Controlling for discrete unmeasured confounding in nonlinear causal models

Patrick Burauel
California Institute of Technology
Pasadena, CA 91125, USA
pburauel@caltech.edu

Frederick Eberhardt
California Institute of Technology
Pasadena, CA 91125, USA
fde@caltech.edu

Michel Besserve
Max Planck Institute for Intelligent Systems
72076 Tübingen, Germany
michel.besserve@tuebingen.mpg.de

Abstract

Unmeasured confounding is a major challenge for identifying causal relationships from non-experimental data. Here, we propose a method that can accommodate unmeasured discrete confounding. Extending recent identifiability results in deep latent variable models, we show theoretically that confounding can be detected and corrected under the assumption that the observed data is a piecewise affine transformation of a latent Gaussian mixture model and that the identity of the mixture components is confounded. We provide a flow-based algorithm to estimate this model and perform deconfounding. Experimental results on synthetic and real-world data provide support for the effectiveness of our approach.

1 Introduction

One of the fundamental challenges of causal inference is the separation of the causal effect from confounding, that is, from statistical dependencies that arise from common causes of the candidate cause and effect. In Pearl's notation [27], this difference is captured by the key contrast between the merely predictive conditional probability $P(Y|X)$ and the causal effect $P(Y|\mathrm{do}(X))$. When confounding variables are observed, confounding can be controlled for by a variety of covariate adjustment techniques [12, 1]. The ability to also deconfound the causal effect in the case of *unobserved* confounding is one of the motivations for the use of randomized controlled trials. The challenge of how to deconfound the causal effect *without experimentation* has given rise to a variety of approaches that require different assumptions for identification. These include instrumental variable approaches [12], approaches based on parametric assumptions (such as in additive noise models [31, 10], linear models [15, 16] or binary Gaussian mixture models [6]), or settings where observed confounding is assumed to be representative of unobserved confounding [2].

In this paper, we contribute to the effort to address unmeasured confounding in purely observational settings by imposing restrictions on the model class. Unlike previous work, we do this by reformulating a confounded cause-effect model as an equivalent latent variable model with a Gaussian mixture prior (see Figure 1). We then leverage the results in [21] that assure identification (up to an affine transformation) of the latent Gaussian mixtures under the assumption of a piecewise affine mapping between latent and observed variables. We show that further constraints on this model specific to our setting (notably causal order) allow to identify causal effects despite (discrete) unobserved confounding. Implementing this approach with a flow-based deep generative model, we show on both synthetic and real data how to estimate the desired causal effects despite unmeasured confounding.

Preprint. Under review.

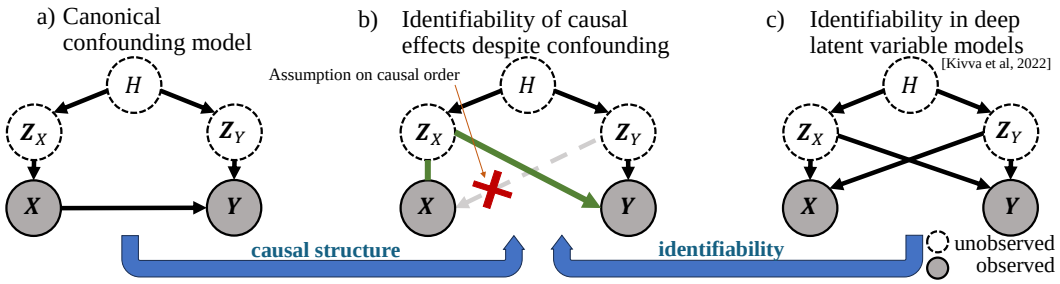

Figure 1: On the left, $\boldsymbol{X}$ causes $\boldsymbol{Y}$ and is confounded by H. On the right, observed variables $\boldsymbol{W} = (\boldsymbol{X}, \boldsymbol{Y})$ are generated by latent variables $\boldsymbol{Z}$, whose identifiability up to affine transformation under model restrictions is shown by [21]. We combine knowledge of causal structure with identifiability results for latent variable models to estimate causal effects despite unmeasured confounding (middle).

Notations. We will use uppercase letters for random variables (e.g. X) and lowercase for deterministic ones (e.g. a realization x of X). Functions and variables that may be vector-valued will be denoted in bold (e.g. $\boldsymbol{X}$, $\boldsymbol{f}$, ...), and $^\top$ denotes transposition. We will use non-bold capital letters for (deterministic) matrices, e.g. A. $P(.)$ denotes a probability distribution, while $p(.)$ denotes the corresponding density with respect to the Lebesgue measure.

2 Background

Canonical cause-effect model in causal inference. In causal inference, the canonical cause-effect model "$\boldsymbol{X}$ causes $\boldsymbol{Y}$" can be represented by a pair of so-called *structural equations* [27]:

$$\boldsymbol{X} \coloneqq \boldsymbol{f}_X(\boldsymbol{Z}_X)\,, \quad \boldsymbol{Y} \coloneqq \boldsymbol{f}_Y(\boldsymbol{X}, \boldsymbol{Z}_Y)\,, \quad \text{with} \quad (\boldsymbol{Z}_X, \boldsymbol{Z}_Y) \sim P_Z(\boldsymbol{Z}_X, \boldsymbol{Z}_Y)\,, \tag{2.1}$$

where the exogenous variables $(\boldsymbol{Z}_X, \boldsymbol{Z}_Y)$ are idiosyncratic error terms representing the influence of external factors on the system, and $(\boldsymbol{f}_X, \boldsymbol{f}_Y)$ are the causal mechanisms associated to each variable. Causal effects of interests are entailed by the mechanism $\boldsymbol{f}_Y$ that describes the influence of $\boldsymbol{X}$ on $\boldsymbol{Y}$. Confounding then posits the existence of a common cause H that influences both idiosyncratic error terms, such that they become dependent when marginalizing with respect to H, leading to

$$P_Z(\boldsymbol{Z}_X, \boldsymbol{Z}_Y) = \textstyle\sum_h P(\boldsymbol{Z}_X|H=h)P(\boldsymbol{Z}_Y|H=h)P(H=h) \neq P_{\boldsymbol{Z}_X}(\boldsymbol{Z}_X)P_{\boldsymbol{Z}_Y}(\boldsymbol{Z}_Y)\,,$$

as depicted in the causal diagram of Figure 1a. Accounting for this dependence is necessary for the unbiased estimation of the causal effect but is difficult as $\boldsymbol{Z}_X$, $\boldsymbol{Z}_Y$ and H are typically unobserved.[1]

Identifiability of latent variable models. The field of *latent variable models* (LVM) [20, 25] addresses the learnability of models mapping latent variables $\boldsymbol{Z}$ to observations $\boldsymbol{W}$ using a so-called mixing function $\boldsymbol{\Psi}$ such that $\boldsymbol{W} = \boldsymbol{\Psi}(\boldsymbol{Z})$, using only samples from the observation distribution $P(\boldsymbol{W})$. Identifiability results provide guaranties that, given infinite data, the ground truth $(\boldsymbol{\Psi}, \boldsymbol{Z})$ can be recovered from $P(\boldsymbol{W})$ in the large sample limit, up to well-characterized ambiguities. We build on results presented by [21], who consider a generative model for observed variables $\boldsymbol{W}$ of the form:

$$\begin{aligned} H &\sim \text{Cat}(K_H, \boldsymbol{\pi})\,, \\ \boldsymbol{Z} \mid H=h &\sim \mathcal{N}(\boldsymbol{\mu_h}, \Sigma_h)\,, \\ \boldsymbol{W} &= \boldsymbol{\Psi}(\boldsymbol{Z}), \end{aligned}$$

where $\text{Cat}(K, \boldsymbol{\pi})$ denotes a categorical distribution with K categories and an associated vector of event probabilities $\boldsymbol{\pi}$. Assuming that $\boldsymbol{\Psi}$ is a piecewise affine injective function (which can be implemented by ReLU networks), [21] show identifiability of $\boldsymbol{\Psi}$ and $\boldsymbol{Z}$ up to an affine transformation [21, Theorem 3.2]. This model is depicted in Figure 1c.

