Recovering Causal Features for Instrumental Variable Regression with Contrastive Learning

Anonymous Author(s)
Affiliation
Address
email

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 (potentially high-dimensional) transformation of it. Particularly, using insights from the independent component analysis (ICA) literature, we propose a general contrastive learning framework to recover the latent treatment up to an affine transformation when it is linearly related to the instrument. We prove that the recovered representation is compatible with classical 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

Standard supervised learning techniques, such as ordinary least squares (OLS), assume that the residuals on the target variable are independent of the features to retrieve the true causal effect. However, this assumption does not generally hold. Consider a setting where we observe a treatment X and an outcome $Y = f_0(X) + \varepsilon$, 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 $\mathbb{E}[Y|X] \neq f_0(X)$, and thus classical supervised learning methods fail to recover the true causal effect. To address this, *Instrumental Variable* (IV) regression [Newey and Powell, 2003] assumes the observation of an *instrument* that affects the outcome only through the treatment variable and is thus independent from the residuals. Originally formulated for linear models, IV methods have been extended to nonlinear settings using universal approximators like neural networks [J. Hartford and Taddy, 2017, Liyuan Xu, 2021] or kernels [Singh et al., 2019].

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, identifiability of nonlinear ICA can be achieved by introducing an auxiliary variable [Hyvärinen et al., 2019, Khemakhem et al., 2020], 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

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

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.

2 Recovering Latent Treatments

2.1 Data Generating Process

We consider a *representation-based* variant of the classical instrumental variable (IV) setting, similar to Saengkyongam et al. [2024], where the corresponding graphical model is shown in Figure 1. We assume that the treatment variable $Z \in \mathcal{Z}$ is unobserved (we denote latent variables as gray nodes), while the outcome $Y \in \mathcal{Y} \subset \mathbb{R}$ and *instrument variable* $A \in \mathcal{A} \subset \mathbb{R}^{d_A}$ are observed. In addition, we observe variable $X \in \mathcal{X}$, which is a transformation of the treatment. Throughout the paper, we assume that our data are generated as follows:

$$\mathcal{S} : \begin{cases} Z := M_0 A + V \\ X := g_0(Z) \\ Y := f_0(Z) + l(V) + \varepsilon \end{cases} \tag{1}$$

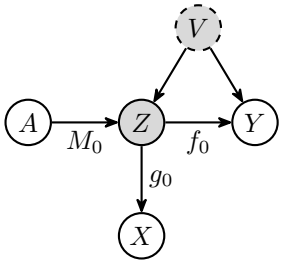


Figure 1: Causal graph of the data generating process

where $g_0 : \mathcal{Z} \to \mathcal{X}$ is a nonlinear injective mixing function, $f_0 : \mathcal{Z} \to \mathbb{R}$ and $l : \mathcal{Z} \to \mathbb{R}$ are assumed measurable and integrable. Treatment Z and effect Y are confounded by unobserved variable V, with $\mathbb{E}[V] = 0$ and $\mathbb{E}[V^2] < \infty$. Residuals ε are assumed independent from V and $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}[\varepsilon^2] < \infty$. We further assume $M_0 \in \mathbb{R}^{d_Z \times d_A}$ to be a matrix of rank d_Z (with $d_A \geq d_Z + 1$). Our goal is to recover the latent treatment Z up to some indeterminacy exploiting A as an auxiliary variable. We argue that this is sufficient for IV regression.

2.2 Identifiability of Latent Treatments

In the general setting, identifying the exact ground-truth latent variable Z from observed data alone is infeasible. We build upon theory from Khemakhem et al. [2020] to show that the latent treatment Z can be recovered up to an affine transformation, particularly in the case where V is an isotropic multivariate Gaussian. 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}. Let $p_\phi(x|a)$ be the posterior distribution of $\tilde{X} := \phi^{-1}(Z)$ given A, then we define *affine identifiability* as (introduced in Saengkyongam et al. [2024]):

Definition 2.1 (Affine identifiability)**.** We say that the latents Z are identifiable up to an affine transformation if there exist an encoder $\phi : \mathcal{X} \to \mathcal{Z}$ such that:

$$p(x|a) = p_\phi(x|a) \Longrightarrow \exists R \in \mathbb{R}^{d_Z \times d_Z} \text{ invertible}, c \in \mathbb{R}^{d_Z} \text{ such that } \phi \circ g_0(z) = Rz + c, \forall z \in \mathcal{Z}\,.$$

In particular, we show that under the assumptions of our data generating process, Z is identifiable up to an affine transformation, where we leverage A as an *auxiliary variable*.

Lemma 2.2. *Assume our data are sampled according to Equation* (1)*, with A being a random variable independent from $V \sim \mathcal{N}(0, \Sigma_V)$, an isotropic multivariate Gaussian random variable. Further assume that the support of A contains at least $d_Z + 1$ affinely independent points. Then Z is identifiable up to an affine transformation.*

Proof Sketch: We rely on an identifiability result from Hyvärinen et al. [2019], which is guaranteed by the conditional independence of the latent Z given instrument A and by the full-rankness of M_0. The full proof is stated in Appendix A.3.

