recoverable up to an affine transformation, which is sufficient to recover the causal effect between the observed transformation of the latent treatment and the target. As a practical instantiation of this framework, we propose two variants of contrastive methods that provably recover the latent treatment (up to some indeterminacy) and demonstrate its consistency with IV methods such as *two-stage least squares* (2SLS) and the *Control Function* (CF) approach.[1]

1.1 Related Work

Other IV methods [J. Hartford and Taddy, 2017] extend the classic *two-stage least squares* approach to nonlinear settings using neural networks. Similarly, Liyuan Xu [2021] propose a method that alternates between the first and second stage, using dictionary learning to extract features from the instruments. An other line of work [Zhang et al., 2023, A. Bennett and Schnabel, 2019, Saengkyongam et al., 2024] aims to recover a function f that satisfies appropriate *moment conditions*, typically expressed as $\mathbb{E}[h(A)(Y - f(X))] = 0$ for every measurable function h, in order to enforce the independence of the residuals from instrument A. However, this leads to an infinite family of conditions, making it impractical to verify directly. As a result, one must choose a function class for h that is expressive enough to approximate the space of all measurable functions and ensure that the moment condition sufficiently captures the independence constraint. As for nonlinear ICA, there exist approaches that ensure identifiability without relying on an auxiliary variable. In the context of time-series data, temporal dependencies have been shown to enable identifiability [Hyvärinen and Morioka, 2016]. Structural assumptions on the mixing function, such as additivity [Lachapelle et al., 2023] or sparsity [Lachapelle et al., 2022, Moran et al., 2022], have also been proposed. Without such auxiliaries, weaker results such as block-identifiability when we have access to style transforming data augmentations [Von Kügelgen et al., 2021] or when working with multimodal data [Daunhawer et al., 2023].

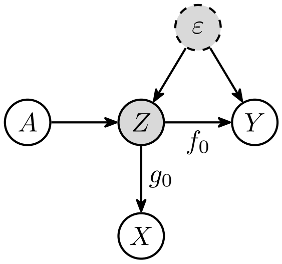
Building on this line of work, Khemakhem et al. [2020b] proposed a novel framework called Independently Modulated Component Analysis (IMCA), which further relaxes the conditional independence assumption while retaining most of the identifiability guarantees established in the auxiliary variable setting.

2 Recovering Latent Treatments

2.1 Data Generating Process

In contrast with the classic IV setting introduced previously, We consider a *representation-based* variant where the treatment $Z \in \mathcal{Z} \subset \mathbb{R}^{d_{\mathcal{Z}}}$ is a low-dimensional latent representation of an observed higher-dimensional variable $X \in \mathcal{X}$. The outcome $Y \in \mathcal{Y} \subset \mathbb{R}$ and *instrument variable* $A \in \mathcal{A} \subset \mathbb{R}^{d_{\mathcal{A}}}$ are observed. Throughout the paper, we assume that our data are generated according to the following SCM:

$$\mathcal{S} : \begin{cases} X := g_0(Z) \\ Y := f_0(Z) + \varepsilon, \end{cases} \tag{1}$$



where ε is a residual term with zero mean and finite variance but correlated with treatment Z, *i.e.*, $\mathbb{E}[\varepsilon|Z] \neq 0$, $g_0 : \mathcal{Z} \to \mathcal{X}$ is a nonlinear injective mixing function, and $f_0 : \mathcal{Z} \to \mathbb{R}$ is the structural function. Since Z is not observed, our first goal is to recover the latent treatment Z up to some indeterminacy exploiting the instrument A as an auxiliary variable.

[1] Implementation and experiments can be found on GitHub.

2.2 Instrument- and Auxiliary Variables

It is well-known that in the general case, nonlinear ICA is infeasible [Hyvärinen and Pajunen, 1999], however, the instrument variable setting as introduced before assumes the observation of a variable A with direct causal influence on Z. Similarly, the nonlinear ICA literature often relies on an observed *auxiliary variable* Hyvärinen et al. [2019] with direct causal influence on latent variable to guarantee its identifiability. We build upon theory from Khemakhem et al. [2020a,b] to show that the latent treatment Z can be recovered up to an affine transformation, with a few assumptions on the distribution of Z which are compatible with the general IV framework. Let us define an encoder $\phi : \mathcal{X} \to \mathcal{Z}$, typically parametrized as a neural network, whose goal is to approximate the inverse mixing function g_0^{-1}. We define *affine identifiability*:

Definition 2.1 (Latent Identifiability)**.** We say that the latent features Z are identified up to an affine transformation and pointwise transformation if there exist an encoder $\phi : \mathcal{X} \to \mathcal{Z}$ such that:

$$\phi \circ g_0(z) = RT(z) + c, \forall z \in \mathcal{Z}$$

with T a pointwise function, R an invertible matrix and $c \in \mathbb{R}^{d_Z}$.

While classic identifiability results usually rely on the mutual independence of the Z components when conditioned on A, which would restrict the types of confounding that we can consider, we build upon the results of Khemakhem et al. [2020b], which demonstrates identifiability in a more general exponential factorial case. Let us first define the conditional exponential family.

Definition 2.2 (Conditional Exponentially Factorial Distribution)**.** We say that a multivariate random variable Z is conditional exponentially factorial if its conditional density has the form

$$p_{T,\lambda}(z|a) := \mu(z) \exp\left(\sum_{i=1}^{d_Z} T_i(z_i)^\top \lambda_i(a) - \Gamma(a) \right), \tag{2}$$

where $T_i : \mathbb{R} \to \mathbb{R}^k$ are called the *sufficient statistics*.