[1] We provide a brief description of the formalism of structural causal models in Appendix B.

3 Theoretical framework for discrete deconfounding

3.1 General setting

Mapping cause-effect models to LVMs. We consider the above cause-effect model in a setting where an observed n-dimensional vector $\boldsymbol{X}$ causes an observed m-dimensional effect vector $\boldsymbol{Y}$, and where, as commonly assumed, exogenous variables have matching dimensions, i.e. $\boldsymbol{Z}_X \in \mathbb{R}^n$ and $\boldsymbol{Z}_Y \in \mathbb{R}^m$.[2] We explore the idea that exogenous variables $\boldsymbol{Z}_X, \boldsymbol{Z}_Y$ and mechanisms $\boldsymbol{f}_X, \boldsymbol{f}_Y$ can be used to construct a corresponding LVM, from which we can then leverage the identifiability results to address unmeasured confounding. The key ideas are the following: We can replace the generative mechanism of $\boldsymbol{Y}$ based on $\boldsymbol{X}$ by one based on $\boldsymbol{Z}_1$ by rewriting

$$\boldsymbol{Y} \coloneqq \boldsymbol{f}_Y(\boldsymbol{X}, \boldsymbol{Z}_Y) = \boldsymbol{f}_Y\left(\boldsymbol{f}_X(\boldsymbol{Z}_X), \boldsymbol{Z}_Y\right) \triangleq \boldsymbol{\Psi}_Y(\boldsymbol{Z}_X, \boldsymbol{Z}_Y). \tag{3.1}$$

If we additionally introduce $\boldsymbol{\Psi}_X(\boldsymbol{Z}_X, \boldsymbol{Z}_Y) \triangleq \boldsymbol{f}_X(\boldsymbol{Z}_X)$ and concatenate the exogenous variables into the latent vector $\boldsymbol{Z} = (\boldsymbol{Z}_X, \boldsymbol{Z}_Y)$, we can build a well-defined mapping $\boldsymbol{\Psi} : \mathbb{R}^{m+n} \mapsto \mathbb{R}^{m+n}$ from exogenous latent variables to observed variables $\boldsymbol{W} = (\boldsymbol{X}, \boldsymbol{Y})$ such that $\boldsymbol{\Psi}(\boldsymbol{Z}) = \big(\boldsymbol{\Psi}_X(\boldsymbol{Z}), \boldsymbol{\Psi}_Y(\boldsymbol{Z})\big)$. This corresponds to the LVM diagram of Figure 1c. Analogous to the causal model in Figure 1a, confounding is induced by a latent variable H that causes both $\boldsymbol{Z}_X$ and $\boldsymbol{Z}_Y$.

Leveraging LVM identifiability to address confounding. Concretely, to connect LVM identifiability to causal deconfounding, we introduce the following assumptions on the cause-effect model.

Assumption 3.1. The function $\boldsymbol{f}_Y : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ is Continuous Deterministic Piecewise Affine (CDPA)[3] and for all $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{z}_Y \mapsto \boldsymbol{f}_Y(\boldsymbol{x}, \boldsymbol{z}_Y)$ is injective.

Additionally, we make an assumption about the relation between $\boldsymbol{Z}_X$ and $\boldsymbol{X}$:

Assumption 3.2. $\boldsymbol{f}_X : \mathbb{R}^n \to \mathbb{R}^n$ is CDPA and invertible.

In combination, these two assumptions will ensure the mapping $\boldsymbol{\Psi}$ belongs to the function class analyzed in [21]. The final key to identifiability is a Gaussian mixture model of the exogenous variables and their confounding induced by H.

Assumption 3.3. The exogenous variables are generated according to the following model:

$$H \sim \mathrm{Cat}(K_H, \boldsymbol{\pi})\,, \tag{3.2}$$

$$L|H \sim \mathrm{Cat}(K_L, p(L|H))\,, \qquad Q|H \sim \mathrm{Cat}(K_Q, p(Q|H))\,, \tag{3.3}$$

$$\boldsymbol{Z}_X|L{=}l \sim \mathcal{N}(\boldsymbol{\mu}_l, \Sigma_l^X)\,, \qquad \boldsymbol{Z}_Y|Q{=}q \sim \mathcal{N}(\boldsymbol{\nu}_q, \Sigma_q^Y)\,, \tag{3.4}$$

where at least one mixture component l that occurs with non-zero probability has Σ_l^X positive definite.

Note that, without loss of generality, we make the separation of the effect of H on the cause vs. the effect side explicit with Eq. (3.3). We now turn to proving that this model setup and the discussed assumptions allow us to identify causal quantities.

3.2 Identifiability

Theorem 3.4. *Under Assumptions 3.1, 3.2, and 3.3 the mixture components and the causal mechanism for the effect $(\boldsymbol{Z}_Y, \boldsymbol{f}_Y)$ in Eq. (3.1) is identifiable up to an invertible affine reparameterization of $\boldsymbol{Z}_Y$. More precisely, let $(\tilde{\boldsymbol{Z}}_Y, \tilde{\boldsymbol{f}}_Y)$ be the latent variable and mechanism obtained by fitting the model to the observation distribution $P(\boldsymbol{X}, \boldsymbol{Y})$, then we have, for some $(m \times m)$ invertible matrix S and some $(m \times 1)$ vector $\boldsymbol{b}$*

$$\boldsymbol{f}_Y(\boldsymbol{x}, \boldsymbol{z}_Y) = \tilde{\boldsymbol{f}}_Y(\boldsymbol{x}, S\boldsymbol{z}_Y + \boldsymbol{b})\,, \quad \textit{and} \quad \tilde{\boldsymbol{Z}}_Y = S\boldsymbol{Z}_Y + \boldsymbol{b}\,.$$

Sketch of the proof (see Appendix A for the complete version). We will consider a latent variable model solution $\tilde{\boldsymbol{\Psi}} : \boldsymbol{Z} \to \boldsymbol{W}$ satisfying all assumptions and fitting the observational distribution $P(\boldsymbol{X}, \boldsymbol{Y})$ perfectly. We study its relationship to the corresponding ground truth mapping $\boldsymbol{\Psi}$

[2] The special cases of scalar cause and/or effect are included.

[3] CDPA functions can be easily implemented by feedforward neural networks with ReLU activation functions.

which generates the observations. This will then be linked to the cause-effect model solution $\tilde{\boldsymbol{f}}_Y$ and its associated ground truth model $\boldsymbol{f}_Y$. The demonstration can be decomposed into three parts:

(1) The identifiability theory in [21, Theorem 3.2] implies that the latents $\boldsymbol{Z}$ can be recovered up to an affine transformation; more formally, the map $\tilde{\boldsymbol{\Psi}}^{-1} \circ \boldsymbol{\Psi}$ associating ground truth latents $\boldsymbol{Z}$ to recovered ones $\tilde{\boldsymbol{Z}}$ is an affine transformation with its linear map represented by a square matrix A. In addition, the constraint on the causal order enforces that $\boldsymbol{\Psi}_X$ is not dependent on $\boldsymbol{Z}_Y$, which imposes a block triangular structure on A, encoding that the true $\boldsymbol{Z}_Y$ does not influence the recovered $\tilde{\boldsymbol{Z}}_X$.

(2) By Assumption 3.3 the mixture components' cross-covariance matrices between $\boldsymbol{Z}_X$ and $\boldsymbol{Z}_Y$ coordinates is zero for both the ground truth $\boldsymbol{Z}$ and recovered $\tilde{\boldsymbol{Z}}$. Identification up to affine transformation and permutation of these mixture components further constrains the relation between ground truth and recovered latents by forcing the matrix A to be block diagonal.