In other words, if we can approximate the ground-truth conditional probability $p(x|a)$ using a suitable learning algorithm, we are guaranteed that the learned encoder approximates the ground-truth latent treatment up to an affine transformation. This affine identifiability of the latent treatment is then sufficient to carry out IV regression.

2.3 Causal Effect Identifiability

Guo and Small [2016] provide assumptions on the instrument for identifiability of a nonlinear causal-effect, in the classical IV setting with observed treatment. We first state them and then discuss how they relate to our setting where the treatment is unobserved.

Assumption 2.3. Let A, Z and Y be sampled according to Equation (1). Furthermore, assume:

(i) **Relevance**: The instrument variable has a direct causal influence on treatment, *i.e.*, $P(Z|A)$ is not constant in A.

(ii) **Quasi-exogeneity**: $A \perp\!\!\!\perp \varepsilon$ and $\mathbb{E}[l(V)|A]$ is constant.

(iii) **Strong Exogeneity**: $A \perp\!\!\!\perp V$.

Remark 2.4. Note that classical IV methods typically assume $\mathbb{E}[l(V)|A] = 0$. However, this assumption restricts l to a very specific class of functions. In practice, relaxing this condition introduces a location indeterminacy in the estimated causal effect. Importantly, this indeterminacy does not affect predictive performance nor the robustness of the estimates to distributional shifts.

If A satisfies assumptions (i) and (ii), we say that it is a *valid instrument* and the ground-truth causal effect f_0 satisfies:

$$\mathbb{E}[Y|A] = \mathbb{E}[f_0(Z)|A] + \textit{constant} \tag{2}$$

This equation is generally ill-posed [Nashed and Wahba, 1974], but it motivates a *two-stage least squares estimator (2SLS)*: first, regress Z on A: $Z_A = \mathbb{E}[Z|A]$, then regress Y on the predicted treatment Z_A with an intercept to account for the constant. In addition, if A fulfills assumption (iii) then f_0 and l satisfy:

$$\mathbb{E}[Y|Z, V] = f_0(Z) + l(V) \tag{3}$$

This equality motivates the *control function* method, where we first regress Z on A obtaining an estimate $Z_A = \mathbb{E}[Z|A]$. Then (since $A \perp\!\!\!\perp V$) the residuals can be consistently estimated from $\hat{V} := Z - Z_A$. As a final step, we perform additive regression of Y on Z and $\hat{V}$ to estimate f_0 and l.

However, in our case, the treatment is not directly observed; instead, we observe a nonlinear transformation $X := g_0(Z)$. Since we previously showed that the true treatment can be recovered up to an affine transformation (under assumptions on g_0 and A that are compatible with those stated above), one can easily verify that, if the assumptions hold for the true treatment Z, they also hold for any invertible affine transformation of Z:

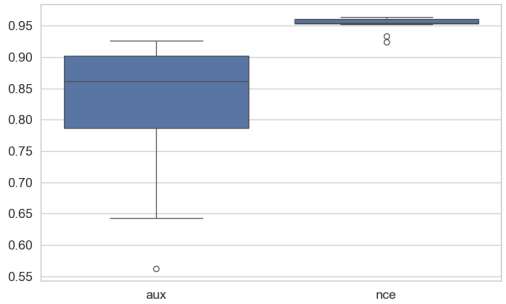
Lemma 2.5. *Let $R \in \mathbb{R}^{d_Z \times d_Z}$ be an invertible matrix, $c \in \mathbb{R}^{d_Z}$ and $\hat{Z} := RZ + c$. If A is a valid instrument for treatment Z according to Assumption 2.3, it is a valid instrument for treatment $\hat{Z}$ along the same assumptions.*

See Appendix A.2 for proofs and additional details on functional assumptions on f_0 and l that ensure their recoverability.

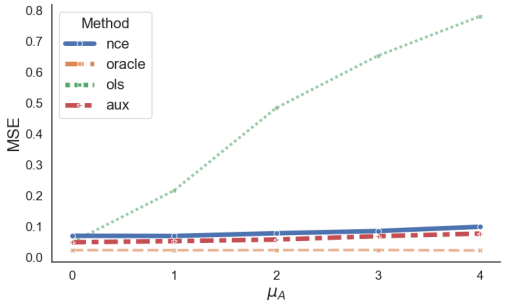
A note on related work: The result we derived above is most closely related to the so-called extrapolation identifiability [Saengkyongam et al., 2024]. Briefly, their method formulates the *moment conditions* in a reproducing kernel Hilbert space (RKHS) to enforce both linearity between the treatment and the instrument, and independence of the treatment residuals from the instrument. Particularly, they leverage *control function* to perform extrapolation on unseen values of instrument A.