Next, we show that under these model assumptions, we can extend the identification result of Khemakhem et al. [2020b] to our instrument variable setting, and show that InfoNCE [van den Oord et al., 2019] is a suitable loss to train an encoder satisfying Definition 2.1.

3 InfoIV

For the data generative process defined in the previous section, we propose a two-phase method to perform IV regression. In Phase 1, the instrument A is used as an auxiliary variable to recover the sufficient statistic of the latent treatment variable Z up to an invertible affine transformation and pointwise transformation (cf. Section 3.1). Specifically, we train an encoder ϕ by minimizing a contrastive loss inspired from InfoNCE [van den Oord et al., 2019], for which we prove that it identifies the true inverse mixing function g_0^{-1} up to an affine transformation and coordinatewise nonlinearities defined by the sufficient statistics. Subsequently, in Section 3.2, we show that we can leverage the learned representations for a 2SLS approach, as well as for extrapolation (Section B.1) via the control function approach similar to the autoencoder-based method proposed by Saengkyongam et al. [2024].

3.1 Recovering Suitable Representations for IV Regression

To recover Z up to a transformation suitable for IV regression, we train an encoder ϕ to maximize the similarity between our estimated latent treatments $\hat{z} := \phi(x)$ and its corresponding instrument a. Accordingly, we modify the well-known InfoNCE loss as follows:

$$\mathcal{L}_{\text{NCE}}(\phi, W) = \mathbb{E}_{A,X}\left[-\log \frac{e^{-\phi(X)WA/\tau}}{\sum_{\tilde{A}\sim P_A} e^{-\phi(X)WA/\tau}} \right], \tag{3}$$

where W is a learnable matrix $\in \mathbb{R}^{d_Z \times d_A}$ and τ is the temperature.

We show that under assumptions of sufficient variability of Z *w.r.t.* the auxiliary variable A, upon convergence of the loss, the corresponding encoder weakly identifies the latent treatment Z.

Theorem 3.1. *Let the conditional $Z \mid A$ follow the conditional factorial distribution introduced in Definition 2.2, with parameters (T, λ). Further, let $g_0 : \mathcal{Z} \to \mathcal{X}$ be a (non-linear) injective mixing function and $X := g_0(Z)$. Consider that the following conditions hold:*

1. *The sufficient statistic $T(z) = (T_i(z_i))_{i=1}^{d_Z}$ is differentiable almost everywhere.*
2. *There exist $d_Z + 1$ distinct points $u^0, ..., u^{d_Z}$ such that the matrix*

$$L_\lambda(\mathbf{u}) = (\lambda(u^1) - \lambda(u^0), ..., \lambda(u^n) - \lambda(u^0)) \quad \textit{is invertible.}$$

3. *We train ϕ^* an encoder with universal approximation capability and $W^* \in \mathbb{R}^{d_Z \times d_A}$ on the loss stated in Equation (3).*

Then in the limit of infinite data, $\phi^*(X)$ identifies Z up to an invertible linear transformation and pointwise nonlinearities defined by its sufficient statistics.

Remark 3.2. Hyvärinen et al. [2019] introduce a related contrastive loss that enables weak identification of latent variables under the same assumptions. Their method trains a logistic regression head on top of the encoder, using as input both the learned latent representation and the instrument, in order to discriminate between positive pairs (sampled from the joint distribution) and negative pairs (sampled independently). In contrast, we show experimentally that our method based on the InfoNCE loss converges faster. We hypothesize that this improvement arises because, at each SGD iteration, our approach compares each point against all other negative pairs within the batch, making it computationally more stable.

3.2 InfoIV-2SLS

The previous result establishes that we can recover the latent treatment up to an invertible linear transformation of the sufficient statistic in the conditional exponential case. We now show that this level of indeterminacy suffices to uniquely identify the causal effect, by extending Theorem A.4 [Newey and Powell, 2003].

Lemma 3.3. *Let (Z, Y) be generated according to Equation (1). Suppose we observe an instrument A that satisfies Assumption A.3 with respect to Z. Let T be differentiable almost everywhere, R an invertible matrix, and c a vector, defining a mapping $\rho : \mathbb{R}^{d_Z} \to \mathbb{R}^{d_Z}$ by $\rho(z) = RT(z) + c$. Then:*

$$\mathbb{E}[f_0 \circ \rho(Z) \mid A] = \mathbb{E}[\hat{f} \circ \rho(Z) \mid A] \quad \Rightarrow \quad f_0 \circ \rho = \hat{f} \circ \rho. \tag{4}$$