(3) The final relation between ground truth and recovered cause-effect model is deduced from the shared structure of $\tilde{\boldsymbol{\Psi}}$ and $\boldsymbol{\Psi}$, and the block diagonality of A. □

Note that the results by [21] alone, allow the ambiguity of the identifiability results to be a general affine transformation without any restriction, which precludes the separation of the causal and the confounded variation in the observed $\boldsymbol{Y}$ and consequently the identification of the causal effect.

Provided the data generating process fits our assumptions, then our result guarantees that, in the infinite sample limit, we retrieve the ground truth causal mechanism up to some ambiguities. We now show that these remaining ambiguities do not affect our ability to estimate causal quantities such as the average treatment effect.

Estimation of causal effects. We now show that Theorem 3.4 implies that the average treatment effect is identifiabile, even though $P(L, H, Q)$ may remain unidentified. Given the graph in Figure 1b, we can see that $\boldsymbol{Z}_Y$ satisfies the backdoor criterion [27], such that we can estimate the following interventional quantities by the adjustment formula:

$$\mathbb{E}\left[\boldsymbol{Y}|\mathrm{do}(\boldsymbol{X}=\boldsymbol{x})\right] = \int \boldsymbol{y}\, p\left(\boldsymbol{y}|\mathrm{do}(\boldsymbol{X}=\boldsymbol{x})\right) d\boldsymbol{y} = \iint \boldsymbol{y}\, p\left(\boldsymbol{y}|\boldsymbol{X}=\boldsymbol{x}, \boldsymbol{z}_Y\right) d\boldsymbol{z}_Y d\boldsymbol{y}\,. \tag{3.5}$$

That is, Theorem 3.4 provides the basis to deconfound the causal effect:

Proposition 3.5. *Under the assumptions of Theorem 3.4, assume additionally strict positivity of $p(\boldsymbol{x}, \boldsymbol{z}_Y)$ for almost all $\boldsymbol{z}_Y$. Then, for any $\boldsymbol{x}$ in the support of $P(\boldsymbol{X})$, $\mathbb{E}\left[\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x})\right]$ is identifiable from the observation of $P(\boldsymbol{X}, \boldsymbol{Y})$ with adjustment formula*

$$\mathbb{E}\left[\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x})\right] = \mathbb{E}_{\boldsymbol{Z}_Y \sim P(\boldsymbol{Z}_Y)}\left[\tilde{\boldsymbol{f}}_Y(\boldsymbol{x}, S\boldsymbol{Z}_Y + \boldsymbol{b})\right] = \mathbb{E}_{\tilde{\boldsymbol{Z}}_Y \sim P(\tilde{\boldsymbol{Z}}_Y)}\left[\tilde{\boldsymbol{f}}_Y(\boldsymbol{x}, \tilde{\boldsymbol{Z}}_Y)\right]\,, \tag{3.6}$$

where $P(\tilde{\boldsymbol{Z}}_Y)$ and $\tilde{\boldsymbol{f}}_Y$ is the solution identified in Theorem 3.4.

See Appendix A for the proof. Importantly, we cannot rely on $\boldsymbol{Z}_1$ as an adjustment variable, as it violates positivity by construction of our model (it is deterministically related to $\boldsymbol{X}$), in line with the point made by [3]. Positivity of $p(\boldsymbol{x}, \boldsymbol{z}_Y)$ is achieved under mild assumptions: it only requires the occurrence of one non-degenerate mixtures component of $\boldsymbol{Z}$ in the observational setting.

Proposition 3.6. *If there exists (l, q) such that $P(L=l, Q=q) > 0$ and both Σ_l^X and Σ_q^Y are positive definite, then the positivity assumption on $p(\boldsymbol{x}, \boldsymbol{z}_Y)$ in Proposition 3.5 is satisfied.*

See Appendix A for the proof. Overall, the positive definite assumptions required on covariance matrices in Theorem 3.4 and Proposition 3.6 emphasize the importance of having independent (Gaussian) noise injected in both mechanism $\boldsymbol{f}_X$ and $\boldsymbol{f}_Y$ for identification.

4 Flow-based implementation

We use flow-based models [25] to estimate the discrete confounding model. Such models learn the (possibly complex) distribution of observed data by using successive transformations of a simpler base distribution. The trained model can then be used to sample from the data distribution. This generative aspect of flow-based models lends itself to our deconfounding application as it allows

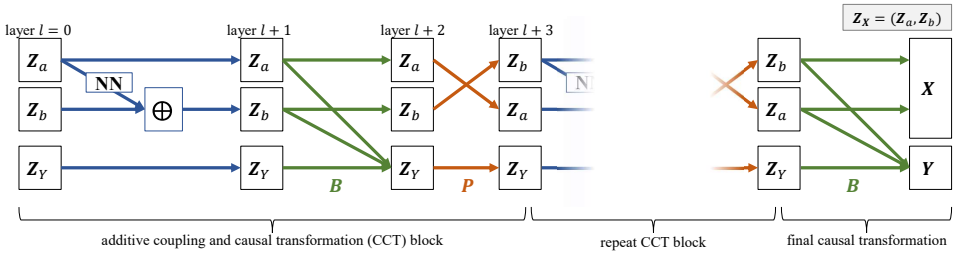

Figure 2: (Flow model implementation) The sequence of transformations that make up one block are composed of an additive coupling bijection from layer l to $l+1$, see lines 5 and 6, a causal transformation with a partly-diagonal structure ($\boldsymbol{Z}_Y$ node does not influence other nodes), see line 7, from $l+1$ to $l+2$, and a permutation layer from $l+2$ to $l+3$. Line numbers refer to Algorithm 1.

us to sample from $P(\tilde{\boldsymbol{Z}}_Y)$, which is the latent variable that blocks the backdoor path and is used in Eq. (3.6). Unlike other generative models such as Variational Autoencoders, flow-based models allow optimization of the exact likelihood of the data, which seems to be critical for their use to estimate causal quantities precisely. Variational Autoencoders with a Gaussian mixture prior [17], as used in experimental section of [21], have proven not to perform as well as flow-based models for the application at hand.[4]

In flow-based models, observed variables $\boldsymbol{w} := (\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^{m+n}$ are expressed as a transformation T of $\boldsymbol{z}$, $\boldsymbol{w} = T(\boldsymbol{z})$, sampled from a base distribution $p(\boldsymbol{z})$. Requiring T to be differentiable and invertible licences the use of the change of variables formula to express the log-likelihood of the data as $\log p_{\boldsymbol{w}}(\boldsymbol{w}) = \log p_{\boldsymbol{z}}(\boldsymbol{z}) + \log |\det J_T(\boldsymbol{z})|^{-1}$ or, using that $\boldsymbol{z} = T^{-1}(\boldsymbol{w})$ and swapping inverse and determinant,

$$\log p_{\boldsymbol{w}}(\boldsymbol{w}) = \log p_{\boldsymbol{z}}(T^{-1}(\boldsymbol{w})) + \log |\det J_{T^{-1}}(\boldsymbol{w})|. \tag{4.1}$$

The log-likelihood of the data can thus be expressed by evaluating the base distribution at the transformed $\boldsymbol{w}$ and accounting for the resulting change in volume by adding the log determinant of the inverse Jacobian of that transformation. To represent the Gaussian mixture structure of the latent variables in our generative model, see Eq. (3.4), we use a Gaussian mixture model as a base distribution.[5] The GMM is characterized by mixture weights (π_k), means ($\boldsymbol{\mu}_k$) and covariances (Σ_k):

$$p(\boldsymbol{z}) = \textstyle\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_k, \Sigma_k), \tag{4.2}$$

where K is the number of mixture components, π_k are the mixture weights, and $\mathcal{N}(\boldsymbol{z}; \mu_k, \Sigma_k)$ with diagonal covariance matrix denotes the Gaussian distribution for component k.