3 InfoIV: A Contrastive Learning Framework for IV

Following our previous identifiability results, we propose a two-phase method to perform IV regression. In the first phase, the instrument A is used as an auxiliary variable to recover the latent treatment variable Z up to an affine transformation. Specifically, we apply a general contrastive learning



(a) R^2 scores per method and V distribution (Gaussian), A is of dimension 8, Z of dimension 6 and X of dimension 10. We run each method with 10 different seeds.



(b) Average MSE score per shift, for *control function* approach. Same dimensions for A, X and Z as described in (a). Y is scalar. Each method is run with 10 different seeds.

Figure 2: Comparison of methods: (a) R^2 scores and (b) MSE under shifts.

framework [Hyvärinen et al., 2019], which provably (i) learns an encoder that identifies the true inverse mixing function g_0 up to an affine transformation, and (ii) approximates the conditional distribution $p(z|a)$, enabling its use in a second phase for unconfounded causal effect estimation (either via 2SLS or the control function method). We also introduce a variant based on InfoNCE [van den Oord et al., 2019], a common contrastive learning objective, and demonstrate its identifiability capacity in Appendix A.3.2. We provide more details to both variants in Appendix A.3.

4 Experiments

For our experiments we sample A, X, Z, Y according to the data generating process introduced in Section 2.1. We consider the isotropic multivariate Gaussian case where $A \sim \mathcal{N}(0, I_{d_A})$ and $V \sim \mathcal{N}(0, I_{d_V})$. We first estimate the latent treatment exploiting the two contrastive methods described in Appendix A.3. In order to verify that we recover the ground-truth latents up to an affine transformation, we compute their R^2-score against our estimated latent features. Both methods approximate the ground-truth latent variables, as attested by the high R^2 scores. However, the NCE method ($R^2 = 0.96$) is more consistent, whereas the Auxiliary Variable method ($R^2 = 0.86$) can even fail for some seeds (minimum observed: 0.57). Based on the estimated latents, we perform as the next step IV regression via the control function method and 2SLS (implementation details in Appendix B). We show that the learned causal-effect is robust to distribution shift, by evaluating it for $A \sim \mathcal{N}(\mu_{test}, I_{d_A})$ with $\mu_{test} \in \{1, 2, \ldots, 5\}$. We evaluate our method against: (i) an oracle model that performs IV regression on the ground-truth Z, (ii) a naive model that performs ordinary-least-squares estimation of Y based on ground-truth Z. As expected, the OLS method yields equivalent results as IV methods when no shifts are applied to the test data distribution. However, as the shift increases, the MSE scores for *two-stage least squares* (Figure 2b) and also for the *control function* method (see Figure 3 in Appendix) remain stable while OLS fails to generalize on the test data distribution. Please refer to Appendix B.1 and B.2 for further details on the data generating process, the neural net architectures and the hyperparameters. We also show examples of learned causal effect in the one dimension case in Appendix B.3.

5 Discussion and Conclusion

Our work demonstrates that contrastive learning is effective for learning latent representations in an IV setting where the treatment is unobserved. In particular, we empirically show that it integrates well with both two-stage least squares and control function approaches. In future work we aim to explore its integration with anchor regression [Dominik Rothenhäusler et al., 2021], which relaxes the standard IV assumption by allowing the instrument to directly influence the outcome. Furthermore, we aim to evaluate our method on high-dimensional and real-world data.

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.

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$_2$$kuqoTcB$.

Nicolai Meinshausen Dominik Rothenhäusler, Peter Bühlmann, and Jonas Peters. Anchor regression: heterogeneous data meet causality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(2):215–246, 2021.

Zijian Guo and Dylan Small. Control function instrumental variable estimation of nonlinear causal effect models. *Journal of Machine Learning Research*, 17(100):1–35, 2016.

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.

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

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, 2020.

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.

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.

Whitney K. Newey, James L. Powell, and Francis Vella. Nonparametric estimation of triangular simultaneous equations models. *Econometrica*, 67(3):565–603, 1999.

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

In the following, we prove the identification result in Lemma 2.2. After defining the exponential family, we recall the identifiability theorem of [Khemakhem et al., 2020] and show that the representation IV setting fulfills its conditions, ensuring that the latent treatment is identifiable up to an invertible affine transformation.

Definition A.1 (Conditional Factorial "simple" Exponential distribution)**.** A multivariate conditional distribution is *simple factorial exponential*, if its density function can be written as:

$$p_{T,\lambda}(z|a) = \prod_i \frac{Q_i(z_i)}{Z_i(a)} \exp[T_i(z_i)\lambda_i(a)]$$

where Q_i is a real-valued function, $Z_i : \mathcal{A} \to \mathbb{R}$ the normalization constant, $T_i : \mathbb{R} \to \mathbb{R}^{d_A}$ are the sufficient statistics and $\lambda_i : \mathbb{R}^{d_A} \to \mathbb{R}^{d_A}$ are such that $\frac{\partial \lambda_i}{\partial u} \neq 0$ for all i.