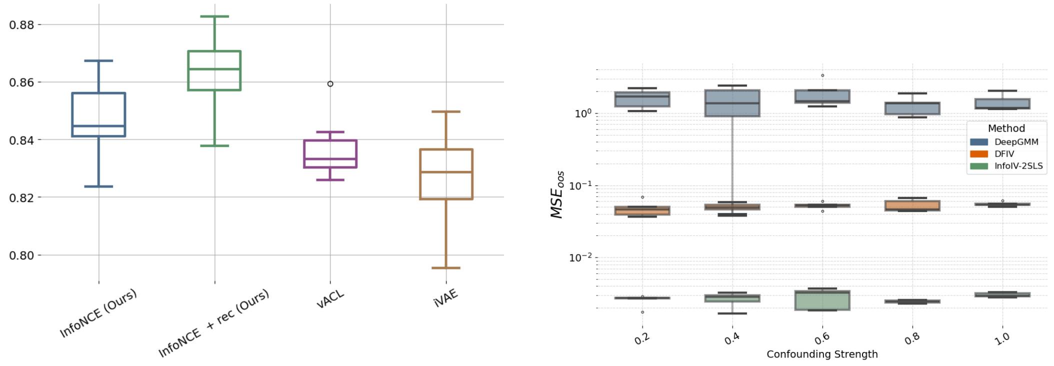
Proof. Since R is invertible and T is differentiable almost everywhere, the mapping ρ is differentiable almost everywhere as well. Hence, both $f_0 \circ \rho$ and $\hat{f} \circ \rho$ are differentiable and satisfy the conditions of Theorem A.4. By the completeness property of the exponential family, the conditional expectation equality implies the functional equality, establishing the claim. □

In summary, Theorem 3.1 and Lemma 3.3 ensure that 2SLS approaches are applicable on the learned representation that we recover based on the loss stated in Equation (3). Additionally to standard IV assumptions, A has to fulfill the IMCA assumption with respect to Z (Definition 2.2). As outlined in Algorithm 0, we first train the encoder and subsequently perform 2SLS. In practice, we perform both regression steps independently with neural networks.

4 Experiments

4.1 Recovering latent treatments

For the experiments shown in the following subsections, we simulate data according to the following data-generating process. The instrumental variable A is drawn independently from a uniform distribution. The latent treatment variable Z is then generated according to a conditionally exponential family distribution as defined in Definition 2.2: $Z := \tilde{\mu}(A) + \mathrm{diag}\left(\tilde{\sigma}_1(A), \ldots, \tilde{\sigma}_{d_Z}(A)\right) \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \Sigma)$ is sampled independently of A. The functions $\tilde{\mu}$ and $\tilde{\sigma}_i$ are nonlinear mappings $\mathbb{R}^{d_A} \to \mathbb{R}^{d_Z}$, implemented as randomly initialized neural networks. The observed treatment is then defined as $X := g_0(Z)$, where g_0 is a neural network with enforced injectivity. We first evaluate Phase 1 of InfoIV, *i.e.*, we evaluate how well we can recover the latent treatments by exploiting the instrument A as a proxy variable. As detailed in Section 3.1 this step is performed by minimizing



(a) Mean Correlation Coefficient (MCC) per method, A is of dimension 10, Z of dimension 8 and X of dimension 12. We sample and evaluate all methods on 20 datasets, each with 5,000 points.

(b) Average o.o.s. MSE score per confounding strength. Same dimensions for A, X and Z as described in (a). Y is scalar. Each method is run with 10 different seeds.

an adaptation of the InfoNCE loss tailored to our setting. We further ablate our method by adding a decoder and a reconstruction term to the loss (cf. Section C.4). Both variants are benchmarked against two baselines: iVAE [Khemakhem et al., 2020a] and vanilla auxiliary contrastive learning (vACL) [Hyvärinen et al., 2019], whose descriptions and implementation details are provided in Section C.3. We see that our InfoNCE variant to perform Phase 1 of Algorithm 0 outperforms both iVAE and vACL (Figure 1a). Our ablation study in which we add a decoder and a reconstruction term to the loss, provides additional benefits, increasing the mean MCC by approximately 0.015. While helping in terms of reconstruction, however, we observe that increasing the weight for the reconstruction term decreases the performance for the estimation of causal effects, as shown in Section C.4.

4.2 Recovering causal effect in high-dimension

To evaluate our method in the context of high-dimensional treatments, we conduct experiments on the dSprites dataset [Matthey et al., 2017], where each 64×64 image is described by five latent factors: scale, rotation, shape, x-position, and y-position. In our setup, the treatments X are the images, while the outcome Y is a scalar function of the latent factors Z, confounded by the y-position variable (details are provided in Section C.2). To evaluate the ability of our method to recover the true causal effect, we compute the out-of-sample mean squared error (o.o.s. MSE) of the predicted outcome against the ground-truth causal effect. We compare InfoIV-2SLS to DeepGMM A. Bennett and Schnabel [2019] and DFIV Liyuan Xu [2021]. We adapt the same training procedure for Phase 1 (train for 50 epochs). We used convolution layers for feature extractor, all methods were run with a similar architecture. Each method on 5,000 data points. We trained InfoIV-2SLS and DFIV for 100 epochs and DeepGMM for 50 epochs since it tended to overfit quickly. We observe that our method outperforms both DFIV and DeepGMM by orders of magnitude for every confounding strength, while DeepIV and KIV failed to converge to reasonable solutions and are therefore excluded from the plot.

5 Discussion and Conclusion

In this paper, we studied a representation-based setting for instrumental variable regression in which we cannot directly access the treatment variable, but only observe a potentially high-dimensional mixing of it. Within this setting, we proved the suitability of a two-phase approach in which we first recover the latent treatments up to an affine transformation via a variant of contrastive learning that leverages the instrument as an auxiliary. We implement our method, InfoIV, which exploits the learned latent variables for IV regression via 2SLS. To recover the latent treatments in Phase 1, we adapt the InfoNCE loss to our setting. We showcase the advantage of InfoIV against *state-of-the-art* methods on causal effect estimation with high-dimensional treatment variable. For future work, we aim to evaluate the extrapolation capacities of InfoIV on vision datasets, as well as explore more principled approaches for 2SLS such as DFIV in Phase 2 of our approach.