Algorithm 1 One `DeconFlow` transformation block, from layer l to $l+3$

1: **Input:** $\boldsymbol{z}^{(l)}$
2: **Output:** $\boldsymbol{z}^{(l+3)}$
3: $\boldsymbol{z}_X^{(l)}, \boldsymbol{z}_Y^{(l)} \leftarrow \text{split}(\boldsymbol{z}^{(l)})$
4: $\boldsymbol{z}_a^{(l)}, \boldsymbol{z}_b^{(l)} \leftarrow \text{split}(\boldsymbol{z}_X^{(l)})$
5: $\boldsymbol{t}^{(l)} \leftarrow f_t(\boldsymbol{z}_a^{(l)})$
6: $\boldsymbol{z}_b^{(l+1)} \leftarrow \boldsymbol{z}_b^{(l)} + \boldsymbol{t}$ (additive coupling)
7: $\boldsymbol{z}^{(l+2)} \leftarrow \boldsymbol{B}\boldsymbol{z}^{(l+1)}$ (causal transform: $\boldsymbol{z}_X \rightarrow \boldsymbol{z}_Y$)
8: $\boldsymbol{z}_X^{(l+3)} \leftarrow \boldsymbol{P}\boldsymbol{z}_X^{(l+2)}$
9: $\boldsymbol{z}_Y^{(l+3)} \leftarrow \boldsymbol{z}_Y^{(l+2)}$

In our causal inference setting, only transformations that respect the causal order of observed variables $\boldsymbol{w}$ are admissible. To ensure that information flows only in the causal direction from $\boldsymbol{x}$ to $\boldsymbol{y}$, we need to restrict the transformations to be lower-triangular. We first introduce a simple one-layer, linear flow, which allows us to introduce the required restriction. In the subsequent section, we introduce a multi-layered model with additive coupling bijections and triangular causal transformations that can express more complex distributions.

[4] We have implemented VAEs with appropriate architectural restrictions in experiments (not reported here) that did not exactly recover the true causal effects even in the simple $m = n = 1$ linear case.

[5] A GMM base distibution in flow-based models has previously been used by e.g. [30].

4.1 One-layer linear flow

In the simplest proof-of-concept model, where we assume we observe 2D Gaussian mixtures in $\boldsymbol{w}$ resulting from linear mechanisms, the transformation T is then a block lower triangular matrix,

$$A = \begin{pmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{pmatrix} . \tag{4.3}$$

The log-likelihood then reduces to $\log p_{\boldsymbol{w}}(\boldsymbol{w}) = \log p_{\boldsymbol{z}}(\boldsymbol{A}^{-1}\boldsymbol{w}) + \sum_{i=1}^{2} \log |a_{ii}|$. We apply this simple model to simulated data with a one-dimensional cause below.

4.2 Additive coupling bijection

To model more complicated distributions of $\boldsymbol{w}$, we propose a flow-based model where one transformation block is composed of an additive coupling layer [4] and a causal tranformation akin to a masked autoregressive layer [26]. Specifically, the transformations in one block are described in Algorithm 1. Superscript (l) denotes layer index, line 3 splits $\boldsymbol{z}_X$ into the first $n/2$ (rounded up if necessary) dimensions (subscript a) and the remaining dimensions (subscript b). The function f_t in line 5 is parameterized by a neural network with ReLU activation function, the transformation matrix in line 7 has a partly-diagonal form,

$$\boldsymbol{B} = \begin{bmatrix} \mathrm{diag}(\boldsymbol{a}) & \boldsymbol{0} \\ \boldsymbol{b} & b_{d,d} \end{bmatrix}$$

with $\boldsymbol{a} = \begin{bmatrix} a_{1,1} & \cdots & a_{d-1,d-1} \end{bmatrix}$ and $\boldsymbol{b} = \begin{bmatrix} a_{d,1} & \cdots & a_{d,d-1} \end{bmatrix}$, and $\boldsymbol{P}$ (only acting on $\boldsymbol{z}_X$, not $\boldsymbol{z}_Y$) in line 8 is a permutation matrix. By restricting $\boldsymbol{B}$ in this way and permuting only $\boldsymbol{z}_X$, we ensure that $\boldsymbol{x}$ influences $\boldsymbol{y}$ (but not vice versa), which reflects the assumed causal structure. Note that lines 5 and 6 differ from widely-used coupling bijections (which would additionally multiply $\boldsymbol{z}_b^{(l)}$ by a factor that is learned by f_t, as proposed in [5]) to ensure that the transformation is piecewise affine, which we require for identifiability. In practice, N_B of such blocks are concatenated as depicted in Figure 2.

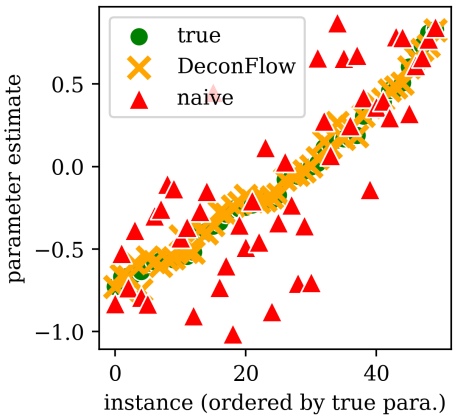

Figure 3: With a one-dimensional cause and one-dimensional confounder, $m = n = 1$, performance can be evaluated by comparing the `DeconFlow`-adjusted slope parameter estimates (orange crosses) to the ground truth (green circles). In addition, we report the naive estimates that are obtained without addressing confounding (red triangles).

We can write the log-likelihood of $\boldsymbol{w}$ given these transformations as

$$\log p_{\boldsymbol{w}}(\boldsymbol{w}) = \log p_{\boldsymbol{z}}(\boldsymbol{z}^{(0)}) + \sum_{l=1}^{L} \sum_{i=1}^{d} \log |a_{ii}^{(l)}| \tag{4.4}$$

where $\boldsymbol{z}^{(0)} = \overline{T}\boldsymbol{w}$ with $\overline{T} = T_{(l=0)} \circ \ldots \circ T_{(l=L)}$ denoting the composition of the transformations described above (similarly for its inverse, $\overline{T}^{-1}$) and $p_{\boldsymbol{z}}$ being a Gaussian mixture model with diagonal covariances, as in Eq. (4.2). The transformation in line 6 is volume-preserving and has a unit Jacobian determinant. Therefore, its logarithm is equal to zero and vanishes in the log likelihood. Since the Jacobian of $\boldsymbol{B}$ is lower-triangular, its determinant is the product of the diagonal elements. We then optimize the log-likelihood in Eq. (4.4) using backpropagation.

4.3 Closing the backdoor path through sampling

Given our model structure, conditioning on $\boldsymbol{Z}_Y$ blocks the backdoor path between $\boldsymbol{X}$ and $\boldsymbol{Y}$. This motivates the following strategy to estimate $\mathbb{E}[Y|\mathrm{do}(\boldsymbol{X} = \boldsymbol{x})]$ from observed data. We transform the observed samples of $\boldsymbol{w}$ to $\boldsymbol{z}$ by inverting $\boldsymbol{\Psi}$ using our trained model. We then sample N_p times from the empirical distribution of $\tilde{\boldsymbol{Z}}_Y$ to compute

$$\overline{\boldsymbol{w}} = (\boldsymbol{x}, \overline{\boldsymbol{y}}) = \frac{1}{N_p} \sum_{\tilde{\boldsymbol{z}}_Y \sim P(\tilde{\boldsymbol{Z}}_Y)}^{N_p} \overline{T}(\boldsymbol{z}_X, \tilde{\boldsymbol{z}}_Y) \,, \tag{4.5}$$

where $\overline{\boldsymbol{x}} = \boldsymbol{x}$ because $\boldsymbol{f}_X$ is invertible. This yields the empirical counterpart to Eq. (3.6),

$$\mathbb{E}[Y|\mathrm{do}(\boldsymbol{X} = \boldsymbol{x})] \approx \overline{\boldsymbol{y}} =: \hat{\theta}(\boldsymbol{x}). \tag{4.6}$$

5 Simulation Study

5.1 Data Generation

Given the generative model, we simulate data from a Generalized Additive Model (GAM, [8]) as follows. First, we randomly generate parameters of the joint distribution $P(L, Q)$ such that there is a correlation between L and Q. Second, we generate $\boldsymbol{Z}_X \sim \mathcal{N}(\boldsymbol{\mu}_{h_X}, \Sigma_{h_X})$ and $Z_Y \sim \mathcal{N}(\mu_{h_Y}, \sigma^2_{h_X})$ where $\boldsymbol{\mu}_{h_X} \sim \mathcal{U}(1, 4)$ and $\mu_{h_Y} \sim \mathcal{U}(0, 1)$, $\Sigma_{h_X} = \boldsymbol{I} \times 0.01$ and $\sigma^2_{h_X} = 0.01$. We focus on the case with $m = 1$, a scalar effect, in the simulation study.