This corresponds to the exponential family with 1-dimension sufficient statistics. We state the main identifiability result of Khemakhem et al. [2020] in the case of "simple" exponential distributions.

Theorem A.2 ([Khemakhem et al., 2020] Identifiability of conditionally exponential latents)**.** *Let $x \in \mathbb{R}^{d_X}$, and $a \in \mathbb{R}^{d_A}$ be observed random variables and $z \in \mathbb{R}^{d_Z}$ a latent variable of exponential conditional distribution w.r.t a $p_{T,\lambda}(z|a)$ as defined in A.1. Let $g_0 : \mathbb{R}^{d_Z} \to \mathbb{R}^{d_X}$ be an injective function such that:*

$$X = g_0(Z) + \varepsilon$$

with ε a random variable independent from z, yielding to the following conditional generative model:

$$p_{g_0,T,\lambda}(x, z|a) = p_{g_0}(x|z)p_{T,\lambda}(z|a) \tag{4}$$

Further assume:

1. *The sufficient statistic $T(z) = (T_i(z_i))_{i=1}^{d_Z}$ is differentiable and its Jacobian is full-rank a.e*
2. *There exist $d_Z + 1$ distinct points $u^0, ..., u^{d_Z}$ such that the matrix*

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

is invertible.

Then $\forall \tilde{g}, \tilde{\lambda}, \tilde{T}$,

$$p_{g_0,\lambda,T}(x|a) = p_{\tilde{g},\tilde{\lambda},\tilde{T}}(x|a) \Longrightarrow g_0(x) = R\tilde{T}(\tilde{g}(x)) + c, \forall x \in \mathcal{X}.$$

where $R \in \mathbb{R}^{d_Z \times d_Z}$ is invertible, $c \in \mathbb{R}^{d_Z}$.

This theorem states that we recover the true latent variables up to an invertible affine transformation and a point-wise nonlinear function T. However, we later show in the proof that sufficient statistic T is the identity in the Gaussian linear case. Second assumption, usually refered as *sufficient variability assumption*, ensures that the individual components of Z are independently modulated by A such that Z is recoverable. This condition is ensured by the fact that M_0 is full-rank.

Remark A.3. The theorem is initially stated in the case of noisy observed data, *i.e.*, $X = g_0(Z) + \varepsilon$. However, as argued in the proof B.2.2 (Step I), the equality in the noisy case implies equality in noise-free probability density (*i.e.*, $X := g_0(Z)$). The rest of the proof relies on a noise-free distribution of X. For simplicity, we conduct experiments in the noise-free case. Nevertheless, the results suggest that the conclusions would extend to the additive noise case, provided the residuals remain independent of the latent features.

Lemma 2.2. *Assume our data are sampled according to Equation* (1)*, with A being a random variable independent from $V \sim \mathcal{N}(0, \Sigma_V)$, an isotropic multivariate Gaussian random variable. Further assume that the support of A contains at least $d_Z + 1$ affinely independent points. Then Z is identifiable up to an affine transformation.*

Proof. Without loss of generality, we prove the result for V a centered multivariate Gaussian with identity covariance matrix. We show that our data generating process verifies all assumptions stated in Theorem A.2. Let us recall that we sample data from the following SCM:

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

with g_0 injective and M_0 full row rank. We have, by independence of A and $V = (V_1, ..., V_{d_z})$:

$$p(z|a) = p_V(z - M_0 a) = \prod_i p_{V_i}(z_i - m^i a)$$

where $\{m^i\}_i$ are the rows of M_0. By hypothesis on the distribution of V, we have:

$$p(z|a) = \prod_i \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z_i - m^i a)^2} \tag{5}$$

$$= \prod_i \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z_i^2 - 2 z_i m^i a + (m^i a)^2)} \tag{6}$$

We recognize a simple exponential distribution as introduced in A.1, with parameters for all $i \in \{1, ..., d_Z\}$:

$$\begin{cases} Q_i(t) = \frac{e^{-t^2/2}}{\sqrt{2\pi}}, & \forall t \in \mathbb{R} \\ Z_i(u) = e^{(m^i u)^2/2}, & \forall u \in \mathbb{R}^{d_A} \\ T_i(t) = t, & \forall t \in \mathbb{R} \\ \lambda_i(u) = m^i u, & \forall u \in \mathbb{R}^{d_A} \end{cases}$$

We verify the second assumption. Let us pick $u^0 = 0_{d_A}$, then $\forall u^1, ..., u^{d_Z} \in supp(A)$ linearly independent:

$$L = (\lambda(u^1), ..., \lambda(u^n)) \tag{7}$$

$$= \begin{bmatrix} m^1 u^1 & m^1 u^2 & \dots & m^1 u^{d_Z} \\ \vdots & \vdots & \vdots & \vdots \\ m^{d_Z} u^1 & m^{d_Z} u^2 & \dots & m^{d_Z} u^{d_Z} \end{bmatrix} \tag{8}$$

$$= M_0 U \tag{9}$$

with $U \in \mathbb{R}^{d_A \times d_Z}$ the matrix formed with columns $\{u^1, ..., u^{d_Z}\}$. Let us set $U := M_0^T$. We note that:

- The matrix $U = M_0^T \in \mathbb{R}^{d_A \times d_Z}$ has rank d_A. If two columns of U were equal, or if a column was zero, then the column space of U would be linearly dependent, implying that $\text{rank}(U) < d_A$. **Therefore, all columns of U are distinct and non-zero**.
- M_0 has full row rank, therefore $L = M_0 M_0^T$ is not only square and symmetric but also positive definite and thus **invertible**.