References

N. Kallus A. Bennett and T. Schnabel. Deep generalized method of moments for instrumental variable analysis. *33rd Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

Chunrong Ai and Xiaohong Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003.

Adin Ben-Israel. The change-of-variables formula using matrix volume. *SIAM J. Matrix Anal. Appl.*, 21(1):300–312, 1999. doi: 10.1137/S0895479895296896.

Serge Darolles, Yanqin Fan, Jean-Pierre Florens, and Eric Renault. Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565, 2011.

Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, and Julia E Vogt. Identifiability results for multimodal contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=U_2kuqoTcB.

Aapo Hyvärinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *30th Conference on Neural Information Processing Systems (NeurIPS)*, 2016.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

Aapo Hyvärinen, Hiroaki Sasaki, and Richard E. Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 89, 2019.

Guido W Imbens and Joshua D Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994.

K. Leyton-Brown J. Hartford, G. Lewis and M. Taddy. Deep iv: A flexible approach for counterfactual prediction. *Proceedings of the 34th International Conference on Machine Learning*, 70: 1414–1423, 2017.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ica: A unifying framework. *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 108, 2020a.

Ilyes Khemakhem, Ricardo P. Monti, Diederik P. Kingma, and Aapo Hyvärinen. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ica. *34th Conference on Neural Information Processing Systems (NeurIPS)*, 2020b.

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. *Proceedings of Machine Learning Research*, 140:1–57, 2022.

Sébastien Lachapelle, Divyat Mahajan, Ioannis Mitliagkas, and Simon Lacoste-Julien. Additive decoders for latent variables identification and cartesian-product extrapolation. *37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Siddarth Srinivasan-Nando de Freitas Arnaud Doucet Arthur Gretton Liyuan Xu, Yutian Chen. Learning deep features in instrumental variable regression. *The Ninth International Conference on Learning Representations*, 2021.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Gemma E. Moran, Dhanya Sridhar, Yixin Wang, and David M.Blei. Identifiable deep generative models via sparse decoding. *Transactions on Machine Learning Research*, 2022.

M.Z Nashed and Grace Wahba. Generalized inverses in reproducing kernel spaces: An approach to regularization of linear operator equations. *SIAM Journal on Mathematical Analysis*, 5(6): 974–987, 1974.

Whitney K. Newey and James L. Powell. Instrumental variable estimation of nonparametrix models. *Econometrica*, 71(5):1565–1578, 2003.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2009.

Sorawit Saengkyongam, Elan Rosenfeld, Pradeep Kumar Ravikumar, Niklas Pfister, and Jonas Peters. Identifying representations for intervention extrapolation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3cuJwmPxXj.

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.

Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. *33rd Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2019.

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.

Rui Zhang, Masaaki Imaizumi, Bernhard Schölkopf, and Krikamol Muandet. Instrumental variable regression via kernel maximum moment loss. *Journal of Causal Inference*, 11(1), 2023.

Appendix

Table of Contents

A Theory

A.1 Identifiability Proofs

We begin with proving that an encoder trained on the InfoNCE loss 3 converges to a solution that identify the ground-truth latent variable up to an affine transformation and a pointwise transformation, in the case where our latent variable follows a factorial conditional distribution (Definition 2.2).

Theorem 3.1. *Let the conditional $Z \mid A$ follow the conditional factorial distribution introduced in Definition 2.2, with parameters (T, λ). Further, let $g_0 : \mathcal{Z} \to \mathcal{X}$ be a (non-linear) injective mixing function and $X := g_0(Z)$. Consider that the following conditions hold:*

1. *The sufficient statistic $T(z) = (T_i(z_i))_{i=1}^{d_Z}$ is differentiable almost everywhere.*
2. *There exist $d_Z + 1$ distinct points $u^0, ..., u^{d_Z}$ such that the matrix*

$$L_\lambda(\mathbf{u}) = (\lambda(u^1) - \lambda(u^0), ..., \lambda(u^n) - \lambda(u^0)) \quad \textit{is invertible.}$$

3. *We train ϕ^* an encoder with universal approximation capability and $W^* \in \mathbb{R}^{d_Z \times d_A}$ on the loss stated in Equation (3).*

Then in the limit of infinite data, $\phi^*(X)$ identifies Z up to an invertible linear transformation and pointwise nonlinearities defined by its sufficient statistics.

Proof. As argued in van den Oord et al. [2019], in the limit of infinite data with ϕ and ψ having universal approximation capacity, if:

$$\phi^*, W^* = \arg\min_{\phi, W} \mathcal{L}_{\text{NCE}},$$

then

$$e^{\phi^*(x)W^*a} \propto \frac{p(x|a)}{p(x)}.$$

Let us recall that we assume g_0 to be injective, therefore it admits a left inverse on its image space contained in $\mathcal{X}$ that we denote g_0^{-1}. Under the assumption that g_0^{-1} has full-rank Jacobian, one can apply the change of variable formula with the volume matrix vol $A := \sqrt{\det A^T A}$ [Ben-Israel, 1999].

$$\begin{aligned} \phi^*(x)W^*a &= \log c + \log p(x|a) - \log p(x) && (5)\\ &= \log c + \log p_Z(g_0^{-1}(x)|a) - \log p_Z\big(g_0^{-1}(x)\big) && (6)\\ &= \log c + \log p(z|a) - \log p(z) && (7) \end{aligned}$$