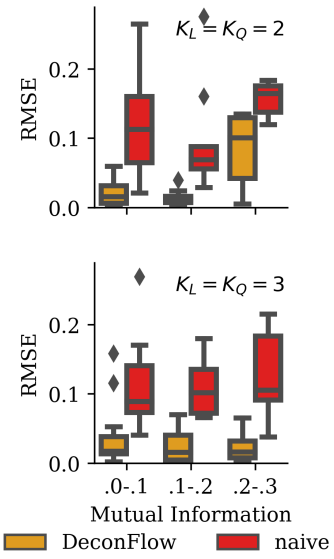

Figure 4: See Section 5.2 for description.

To generate $\boldsymbol{X}$ and Y, we then parameterize the influence of $\boldsymbol{Z}_X$ on X and Y as well as the influence of Z_Y on Y with random CDPA functions,

$$\boldsymbol{X} = \tau_1(\boldsymbol{Z}_X), \text{ and } Y = \beta\tau_2(\boldsymbol{Z}_X) + \tau_3(Z_Y) + \varepsilon, \tag{5.1}$$

where β is the true causal effect, and $\varepsilon \sim \mathcal{N}(0, 0.01)$. Inspired by [9], the functions τ_1, τ_2, and τ_3 are randomly initialized residual-flow type neural networks designed to generate an invertible piecewise affine transformation of data. The architecture consists of an initial linear layer, followed by a series of five ResNet blocks, and concludes with a final linear layer to produce the transformed output. Each ResNet block contains two linear layers with LeakyReLU activations and a skip connection, which adds the input of the block to its output. Note that the model class described in Eq. 5.1 is not covering the whole set of models considered in the theory. Notably, the effects of $\boldsymbol{Z}_X$ and Z_Y on Y are not required to be additive for our theoretical results to hold.

Evaluation metric in linear case with $n = m = 1$. When τ_1, τ_2, and τ_3 are identity mappings, we evaluate the ability of our method to deconfound by comparing the estimated slope parameter with the true causal effect β. In the linear case, the estimated parameter can be read off the estimate of the transformation matrix A in (4.3): $\hat{\beta} = \frac{a_{21}}{a_{11}}$.

Evaluation metric in the nonlinear case. When τ_1, τ_2, and τ_3 are random injective mappings, we evaluate the ground truth $\theta^*(\boldsymbol{x}) := \mathbb{E}[Y|do(\boldsymbol{X} = \boldsymbol{x})]$ using Eq. (3.6) but for the ground truth model. We compare $\theta^*(\boldsymbol{x})$ with the estimate defined in Eq. (4.6):

$$\mathrm{RMSE} = \sqrt{\mathbb{E}_{\boldsymbol{x}\sim P(\boldsymbol{X})}\left[\left(\hat{\theta}(\boldsymbol{x}) - \theta^*(\boldsymbol{x})\right)^2\right]} \tag{5.2}$$

For comparison, we report a baseline RMSE that is obtained when the conditional density is erroneously used as a causal effect estimate:

$$\mathrm{RMSE}_{\text{naive}} = \sqrt{\mathbb{E}_{\boldsymbol{x}\sim P(\boldsymbol{X})}\left[\left(\mathbb{E}(Y|\boldsymbol{x}) - \theta^*(\boldsymbol{x})\right)^2\right]}. \tag{5.3}$$

5.2 Results

Linear one-layer, identity mapping. First we generate 10,000 samples for the simple setting when $n = m = 1$, and τ_1, τ_2, τ_3 all being identity mappings, with $K_L = K_Q = 2$, and apply the simple one-layer linear flow described in Section 4.1. In this case, the observed data *is* a Gaussian mixture. Therefore, we have a setting in which the estimation procedure focuses solely on disentangling causal from confounded variation without additionally learning the mapping from observed data to

a Gaussian mixture model. This setting serves as proof-of-concept of the deconfounding strategy. Results are shown in Figure 3. It can be seen that the naive parameter estimates that are obtained by regressing observed Y on observed X are biased in arbitrary directions. Using `DeconFlow`, we recover estimates of $\mathbb{E}[Y|\mathrm{do}(X = x)]$, which we regress on x to compute the deconfounded parameter estimates that almost perfectly match the ground truth.[6]

Nonlinear, invertible piecewise affine transformations. Next we generate data with $n = 5$, $m = 1$ and τ_1, τ_2, τ_3 random invertible piecewise affine functions (as described in Section 5.1) and $K_L = K_Q = k$ for $k \in \{2, 3\}$, 10,000 observations. Figure 4 shows RMSE, see Eq. (5.2), and $\mathrm{RMSE}_{\mathrm{naive}}$, see Eq. (5.3). The x-axis shows mutual information between discrete variables L and Q as a measure for the strength of confounding. `DeconFlow` decreases the error incurred when estimating $\mathbb{E}[Y|\mathrm{do}(\boldsymbol{X} = \boldsymbol{x})]$ without observing the discrete confounder substantially. What we achieve here is the estimation of a nonlinear causal quantity, $\mathbb{E}[Y|\mathrm{do}(\boldsymbol{X} = \boldsymbol{x})]$, without observing the latent quantity that induces the discrepancy between it and $\mathbb{E}[Y|\boldsymbol{x}]$.[7]

6 Application

We use data on twin births in the USA collected around 1990, which has been used before by [23] to illustrate causal inference methods. It contains measures of birth weight of newborn twins with about two dozen additional control covariates, such as parental education, number of prenatal visits, etc. for about 32,000 twins (and their parents). See Appendix C for a complete list of variables. The dataset lends itself to our setting because most of the variables are discrete and can serve as confounders. At the same time, some ordinal variables are also recorded. We choose as causes those ordinal variables so that we can approximate them with continuous variables by adding uniformly distributed noise. We do this because our model requires continuous cause variables and discrete confounding variables.

From the set of covariates $\{X_1, \ldots, X_K\}$ we select the three ordinal variables that are directly related to the mother as observed causes: *mother's age, gestation type*, and *mother's education*, and denote them by $\boldsymbol{X} = \{X_1, X_2, X_3\}$. We use *birth weight of the first-born twin* as target variable, Y, and treat all remaining covariates as confounders, denoted by $\boldsymbol{V} = \{X_4, \ldots, X_K\}$. This allows us to estimate "true" causal effects when we treat the confounders as observed, and test whether `DeconFlow` can recover these given only the data about $\boldsymbol{X}$ and Y.