Finally, we verify that the first assumption holds. T is differentiable a.e. and

$$\frac{\partial T_i}{\partial z_j}(z) = \begin{cases} 1, & i = j \\ 0 & \text{else} \end{cases}$$

So its Jacobian is the identity matrix on $\mathbb{R}^{d_Z \times d_Z}$ and is thus full-rank.

□

A.2 Causal Effect Identifiability

As mentioned, our method assumes that the treatment is a latent variable (unobserved) but recoverable up to an affine transformation using contrastive learning. We argue that this indeterminacy does not affect the indentifiability of the causal-effect. Particularly, if the IV assumptions stated in Section 2.3 hold for latent Z, they also hold for a non-constant affine transformation of Z.

Lemma 2.5. *Let $R \in \mathbb{R}^{d_Z \times d_Z}$ be an invertible matrix, $c \in \mathbb{R}^{d_Z}$ and $\hat{Z} := RZ + c$. If A is a valid instrument for treatment Z according to Assumption 2.3, it is a valid instrument for treatment $\hat{Z}$ along the same assumptions.*

Proof. **Relevance**. Let $\hat{Z} := RZ + c$ with $R \in \mathbb{R}^{d_Z \times d_Z}$ invertible and $c \in \mathbb{R}^{d_Z}$. Using the change of variable formula yields:

$$p(\hat{z}|a) = |\det R^{-1}| \cdot p(z|a). \tag{10}$$

By hypothesis $\exists z$ such that $\frac{\partial p(.|a)}{\partial a}(z) \neq 0$ and R is non-zero. Therefore, the relevance assumption also holds for $\hat{Z}$.

Quasi-exogeneity. Let us verify that the assumption ii) holds for A w.r.t the residuals of $\hat{Z}$. We have:

$$\begin{aligned}\hat{Z} &= RZ + c \\ &= RM_0A + RV + c.\end{aligned}$$

We identify the residuals of $\hat{Z}$ as: $\hat{V} := RV + c$, similarly the control function in the unobserved case is now $\tilde{l} = l \circ \tau$, with $\tau(z) := R^{-1}z - c$. We then have $\mathbb{E}[\tilde{l}(V)|A] = \mathbb{E}[l(V)|A] = \textit{constant}$. **Exogeneity**. Finally, $z \mapsto Rz + c$ is an affine and invertible transformation and thus a measurable function. since $V \perp\!\!\!\perp A$ it follows that $\hat{V} \perp\!\!\!\perp A$. □

Let us verify Equation (2). Assumption 2.3 ii) implies that $\mathbb{E}[\varepsilon|A] = 0$, we then have:

$$\begin{aligned}\mathbb{E}[Y|A] &= \mathbb{E}[f_0(Z) + l(V) + \varepsilon|A] \\ &= \mathbb{E}[f_0(Z)|A] + \mathbb{E}[l(V)|A] \\ &= \mathbb{E}[f_0(Z)|A] + \textit{constant}.\end{aligned}$$

This constant term emerges because $\mathbb{E}[l(V)]$ may have a non-zero mean, while classic IV setting usually assumes that the expectation of residuals conditioned on the treatment is 0, which would highly restrict function l in our case. However, this difference does not invalidate IV. Let $c := \mathbb{E}[l(V)|A]$ and $\tilde{\varepsilon} := l(V) - c + \varepsilon$, we have:

$$\begin{aligned}Y &= f_0(Z) + l(V) + \varepsilon \\ &= \big(f_0(Z) + c\big) + \big(l(V) - c + \varepsilon\big) \\ &= f_1(Z) + \tilde{\varepsilon},\end{aligned}$$

where $f_1(z) := f_0(z) + c$. We have $\mathbb{E}[\tilde{\varepsilon}|A] = 0$ and therefore:

$$\mathbb{E}[Y|A] = \mathbb{E}[f_1(Z)|A].$$

This additive constant c yields a location indeterminacy: the true causal-effect f_0 is only recovered up to an additive constant.