We use the change of variable formula to go from 5 to 6 and notice that the Jacobian volumes cancel themselves. We define c the proportionality constant that is not dependent on a or x. At line 7 we simply set $z := g_0^{-1}(x)$. By assumption, $\{Z_i\}_{i=1,...,d_Z}$ given A follow an exponential distribution (Definition 2.2), thus, following the proof of Khemakhem et al. [2020b][Theorem 9]:

$$\begin{aligned} \phi^*(x)W^*a &= \log p_{T,\lambda}(z|a) - \log p_Z\big(z\big) + \log c && (8)\\ &= \log c + T(z)\lambda(a) + \log \mu(z) - \Gamma(a) - p(z), && (9) \end{aligned}$$

By collecting these equations for every a_k, $k \in \{0, ..., d_Z\}$ as defined in assumption 3. and taking out the case a_0, we obtain for all $k \in \{1, ..., d_Z\}$:

$$\phi^*(x)W^*(a_k - a_0) = T(z)\big(\lambda(a_k) - \lambda(a_0)\big) + \big(\Gamma(a_0) - \Gamma(a_k)\big), \tag{10}$$

which yield the following matrix form:

$$\phi^*(x)\Psi = T(z)L + C, \tag{11}$$

with Ψ a $\mathbb{R}^{d_Z \times d_A}$ matrix whose k-th row is given by $a_k - a_0$ which is non-zero by assumption, L is defined as in assumption 3 and C is a vector of dimension d_Z whose k-th element is given by $\Gamma(a_0) - \Gamma(a_k)$. By assumption, L is invertible thus we can multiply both side by its inverse, which yields the following result:

$$\phi^*(x)R = T(z) + \tilde{C}, \tag{12}$$

with $\tilde{C} := CL^{-1}$ and $R := \Psi L^{-1}$.

Finally, by assumption T has full-rank Jacobian and is thus non-degenerate. As a consequence, the mapping $z \mapsto zR$ has to cover the full-space and thus cannot be degenerate. Since R is a square matrix we deduce its invertibility.

□

After stating the identifiability of InfoNCE in the factorial conditional exponential case, we can now state its more refined identifiability in the Gaussian case. Since this result is required for extrapolation, we first recite the corresponding theorem enabeling extrapolation of Saengkyongam et al. [2024].

Theorem A.1 (Saengkyongam et al. [2024], Theorem 4)**.** *Assume Setting 21 with f_0 and l differentiable. Let ϕ be an encoder that identifies g_0^{-1} up to an affine transformation. Let:*

$$(W_\phi, \alpha_\phi) := \underset{W \in \mathbb{R}^{d_Z \times d_A}, \alpha \in \mathbb{R}^{d_Z}}{\arg\min} \mathbb{E}[\|\phi(X) - (WA + \alpha)\|^2]. \tag{13}$$

and the estimated noise term $V_\phi := \phi(X) - (W_\phi A + \alpha_\phi)$. Finally, let ν and ψ be the the estimated functions obtained from additive regression of Y on $\phi(X)$ and V_ϕ. Then:

$$\forall a^* \in \mathcal{A}, \mathbb{E}[Y|do(A = a^*)] = \mathbb{E}[\nu(W_\phi a^* + \alpha_\phi + V_\phi)] - (\mathbb{E}[\nu(\phi(X))] - \mathbb{E}[Y]). \tag{14}$$

Corollary A.2. *Assume $Z := M_0 A + V$ with M_0 full-rank and $V \sim \mathcal{N}(0, \Sigma)$. Let $X := g_0(Z)$ with g_0 an injective function. Assume that there exist $d_Z + 1$ linearly independent distinct points in $supp(A)$. Then, in the limit of infinite data an encoder ϕ^* trained to minimize loss Equation (3) provides a consistent estimator of Z up to an invertible affine transformation.*

Proof. Let us recall that we sample data from the following SCM:

$$\mathcal{S} : \begin{cases} V \sim \mathcal{N}(0, \Sigma) \\ Z := M_0 A + V \\ X := g_0(Z) \end{cases}$$

with g_0 injective and M_0 full row rank. We have:

$$p(z|a) = p_V(z - M_0 a) \tag{15}$$

$$= (2\pi)^{-d_Z/2} \det(\Sigma)^{-1/2} \exp\big[-\frac{1}{2}(z - M_0 a)^T \Sigma^{-1} (z - M_0 a)\big] \tag{16}$$

$$= (2\pi)^{-d_Z/2} \det(\Sigma)^{-1/2} \exp -\frac{1}{2}\big[z^T \Sigma^{-1} z - z^T \Sigma^{-1} M_0 a - a^T M_0^T \Sigma^{-1} z + a^T M_0^T \Sigma^{-1} M_0 a\big] \tag{17}$$

$$= (2\pi)^{-d_Z/2} \det(\Sigma)^{-1/2} \exp\big[-\frac{1}{2} z^T \Sigma^{-1} z\big] \exp\big[z \Sigma^{-1} M_0 a\big] \exp\big[-\frac{1}{2} a^T M_0^T \Sigma^{-1} M_0 a\big] \tag{18}$$

$$= \mu(z) \exp\big[z \Sigma^{-1} M_0 a - \Gamma(a)\big] \tag{19}$$

where we go from Eq. 17 to 18 by noticing that the two terms are scalar and the transpose of the other, in Eq. 19 we set $\mu(z) := (2\pi)^{-d_Z/2} \det(\Sigma)^{-1/2} \exp\big[-\frac{1}{2} z^T \Sigma^{-1} z\big]$ and $\Gamma := \frac{1}{2} a^T M_0^T \Sigma^{-1} M_0 a$. This derivation allows us to identify a conditional exponential family with parameters (T, λ), as introduced in Definition 2.2. In particular, we obtain $\forall i = 1, ..., d_Z$:

$$\begin{cases} T_i(t) = t, & \forall t \in \mathbb{R} \\ \lambda_i(u) = \Sigma^{-1} M_0 u, & \forall u \in \mathbb{R}^{d_A} \end{cases}$$