Predicting Y using least-squares regression, we estimate the parameter vector for $\boldsymbol{X}$ once when controlling for $\boldsymbol{V}$ (denoted $\boldsymbol{\beta}^*$) and once when not controlling for $\boldsymbol{V}$ (denoted $\hat{\boldsymbol{\beta}}$). We run our deconfounding approach as described in Section 4.3 using only $\{\boldsymbol{X}, Y\}$, which yields our estimate of $\hat{\theta}(\boldsymbol{x}) = \mathbb{E}[Y|\mathrm{do}(\boldsymbol{X} = \boldsymbol{x})]$. We then regress $\hat{\theta}(\boldsymbol{x})$ on $\boldsymbol{X}$ to estimate our debiased parameter vector, $\tilde{\boldsymbol{\beta}}$. We can evaluate whether our method can account for the confounders $\boldsymbol{V}$ (that are unobserved from its perspective) by comparing $\boldsymbol{\beta}^*$ with $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$.

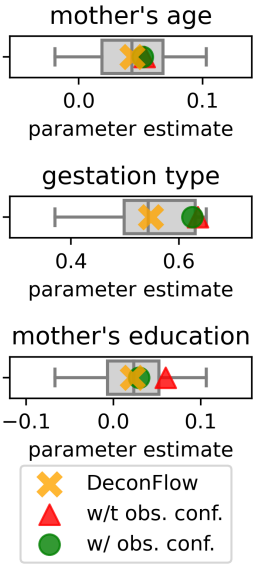

Figure 5: See Section 6 for description.

We run `DeconFlow` for multiple seeds and hyperparameters. In Figure 5, for each of the three cause variables (*mother's age, gestation type*, and *mother's education*), we report *i*) the slope parameter of that cause variable in a regression of Y on the three causes (red triangle), *ii*) the slope parameter of that cause variable in a regression of Y on the three causes and the observed confounders (green dot), *iii*) the average slope parameter of that cause in a regression of the `DeconFlow`-adjusted target variable $\tilde{Y}$ on the three causes for 32 runs of `DeconFlow` (orange cross), as well as a boxplot of the underlying distribution of this parameter. For causes *mother's age* and *mother's education*, we observe that our method yields mean parameter estimates that are closer to $\boldsymbol{\beta}^*$ than $\hat{\boldsymbol{\beta}}$. For *gestation type*, we find $\tilde{\boldsymbol{\beta}}$ to be lower than both $\boldsymbol{\beta}^*$ and $\hat{\boldsymbol{\beta}}$.

[6] Experiments are run on AWS Deep Learning AMI, with 36 vCPUs, runtime about 3 hours.
[7] Experiments are run on AWS Deep Learning AMI, with 96 vCPUs, runtime about 20 hours.

While we consider similar $\boldsymbol{\beta}^*$ and $\tilde{\boldsymbol{\beta}}$ as evidence that our method accounts for $\boldsymbol{V}$ without observing it, we stress that $\boldsymbol{\beta}^*$ might in fact differ from the true parameter vector because of residual confounding that is not captured by $\boldsymbol{V}$. That is, a discrepancy between $\boldsymbol{\beta}^*$ and $\tilde{\boldsymbol{\beta}}$ might indicate the existence of additional confounders unmeasured in the dataset, rather than a shortcoming of our method. For instance, the discrepancey between $\tilde{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^*$ for *gestation type* could be due to additional unmeasured confounders.

7 Discussion

While there is a large literature on using measured confounders to deconfound causal effect estimates (see e.g. [1]), or to gauge the sensitivity to unmeasured confounders by benchmarking against *measured* confounders in treatment effect estimation [2] or policy learning [18, 24], work on accounting for unmeasured confounders without such benchmarks is scarce. In the following we provide a brief overview of related work that addresses unmeasured confounding without access to observed confounders.

One way to tackle unmeasured confounding is to make assumptions on the independence of causal mechanisms (ICM) [28, 14]. For instance, [15, 16] formalize ICM in multivariate linear models to estimate a degree of confounding. ICM can also be seen as motivating additive noise models as used in [13], which is similar to our approach in the sense that a latent confounder is learned from observed variables. However, this method does not allow for both a causal *and* a confounding effect between the two variables.

Even without implicit or explicit motivation through ICM, restricing model classes can help to address unmeasured confounding. For instance, assuming linear relations and non-Gaussian variables yields identifiability of a number of causal properties [29]. In this model class, [10] show how independent component analysis (ICA) with an overcomplete basis (recovering more source variables than there are observed signals), can help to theoretically identify, up to some remaining ambiguity, the latent confounder and causal effect. However, practical algorithms that reliably estimate an overcomplete basis are lacking and require additional assumptions (such as sparsity of the mixing matrix). Methods for (nonlinear) ICA with equal number of sources and signals include e.g. [19, 11] but these require observed auxiliary information (such as environment variables) or assumptions like ICM [7]. None of these methods can address unmeasured confounding in a principled and practical way, which is the goal of our proposed method.

Limitations. As all causal inference techniques, the proposed methodology relies on assumptions that, if not satisfied, can cast doubt on causal effect estimates that are produced using the method. While the discrete nature of the confounding we are considering has applications in a variety of domains (e.g., controlling for batch effects in high-throughput sequencing data [22]), it is a substantial assumption that needs to be taken into account by practitioners. Furthermore, we restrict the latent variables to follow a Gaussian mixture model and the function mapping from latent to observed variables to be piecewise affine and injective. While this is a very flexible model class, how our causal effect identification result generalizes to the case where the ground truth model does not strictly belong to this class remains an open question.

8 Conclusion

We propose a method to address unmeasured discrete confounding in nonlinear cause-effect models. By mapping a confounded causal model to an equivalent latent variable model, we can leverage identifiability results in the literature on such models. We demonstrate that, under specific assumptions, it is possible to identify causal effects despite the presence of unmeasured confounders. We introduce a flow-based algorithm that can correct for this type of unmeasured confounding. The empirical results on both synthetic and real-world data provide evidence of the effectiveness of our approach.

As such, this work is an effort at building a bridge between the literature on causal inference that uses constraints on function classes and deep latent variable models. The usefulness of deep latent variable models have successfully been shown in a variety of applications and has spurned an interested in analyzing their identifiability properties, whose connections to causal inference problems we explore here.

Future work may investigate how the proposed strategy can be extended to more complex causal graphs, other model classes, and other estimable causal quantities such as counterfactuals.

References

[1] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018. [Cited on pages 1 and 9.]

[2] Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1):39–67, 2020. [Cited on pages 1 and 9.]

[3] Alexander D'Amour. On multi-cause causal inference with unobserved confounding: Counterexamples, impossibility, and alternatives. *arXiv preprint arXiv:1902.10286*, 2019. [Cited on page 4.]

[4] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014. [Cited on page 6.]

[5] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016. [Cited on page 6.]

[6] Spencer L Gordon, Bijan Mazaheri, Yuval Rabani, and Leonard Schulman. Causal inference despite limited global confounding via mixture models. In *Conference on Causal Learning and Reasoning*, pages 574–601. PMLR, 2023. [Cited on page 1.]

[7] Luigi Gresele, Julius Von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. Independent mechanism analysis, a new concept? *Advances in neural information processing systems*, 34:28233–28248, 2021. [Cited on page 9.]

[8] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009. [Cited on page 7.]

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. [Cited on page 7.]

[10] Patrik O Hoyer, Shohei Shimizu, Antti J Kerminen, and Markus Palviainen. Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362–378, 2008. [Cited on pages 1 and 9.]

[11] Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ica of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pages 460–469. PMLR, 2017. [Cited on page 9.]

[12] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015. [Cited on page 1.]

[13] Dominik Janzing, Jonas Peters, Joris Mooij, and Bernhard Schölkopf. Identifying confounders using additive noise models. *arXiv preprint arXiv:1205.2640*, 2012. [Cited on page 9.]

[14] Dominik Janzing and Bernhard Schölkopf. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010. [Cited on page 9.]

[15] Dominik Janzing and Bernhard Schölkopf. Detecting confounding in multivariate linear models via spectral analysis. *Journal of Causal Inference*, 6(1):20170013, 2018. [Cited on pages 1 and 9.]