Let us now verify Equation (3). We recall that $\varepsilon \perp\!\!\!\perp V$, thus:

$$\begin{aligned}\mathbb{E}[Y|Z,V] &= \mathbb{E}[\mathbb{E}[f_0(Z) + l(V) + \varepsilon|Z,V] \\ &= \mathbb{E}[f_0(Z) + l(V)|Z,V] \\ &= f_0(Z) + l(V).\end{aligned}$$

As mentioned before, our difference with the classic IV setting lies in that we do not directly observe treatment Z but rather a nonlinear mixing of it $X = g_0(Z)$. Luckily we demonstrated in 2.2 that latent treatment was recoverable up to affine transformation in our IV setting. Furthermore, Newey et al. [1999] argues that f_0 and l are identifiable up to an additive constant if they are differentiable, assuming that we observe the treatment. Since we only observe treatment and thus residuals up to an unknown affine invertible transformation κ, we are actually looking to recover $f_0 \circ \kappa^{-1}$ and $l \circ \kappa^{-1}$, which are obviously also differentiable. As a consequence we recover the true causal effect up to an affine transformation, inverting the transformation indeterminacy of our retrieved latent features.

A.3 Latent Estimation

Following the identifiability result from previous section, we propose to estimate g_0^{-1} and $p(z|a)$ using contrastive methods. We train an encoder $\phi : \mathcal{X} \to \mathcal{Z}$, parametrized as a neural network, to approximate the inverse of ground-truth mixing function g_0. We also train a decoder $\psi : \mathcal{Z} \to \mathcal{X}$ to reconstruct X from the estimated latent $\hat{Z} := \phi(X)$ using reconstruction loss:

$$\mathcal{L}_{recon} = \mathbb{E}[\|X - \psi(\phi(X)\|^2]$$

This loss term ensures that our encoder is injective/invertible, but is generally not sufficient to identify latents Z [Khemakhem et al., 2020]. In addition to the reconstruction loss, we present two contrastive methods exploiting auxiliary variable A to ensure identifiability and learn $p(z|a)$.

A.3.1 Auxiliary Variable

The first approach was introduced by Hyvärinen et al. [2019] and we state its identifiability property in this section. It consists in first creating two datasets:

$$\mathcal{D}_{pos} = (X, A) \text{ and } \mathcal{D}_{neg} = (X, \tilde{A})$$

where $(X, A) \sim P_{X,A}$ and $\tilde{A} \sim P_A$ is sampled independently from $X \sim P_X$. A nonlinear logistic head $h : \mathcal{Z} \times \mathcal{A} \to \{0, 1\}$ parametrized as a neural network, is trained to discriminate between the two datasets:

$$h(x, a) := \sum_{i=1}^{d_z} \psi_i(\phi_i(x), a) \tag{11}$$

where ϕ_i is the *i-th* element of our encoder ϕ and $\psi := (\psi_1, ..., \psi_{d_Z})$ is a neural network. We optimize it using cross-entropy loss to estimate the probability that a given pair belong to the positive or negative pair:

$$\mathcal{L}_{\text{CL}} = \mathbb{E}_{(X,A)\sim P_{X,A}} \left[-\log h(g(X), A)\right] + \mathbb{E}_{X\sim P_X, \tilde{A}\sim P_A} \left[-\log(1 - h(g(X), \tilde{A}))\right]. \tag{12}$$

The following theorem states that an encoder trained on this loss will converge to the true inverse mixing function up to an affine transformation.

Theorem A.4 ([Hyvärinen et al., 2019], *Theorem 3*)**.** *Assume:*

1. $X := g_0(Z)$*, with* g_0 *invertible.*
2. $\{Z_i\}_{i=1,...,d_Z}$ *are conditionally independent and exponential given auxiliary variable A, Definition A.1,* $Z|A = a \sim p_{T,\lambda}(z|a)$
3. *The sufficient statistic* $T(z) = (T_i(z_i))_{i=1}^{d_Z}$ *is differentiable and its Jacobian is full-rank a.e.*
4. *There exist* $n + 1$ *distinct points* $u^0, ..., u^n$ *such that the matrix.*

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

   is invertible.

5. *The regression head* r *has universal approximation capacity and is trained on Cross-Entropy Loss to discriminate between* $\mathcal{D}_{pos}$ *and* $\mathcal{D}_{neg}$*.*

Then in the limit of infinite data, ϕ *identifies ground-truth latent* Z *up to an invertible affine transformation.*

Remark A.5. Theorem A.4 is initially stated in the general exponential case $k \geq 1$ while we are in the simple case with $k = 1$. Moreover, the theorem states that we recover up the true inverse mixing function up to a point-wise transformation (defined by the sufficient statistic), which is the identity in the location-scale Gaussian case (see proof of Lemma 2.2 where we state that the sufficient statistic is the identity).