It remains to prove that this parametrization validates the assumptions of Theorem 3.1. Let us choose $u^0, \ldots, u^{dz}$ in $\text{supp}(A)$, assumed to exist, such that these $d_Z + 1$ points are distinct and linearly independent. Define

$$U \in \mathbb{R}^{d_A \times dz}, \qquad U = \big(u^1 - u^0, \ldots, u^{dz} - u^0\big).$$

By construction, the columns of U are linearly independent, so U has full column rank, *i.e.*, $\operatorname{rank}(U) = d_Z$.

Since Σ is invertible, we have $\operatorname{rank}(L) = \operatorname{rank}(M_0 U)$. Moreover, M_0 is assumed to be full row rank of dimension d_Z. Therefore,

$$\operatorname{rank}(M_0 U) = \min\{\operatorname{rank}(M_0), \operatorname{rank}(U)\} = \min\{d_Z, d_Z\} = d_Z.$$

Thus $M_0 U$ is square and invertible, which implies that L is also invertible. This verifies the full-rank condition required in Theorem 3.1. □

A.2 Standard Instrument Variable Setting

Instrument variable (IV) regression assumes that we observe a treatment $X \in \mathcal{X} \subset \mathbb{R}^{d_X}$ and an outcome $Y \in \mathcal{Y}$ generated according to the following structural causal model (SCM)

$$Y := f_0(X) + \varepsilon, \tag{20}$$

where f_0 denotes the structural function and ε is a residual term with zero mean and finite variance. In contrast to the standard supervised learning setting—where ε are assumed to be *i.i.d.* and independent of X—the IV framework allows for the presence of confounder, which implies that the residual term is correlated with the treatment, *i.e.*, $\mathbb{E}[\varepsilon|X] \neq 0$. To account for the confounding variable, we assume that we observe an *instrument variable* $A \in \mathbb{R}^{d_A}$ which satisfies the following conditions.

Assumption A.3. An instrument $A \in \mathbb{R}^{d_A}$ satisfies the following conditions: (i) A has a direct causal influence on treatment (**Relevance**), *i.e.*, $P(X|A)$ is not constant in A. (ii) A is uncorrelated with the confounder (**Exogeneity**), *i.e.*, $\mathbb{E}[\varepsilon|A] = 0$.

Based on Assumption A.3 the ground-truth structural function satisfies $\mathbb{E}[Y|A] = \mathbb{E}[f_0(X)|A]$, which allows us to derive the following prominent result, which we recite for completeness.

Theorem A.4 (Newey and Powell [2003])**.** *Assume X, Y generated according to Equation (20), and let A be an instrument satisfying Assumption A.3. Further assume that the distribution of X conditional on A is exponential. Then, if f_0 and $\hat{f}$ are differentiable, $\mathbb{E}[f_0(X)|A] = \mathbb{E}[\hat{f}(X)|A]$ implies $f_0 = \hat{f}$.*

Simply put, if an estimator $\hat{f}$ reproduces the ground-truth conditional expectation of the structural function given A, then it coincides with f_0. Since directly minimizing this conditional expectation is generally ill-posed [Nashed and Wahba, 1974], more practical estimators have been derived.

B Methods

Algorithm 0: InfoIV (Sketch)

```
input : Data drawn from $P(A, X, Y)$
// Phase 1 (Representation Learning)
Obtain $\phi^*, W^* = \arg\min_{\phi,W} \mathcal{L}_{\text{NCE}}(\phi, \psi)$
Estimate latent treatment $\hat{Z} = \phi^*(X)$
// Phase 2a (2SLS)
Estimate $\mathbb{E}[\hat{Z}|A]$— obtaining $\hat{Z}_A$
Estimate $\hat{f}_0$ from the regression of $Y$ on $\hat{Z}_A$
// Phase 2b (Control Function)
Estimate $\mathbb{E}[\hat{Z}|A]$— obtaining $\hat{Z}_A$
Obtain $\hat{V} = \hat{Z} - \hat{Z}_A$
Estimate $\hat{f}_0, \hat{l}$ from the additive regression of $Y$ based on $\hat{Z}$ and $\hat{V}$
```

In this section, we provide a detailed description of how we adapt the control function method to our latent IV setting. Detailed on the 2SLS adaptation of InfoIV are provided in the main text (Section 3.2). Assume that we have trained an encoder ϕ to invert the true mixing function g_0 using the adapted InfoNCE loss introduced in Section 3.1.