[16] Dominik Janzing and Bernhard Schölkopf. Detecting non-causal artifacts in multivariate linear regression models. In *International Conference on Machine Learning*, pages 2245–2253. PMLR, 2018. [Cited on pages 1 and 9.]

[17] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: an unsupervised and generative approach to clustering. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 1965–1972, 2017. [Cited on page 5.]

[18] Nathan Kallus and Angela Zhou. Minimax-optimal policy learning under unobserved confounding. *Management Science*, 67(5):2870–2890, 2021. [Cited on page 9.]

[19] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020. [Cited on page 9.]

[20] Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. [Cited on page 2.]

[21] Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Identifiability of deep generative models without auxiliary information. *Advances in Neural Information Processing Systems*, 35:15687–15701, 2022. [Cited on pages 1, 2, 3, 4, 5, and 12.]

[22] Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, 2010. [Cited on page 9.]

[23] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017. [Cited on page 8.]

[24] Myrl G Marmarelis, Fred Morstatter, Aram Galstyan, and Greg Ver Steeg. Policy learning for localized interventions from observational data. In *International Conference on Artificial Intelligence and Statistics*, pages 4456–4464. PMLR, 2024. [Cited on page 9.]

[25] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021. [Cited on pages 2 and 4.]

[26] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, 30, 2017. [Cited on page 6.]

[27] Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009. [Cited on pages 1, 2, 4, and 14.]

[28] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017. [Cited on page 9.]

[29] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006. [Cited on page 9.]

[30] Vincent Stimper, Bernhard Schölkopf, and José Miguel Hernández-Lobato. Resampling base distributions of normalizing flows. In *International Conference on Artificial Intelligence and Statistics*, pages 4915–4936. PMLR, 2022. [Cited on page 5.]

[31] Tatsuya Tashiro, Shohei Shimizu, Aapo Hyvärinen, and Takashi Washio. Parcelingam: a causal ordering method robust against latent confounders. *Neural computation*, 26(1):57–83, 2014. [Cited on page 1.]

Appendices

A Proof of main text results

Theorem 3.4. *Under Assumptions 3.1, 3.2, and 3.3 the mixture components and the causal mechanism for the effect $(\boldsymbol{Z}_Y, \boldsymbol{f}_Y)$ in Eq. (3.1) is identifiable up to an invertible affine reparameterization of $\boldsymbol{Z}_Y$. More precisely, let $(\tilde{\boldsymbol{Z}}_Y, \tilde{\boldsymbol{f}}_Y)$ be the latent variable and mechanism obtained by fitting the model to the observation distribution $P(\boldsymbol{X}, \boldsymbol{Y})$, then we have, for some $(m \times m)$ invertible matrix S and some $(m \times 1)$ vector $\boldsymbol{b}$*

$$\boldsymbol{f}_Y(\boldsymbol{x}, \boldsymbol{z}_Y) = \tilde{\boldsymbol{f}}_Y(\boldsymbol{x}, S\boldsymbol{z}_Y + \boldsymbol{b}), \quad \textit{and} \quad \tilde{\boldsymbol{Z}}_Y = S\boldsymbol{Z}_Y + \boldsymbol{b}.$$

Proof. **Step 1**: *Affine identifiability.*

The above model can be rewritten as a piecewise affine injective mapping

$$\boldsymbol{\Psi}: \quad \mathcal{Z} \to \mathcal{X} \times \mathcal{Y}, \tag{A.1}$$

$$\begin{bmatrix} \boldsymbol{z}_X \\ \boldsymbol{z}_Y \end{bmatrix} \mapsto \begin{bmatrix} \boldsymbol{f}_X(\boldsymbol{z}_X) \\ \boldsymbol{f}_Y(\boldsymbol{f}_X(\boldsymbol{z}_X), \boldsymbol{z}_Y) \end{bmatrix}. \tag{A.2}$$

Therefore we get affine identifiability from [21, Theorem 3.2].

Step 2: *Form restriction on the affine transformation due to partial observation.*[8] Assume another solution $\tilde{f}$, it can also be rewritten as an injective mapping

$$\tilde{\boldsymbol{\Psi}}: \quad \mathcal{Z} \to \mathcal{X} \times \mathcal{Y}, \tag{A.3}$$

$$\begin{bmatrix} \boldsymbol{z}_X \\ \boldsymbol{z}_Y \end{bmatrix} \mapsto \begin{bmatrix} \tilde{\boldsymbol{f}}_X(\boldsymbol{z}_X) \\ \tilde{\boldsymbol{f}}_Y(\boldsymbol{f}_X(\boldsymbol{z}_X), \boldsymbol{z}_Y) \end{bmatrix}. \tag{A.4}$$

By affine identifiability, $\tilde{\boldsymbol{\Psi}}^{-1} \circ \boldsymbol{\Psi}$ is an affine map $\boldsymbol{z} \mapsto A\boldsymbol{z} + \boldsymbol{b}$. From the above we deduce that[9]

$$A = \begin{bmatrix} T & 0 \\ U & S \end{bmatrix}. \tag{A.5}$$

with U an $m \times n$ row vector, T an invertible matrix and S a non-vanishing scalar (due to invertibility of both functions).

Step 3: *Further form restriction due to non-degeneracy of intra-mixture component covariances.* Let us consider the ground truth distribution of $\boldsymbol{Z}$: due to Assumption. 3.3 it is a Gaussian mixture, whose mixture components are indexed by $\{(l, q)\}_{l=1..K_L; q=1..K_Q}$ and whose associated covariances are of block diagonal of the form

$$\Sigma_{l,q} = \begin{bmatrix} \Sigma_l^X & \mathbf{0} \\ \mathbf{0} & \Sigma_q^Y \end{bmatrix}.$$

Moreover, this is the same for the retrieved latent $\tilde{\boldsymbol{Z}}$, up a permutation of indices $(l, q) \mapsto \sigma(l, q)$ and the affine transformation introduced above (e.g. using Theorem C.2 in [21], stating that the mixture components are identified up to a permutation and affine transformation). As a consequence we get, for any index (l, q), that the corresponding mixture component covariance $\widetilde{\Sigma}_{\sigma(l,q)}$ correspond $\Sigma_{l,q}$ after linear transformation of the Gaussian distribution by matrix A, i.e.

$$\widetilde{\Sigma}_{\sigma(l,q)} = A\Sigma_{l,q}A^\top = \begin{bmatrix} T & 0 \\ U & S \end{bmatrix} \begin{bmatrix} \Sigma_l^X & 0 \\ 0 & \Sigma_q^Y \end{bmatrix} \begin{bmatrix} T^\top & U^\top \\ 0 & S^\top \end{bmatrix} \tag{A.6}$$

$$= \begin{bmatrix} T & 0 \\ U & S \end{bmatrix} \begin{bmatrix} \Sigma_l^X T^\top & \Sigma_l^X U^\top \\ 0 & \Sigma_q^Y S^\top \end{bmatrix} \tag{A.7}$$

$$= \begin{bmatrix} T\Sigma_l^X T^\top & T\Sigma_l^X U^\top \\ U\Sigma_l^X T^\top & S\Sigma_q^Y S^\top + U\Sigma_l^X U^\top \end{bmatrix}. \tag{A.8}$$

[8] Restriction on the ambiguity that results because we only recover $g, \mathcal{Z}$ up to affine transformation. The point here is that it is a very special ambiguity, namely one where $\boldsymbol{A}$ is diagonal.