A.3.2 InfoNCE

Our second proposed approach is similar but leverages InfoNCE loss [van den Oord et al., 2019]. We train an encoder ϕ to maximize the similarity between our estimated latents $z := \phi(x)$ and its corresponding instrument a, and minimize the similarity with $\tilde{a}$ the other instruments from the batch. In essence, in contrast to the previous approach, we do not only compare a positive to a negative pair, but we have $B - 1$ negative pairs in practice, where B is the batch size. The loss is given by:

$$\mathcal{L}_{\text{NCE}} = \mathbb{E}\left[\log \frac{e^{\text{sim}_W(\phi(X),A)}}{\sum_{\tilde{A}\sim P_A} e^{\text{sim}_W(\phi(X),\tilde{A})}}\right] \tag{13}$$

with $\text{sim}_W(Z, A) = Z^T W A$ where $W \in \mathbb{R}^{d_Z \times d_A}$ is a learnable matrix. We later show that in the event of infinite data, an encoder trained on NCE loss identifies the latent up to an affine transformation.

Theorem A.6. *Assume we train an invertible encoder ϕ to minimize $\mathcal{L}_{NCE}$. Then under Assumptions 1-4 of Theorem A.4, ϕ identifies ground-truth latent Z up to an invertible affine transformation and point-wise invertible transformation.*

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

$$W^*, \phi^* = \arg\min_{\phi, W} \mathcal{L}_{\text{NCE}}, \tag{14}$$

then

$$e^{\phi^*(x)W^*a} \propto \frac{p(x|a)}{p(x)}. \tag{15}$$

$$\begin{aligned}
\phi^*(x)W^*a &= \log c + \log p(x|a) - \log p(x) && (16)\\
&= \log c + \log p_Z(g_0^{-1}(x)|a) - \log p_Z\big(g_0^{-1}(x)\big) && (17)\\
&= \log c + \log p_Z(z|a) - \log p_Z\big(z\big) && (18)
\end{aligned}$$

We use the change of variable formula to go from (17) to (18) and notice that the Jacobian determinants cancel themselves. We define c the proportionality constant that is not dependent on a or x. At line (19) we simply set $z := \phi^{-1}(x)$. By assumption, $\{Z_i\}_{i=1,\dots,d_Z}$ given A follow an exponential distribution as introduced in, thus:

$$\begin{aligned}
\phi^*(x)W^*a &= \log p_{T,\lambda}(z|a) - \log p_Z\big(z\big) + \log c && (19)\\
&= \log c + T(z)\lambda(a) + \log Q(z) - \log Z(a) - T_0(z), && (20)
\end{aligned}$$

with $T_0(z) := p_Z(z)$. By collecting these equations for every a_k, $k \in \{0, ..., d_Z\}$ as defined in assumption 4 of Theorem A.4 and taking out the case a_0, we obtain for all $k \in \{1, ..., d_Z\}$:

$$\phi^*(x)W^*(a_k - a_0) = T(z)\big(\lambda(a_k) - \lambda(a_0)\big) + \log \frac{Z(a_0)}{Z(a_k)}, \tag{21}$$

which yield the following matrix form:

$$\phi^*(x)W^*A = T(z)L + C, \tag{22}$$

with A 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 Theorem A.4 assumption 4 and C is a vector of dimension d_Z whose *k-th* element is given by $\log \frac{Z(a_0)}{Z(a_k)}$. By assumption, L is invertible thus we can multiply both side by its inverse, which yields the following result:

$$\phi^*(x)R = T(z) + \tilde{C}, \tag{23}$$

with $\tilde{C} := CL^{-1}$ and $R := W^*AL^{-1}$.

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.

□

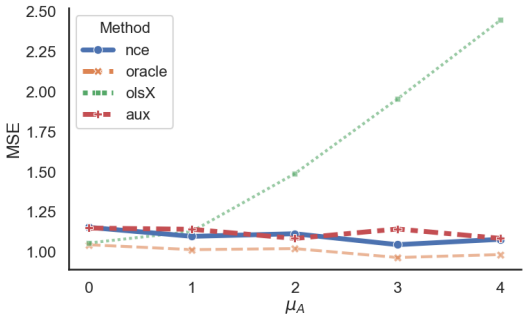


Figure 3: Average MSE score per distribution shift for *two-stages least squares* method. Let us recall that all models are trained on data generated with $A \sim \mathcal{N}(0, I_{d_A})$ and evaluated on data generated with $A \sim \mathcal{N}(\mu_A, I_{d_A})$, with $\mu_A \in \{0, 1, 2, 3, 4\}$. A is of dimension 8, Z of dimension 6 and X of dimension 10. Y is scalar. Each method is run with 10 different seeds.

Our theorem states the identifiability of ground-truth latent variables up to an affine invertible transformation and a point-wise invertible transformation defined by T, the sufficient statistic associated with the distribution of $Z|A$. However, as stated in Remark A.5 the sufficient statistic is the identity in the location-scale Gaussian case.