B.1 InfoIV-CF

We now show that Phase 1 of InfoIV also recovers suitable features for extrapolation tasks, where we aim to predict the result of an intervention on an action variable A, when this intervention was not observed in the training support. Using do-notation [Pearl, 2009], this corresponds to estimating $\mathbb{E}[Y|do(A := a^*)]$. In particular, we build upon the results of Saengkyongam et al. [2024] who relied on an autoencoder trained via moment constraints to obtain the latent features. Saengkyongam et al. [2024] show that one can extrapolate over unseen values of A if we restrict the effect of A on Z to be linear. In particular, let us consider the following SCM:

$$\mathcal{S}_1 : \begin{cases} Z := M_0 A + V \\ X := g_0(Z) \\ Y := f_0(Z) + l(V) + \varepsilon, \end{cases} \tag{21}$$

with $A \perp\!\!\!\perp V, \varepsilon$ whose support's interior is convex. Here, ε is a noise term with zero mean and finite variance independent from Z. We further assume $M_0 \in \mathbb{R}^{d_Z \times d_A}$ to be full-rank and g_0 injective. Note that in comparison to Equation (1), the dependence to the confounder V is modeled explicitly, while previously it was absorbed in the noise term.

Most relevant for us is that Saengkyongam et al. [2024] show that if we can train an encoder ϕ that *recovers Z up to an affine-transformation*, then one can leverage the control function approach to estimate the true causal-effect f_0 and perform extrapolation on A, *i.e.*, estimate $\mathbb{E}[Y|do(A := a^*)]$ for all $a^* \in \mathcal{A}$[2]. Consequently, we need to show that we can recover the latent treatment Z up to an affine-transformation for the SCM above.

Corollary A.2. *Assume $Z := M_0 A + V$ with M_0 full-rank and $V \sim \mathcal{N}(0, \Sigma)$. Let $X := g_0(Z)$ with g_0 an injective function. Assume that there exist $d_Z + 1$ linearly independent distinct points in $supp(A)$. Then, in the limit of infinite data an encoder ϕ^* trained to minimize loss Equation (3) provides a consistent estimator of Z up to an invertible affine transformation.*

As can be noted, in comparison to 2SLS, we need to restrict the function from A to Z to be linear and need to add some distributional assumptions to ensure that the extrapolation task is well-defined. Similar to 2SLS, all regression steps are performed independently based on neural networks. This concludes our theoretical results. Next, we empirically evaluate the different components of InfoIV.

C Experiments

C.1 IMCA Data Generative Process

Injectivity of g_0. Our identifiability result stated in Theorem 3.1 relies on the assumption that the ground-truth mixing function g_0 is injective. To enforce this property in our data-generating process, we use LeakyReLU activations and initialize the weight matrices of the linear layers to be full-rank. Particulary, g_0 has 2 hidden layers of dimension $[32, 64]$. Similarly, ground-truth causal effect f_0 is a 2 hidden layers neural network with tanh activations.

C.2 dSprites Data Generative Process

We now describe the data generative process for dSprites data. We first sample a proxy between instrument and treatment in order to avoid inverting the causal direction by defining the instrument as a function of the treatment.

1. Sample a proxy variable Q uniformly in a ball around the extremal values of Z.
2. Map Q to the nearest existing latent value to define the latent treatment Z.
3. Compute the instrument A as a nonlinear mapping of the components of Q except for the one associated with position-y.
4. Obtain the observed treatment X as the corresponding images from the dSprites dataset.

[2] For completeness, we recite a shortened version of their theorem in Section A.1.

5. Define the outcome as

$$Y = f_{\text{struct}}(Z) + \rho(posY - 0.5) + \eta,$$

where f_{struct} is a randomly initialized neural network, ρ is the confounding strength, and η is Gaussian noise.

C.3 Baseline Methods

Latent recovery We perform evaluation of our latent recovery method against three existing methods: vanilla auxiliary constrastive learning (vACL) [Hyvärinen et al., 2019], iVAE [Khemakhem et al., 2020a] and first stage of Rep4Ex-CF [Saengkyongam et al., 2024]. We use the same encoder and decoder architecture for each method, as well as the neural network architecture for each method to estimate the causal effects. Additionally, vACL includes a logistic regression head that we implement as an MLP with two hidden layers with ReLU activation, trained on cross-entropy loss. All three methods are implemented in our code that is appended to the submission. The network architecture for each method consists of the following blocks: 3 blocks of Linear - Batch normalization - LeakyRelu layers, with dropout at a rate of 0.2. The hidden dimensions are fixed at 16, 32 and 64 throughout both IMCA and extrapolation experiments.

IV baseline comparison We use the implementation of DeepGMM A. Bennett and Schnabel [2019], KIV Singh et al. [2019] and DFIV Liyuan Xu [2021] provided in `https://github.com/liyuan9988/DeepFeatureIV`. We include an adapted version in our code, particularly new model specs as well as our data generative process. For the dSprites experiments we use an Image extractor (Table 2) for both DeepGMM and DFIV with a similar architecture as the encoder used for first stage of our method.

ConvBlockDown($C_{in} \to C_{out}$)	**Operations**
Conv2d($C_{in} \to C_{out}$, kernel=3, stride=2, padding=1)	Downsampling conv
BatchNorm2d(C_{out})	Normalization
Activation (LeakyReLU(0.2) by default)	Non-linearity
Dropout2d(0.1)	Regularization

Table 1: Definition of ConvBlockDown.

Layer	**Output Shape**
Input (1 × 64 × 64)	1 × 64 × 64
ConvBlockDown(1→32)	32 × 32 × 32
ConvBlockDown(32→64)	64 × 16 × 16
ConvBlockDown(64→128)	128 × 8 × 8
ConvBlockDown(128→256)	256 × 4 × 4
Flatten	4096
Dense(4096 → 6)	6

Table 2: Image feature extractor used for DeepGMM, DFIV, and InfoIV in the dSprites experiment.