Lemma A.7. *Assume our data are sampled according to Equation* (1)*, with* $A \sim \mathcal{N}(0, \Sigma_A)$ *and* $V \sim \mathcal{N}(0, \Sigma_V)$ *two independent isotropic multivariate Gaussian random variables. Then an encoder trained on* $\mathcal{L}_{NCE}$ *on infinite data identifies the ground-truth latent up to an invertible affine transformation.*

B Experiments

B.1 Data generation

Let us recall that we sample X, Y, Z, A according to the following SCM:

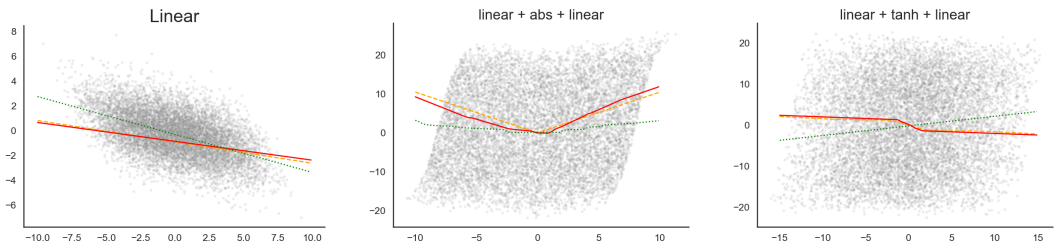
$$\mathcal{S} : \begin{cases} A, V \sim \mathcal{N}(0, I_{d_A}), \mathcal{N}(0, I_{d_Z}) \\ Z := M_0 A + V \\ X := g_0(Z) \\ Y := f_0(Z) + l(V) + \varepsilon \end{cases} \tag{24}$$

In our experimental setup, g_0 is implemented as a two-layer neural network with a hidden dimension of 16. To ensure injectivity, we employ LeakyReLU activations with a negative slope of -0.2 and sample the weight matrices of each layer to be full-rank. Similarly, M_0 is constructed to be full-rank using a QR-decomposition-based procedure: two random matrices are independently sampled, the Q factors from their QR decompositions are extracted, and their product is used as M_0. Since the Q factor of a QR decomposition is orthogonal (and hence invertible), this procedure guarantees that M_0 is full rank by construction. One could just pick Q as M_0, however this would lead to orthogonal matrices only. f_0 and l are two-layers neural network with hidden dimension 16. To preserve their differentiability, we use Tanh as activation functions. We sample 10,000 data points from our training SCM and 2,000 data points for each validation set (with shifted instrument).

As shown in Figure 3, both NCE and Auxiliary Variable methods learn relevant features to perform *control function* in a second step.

B.2 Model and Training specs

Our encoder is a three-layer neural network with 32 hidden units. The NCE method (see Appendix A.3.2) requires only an additional trainable matrix of size $d_A \times d_Z$, whereas the auxiliary variable method (see Appendix A.3.1) uses a logistic head implemented as a two-layer neural network with 16 hidden units, Tanh activations, and a linear output layer. Encoders are trained for 50 epochs. For the final IV regression step, we use an additive model $f_\theta + \nu_\theta$ in the control function



(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*)

approach, where both f_θ and ν_θ are three-layer neural networks with 32 hidden units and Tanh activations. In the two-stage least squares approach, only f_θ is used, with the same architecture. During IV training, the parameters used for latent estimation are frozen, and the model is trained for an additional 50 epochs. The oracle models share the exact same architecture as the corresponding IV models (control function or two-stage). The naive baseline is a single three-layer network with 32 hidden units. Both baselines are trained on 100 epochs.

Finally, we use Adam optimizer with $(\beta_1, \beta_2) = (0.9, 0.999)$. As our experiment are conducted in relatively low dimensions, both training and inference are done on an *Apple M4 Pro* chip.

B.3 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 learnt 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 Appendix A.2, 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.

B.4 Latent IV methods adaptation

In this section, we provide a detailed description of how we adapt the control function and two-stage least squares (2SLS) methods to our latent IV setting. Assume that we have trained an encoder ϕ to invert the true mixing function g_0 using one of the strategies described in Appendix A.3.

We first estimate the linear relationship between the observed instrument A and the encoded latent $\hat{Z} := \phi(X)$ up to an affine transformation:

$$\hat{M}, \hat{c} = \arg\min_{M,c} \|\hat{Z} - MA - c\|^2.$$

B.4.1 Control function approach

1. Compute the predicted latent features $\hat{Z} = \phi(X)$.
2. Estimate the control variable $\hat{V}$ as

$$\hat{V} = \hat{Z} - \hat{M}A - \hat{c}.$$

3. Perform additive regression on Y, estimating

$$\hat{\mu}, \hat{\nu} = \arg\min_{\mu,\nu} \|Y - \mu(\hat{Z}) - \nu(\hat{V})\|^2.$$

B.4.2 Two-Stage Least Squares

1. Compute the predicted latent features from A:

$$\hat{Z}_A = \hat{M}A + \hat{c}.$$

2. Regress Y on $\hat{Z}_A$ to estimate:

$$\hat{h} = \arg\min_{h} \|Y - h(\hat{Z}_A)\|^2.$$

C 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, Ilyes Khemakhem [2020] 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.