GPU The extrapolation and causal effect estimation experiments on tabular data are run on an *Apple M4 Pro chip*, while the dSprites experiments are run on an *Nvidia A100 Tensor Core GPU*.

C.4 InfoIV Hyperparameters Tuning

One advantage of our method over autoencoder-based approaches is that it depends on only a single hyperparameter: the temperature in the InfoNCE loss. We tune this parameter by evaluating the validation MCC Figure 2, and notice the best performance is achieved with a temperature of 0.3, which we use for all subsequent experiments.

As mentioned earlier, we also explored adding a reconstruction term to our loss by training a decoder (mirrored architecture to the encoder) to reconstruct the input X. The resulting loss is:

$$\mathcal{L}(\phi, \psi, W) = \mathcal{L}_{\text{NCE}}(\phi, W) + \lambda_{\text{rec}} \, \|\psi \circ \phi(X) - X\|^2.$$

We conducted a study on the IMCA dataset, evaluating the learned latents against the ground truth using the MCC metric for different values of λ_{rec}. The latent features were then used in the second step of InfoIV-2SLS for causal effect estimation, which we evaluated using the out-of-sample MSE

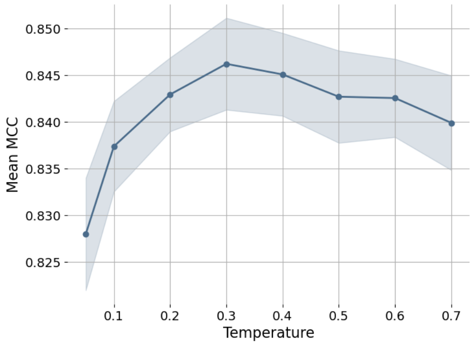


Figure 2: Average MCC on validation set (1,000 samples) over 20 runs for each temperature value. $d_A = 10$, $d_Z = 8$, $d_X = 12$. The training set includes 5,000 data points and the encoders are trained for 50 epochs on InfoNCE loss solely. Light blue area represents the 90% confidence interval.

(MSE_{oos}, Figure 3). While values of $\lambda_{\text{rec}} > 1$ generally improve the consistency of the learned representation (increasing MCC by up to 0.2), they also lead to a deterioration in causal effect estimation, raising the MSE by an average of 1.5×10^{-2}.

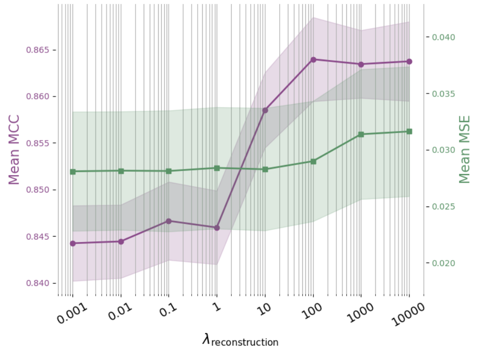
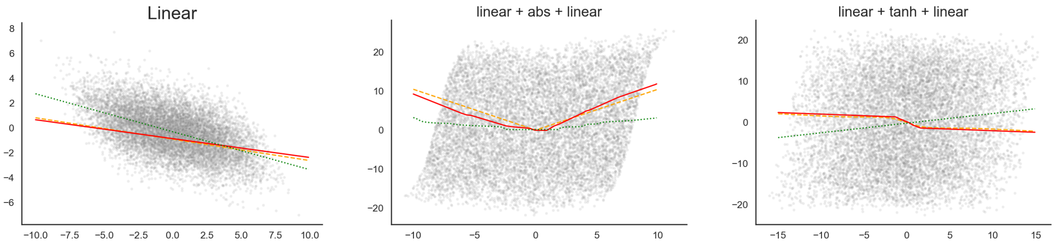


Figure 3: Average MCC (*in purple*) and out-of-sample MSE (*in green*) per reconstruction regularization parameter. The temperature for the InfoNCE loss term is fixed at 0.3.

C.5 Causal Effect Estimation in Dimension One

We additionally evaluate our method in a setting where both Z and X are scalar, while A is sampled from a two-dimensional uniform distribution. Figure 4 shows the learned causal-effect. We consider three scenarios: a) corresponds to the case of a linear causal effect; b) corresponds to a nonlinear causal effect implemented as a linear layer with hidden dimension 16, followed by a $\tanh$ activation and a final linear layer; and c) corresponds to a similar architecture where the nonlinear activation is the absolute value function instead of $\tanh$. In all cases, we first estimate the latent variable Z using the InfoNCE variant and then apply the *control function* technique for causal effect estimation. As discussed in Section B.1, our method recovers the ground-truth causal effect f_0 up to an affine indeterminacy that arises from latent variable estimation. To account for this, we learn an affine transformation that aligns the estimated latent representation with the ground-truth Z, and we report the causal effect after applying this transformation. For comparison, we also fit an OLS model mapping the ground-truth Z to the outcome Y. The OLS estimator fails to recover the causal effect, as Z is confounded with the residual variation in Y. Importantly, despite the affine indeterminacy, our method still yields a valid estimate of the causal relationship from the observed X to Y.



(a)

(b)

(c)

Figure 4: Estimated causal effect with NCE loss and control function (*in red*), ground-truth causal effect (*in orange*), OLS model (*in green*), (Z;Y) (*in grey*)