[9] This is because $\boldsymbol{\Psi}$ is lower triangular, therefore $\tilde{\boldsymbol{\Psi}}$ is lower triangular, therefore $\tilde{\boldsymbol{\Psi}}^{-1}$ is lower triangular, and therefore $\tilde{\boldsymbol{\Psi}}^{-1} \circ \boldsymbol{\Psi}$ is lower triangular.

where the off diagonal blocks must again be equal to zero by Assumption 3.3 applied to the covariance of the mixture component of the obtained solution $\widetilde{\Sigma}_{\sigma(l,q)}$. Exploiting this assumption further, let us choose l such that Σ_l^X is positive definite. In that case, we can write for the off-diagonal block

$$U\Sigma_l^X T^\top = 0 \tag{A.9}$$

$$U\Sigma_l^X = 0 \text{ because } T^\top \text{ is invertible} \tag{A.10}$$

$$U = 0 \text{ because } \Sigma_l^X \text{ is positive definite and therefore invertible.} \tag{A.11}$$

Consequently,

$$A = \begin{bmatrix} T & 0 \\ 0 & S \end{bmatrix}, \tag{A.12}$$

which entails identifiability up to scalar affine reparametrization of Z_2 and affine invertible transformation of Z_1.

More precisely, for all $\boldsymbol{z}_1, \boldsymbol{z}_2$, the composition of $\tilde{\boldsymbol{\Psi}}^{-1}$ with $\boldsymbol{\Psi}$ is ambiguous up to a diagonal affine transformation:

$$\begin{bmatrix} \tilde{\boldsymbol{z}}_X \\ \tilde{\boldsymbol{z}}_Y \end{bmatrix} = \tilde{\boldsymbol{\Psi}}^{-1} \circ \boldsymbol{\Psi}(\boldsymbol{z}_X, \boldsymbol{z}_Y) = \begin{bmatrix} T\boldsymbol{z}_X + \boldsymbol{b}_1 \\ S\boldsymbol{z}_Y + \boldsymbol{b}_2 \end{bmatrix}$$

Leading to

$$\boldsymbol{\Psi}(\boldsymbol{z}_X, \boldsymbol{z}_Y) = \tilde{\boldsymbol{\Psi}}(T\boldsymbol{z}_X + \boldsymbol{b}_X, Sz_Y + \boldsymbol{b}_Y)$$

For the $\boldsymbol{X}$ component this gives

$$\boldsymbol{f}_X(\boldsymbol{z}_X) = \tilde{\boldsymbol{f}}_X(T\boldsymbol{z}_X + \boldsymbol{b}_X)$$

such that

$$\boldsymbol{f}_X^{-1}(\boldsymbol{x}) = T^{-1}\left(\tilde{\boldsymbol{f}}_X^{-1}(\boldsymbol{x}) - \boldsymbol{b}_X\right)$$

because $(f \circ g)^{-1} = g^{-1} \circ f^{-1}$. And for the $\boldsymbol{Y}$ component this gives,

$$\boldsymbol{f}_Y(\boldsymbol{f}_X(\boldsymbol{z}_X), \boldsymbol{z}_Y) = \tilde{\boldsymbol{f}}_Y(\tilde{\boldsymbol{f}}_X(T\boldsymbol{z}_X + \boldsymbol{b}_X), S\boldsymbol{z}_Y + \boldsymbol{b}_Y)$$

Finally we get the following relation for the causal mechanism

$$\boldsymbol{f}_Y(\boldsymbol{x}, \boldsymbol{z}_Y) = \tilde{\boldsymbol{f}}_Y(\boldsymbol{f}_X(\boldsymbol{z}_X), S\boldsymbol{z}_Y + \boldsymbol{b}_Y) = \tilde{\boldsymbol{f}}_Y(\boldsymbol{x}, S\boldsymbol{z}_Y + \boldsymbol{b}_Y)$$

□

Proposition 3.5. *Under the assumptions of Theorem 3.4, assume additionally strict positivity of $p(\boldsymbol{x}, \boldsymbol{z}_Y)$ for almost all $\boldsymbol{z}_Y$. Then, for any $\boldsymbol{x}$ in the support of $P(\boldsymbol{X})$, $\mathbb{E}\left[\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x})\right]$ is identifiable from the observation of $P(\boldsymbol{X}, \boldsymbol{Y})$ with adjustment formula*

$$\mathbb{E}\left[\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x})\right] = \mathbb{E}_{\boldsymbol{Z}_Y \sim P(\boldsymbol{Z}_Y)}\left[\tilde{\boldsymbol{f}}_Y(\boldsymbol{x}, S\boldsymbol{Z}_Y + \boldsymbol{b})\right] = \mathbb{E}_{\tilde{\boldsymbol{Z}}_Y \sim P(\tilde{\boldsymbol{Z}}_Y)}\left[\tilde{\boldsymbol{f}}_Y(\boldsymbol{x}, \tilde{\boldsymbol{Z}}_Y)\right], \tag{3.6}$$

where $P(\tilde{\boldsymbol{Z}}_Y)$ and $\tilde{\boldsymbol{f}}_Y$ is the solution identified in Theorem 3.4.

Proof. Consider a given $\boldsymbol{x}$ in the support of $p(\boldsymbol{X})$, the above backdoor adjustment require $p(\boldsymbol{y}|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{z}_Y)$ to be well defined for almost any $\boldsymbol{z}_Y$. Given our generative model of Section 5.1, this amounts to having $\boldsymbol{f}$ unambiguously defined for almost any $\boldsymbol{z}_Y$. As $\boldsymbol{f}_Y$ is only unambiguously identified on the support of the observational distribution $p(\boldsymbol{x}, \boldsymbol{z}_Y)$, it is necessary and sufficient to have strict positivity of $p(\boldsymbol{x}, \boldsymbol{z}_Y)$ for almost all $\boldsymbol{z}_Y$. The adjustment formula using $\boldsymbol{Z}_Y$ is given by

$$\mathbb{E}\left[\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x})\right] = \mathbb{E}_{\boldsymbol{Z}_2 \sim P(\boldsymbol{Z}_Y)}\left[\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{Z}_Y)\right]$$

Using Theorem 3.4 we can rewrite the expression of function $\boldsymbol{f}$ such that

$$\mathbb{E}\left[\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x})\right] = \mathbb{E}_{\boldsymbol{Z}_Y \sim P(\boldsymbol{Z}_Y)}\left[\tilde{\boldsymbol{f}}_Y(\boldsymbol{x}, S\boldsymbol{Z}_Y + \boldsymbol{b})\right].$$

Moreover, we can replace the (unknown) latent variable distribution $P(\boldsymbol{Z}_2)$ with the estimated latent variable distribution $P(\tilde{\boldsymbol{Z}}_2)$ to obtain the result

$$\mathbb{E}\left[\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x})\right] = \mathbb{E}_{\tilde{\boldsymbol{Z}}_Y \sim P(\tilde{\boldsymbol{Z}}_Y)}\left[\tilde{\boldsymbol{f}}_Y(\boldsymbol{x}, \tilde{\boldsymbol{Z}}_Y)\right]. \tag{A.13}$$

□

Proposition 3.6. *If there exists (l, q) such that $P(L = l, Q = q) > 0$ and both Σ_l^X and Σ_q^Y are positive definite, then the positivity assumption on $p(\boldsymbol{x}, \boldsymbol{z}_Y)$ in Proposition 3.5 is satisfied.*

Proof. As $p(\boldsymbol{x}, \boldsymbol{z}_Y)$ is the pushforward of $p(\boldsymbol{z}_X, \boldsymbol{z}_Y)$ by an invertible, continuous, differentiable almost everywhere, function $\boldsymbol{\Psi}$ defined in the proof of Theorem 3.4. Therefore, $p(\boldsymbol{x}, \boldsymbol{z}_Y)$ is strictly positive if and only if $p(\boldsymbol{z}_X = \boldsymbol{f}_X^{-1}(\boldsymbol{x}), \boldsymbol{z}_Y)$ is strictly positive. Since $p(\boldsymbol{z}_X, \boldsymbol{z}_Y)$ is a Gaussian mixture, it is sufficient to have at least one non-degenearate mixture component occurring with non-zero probability strict positivity (see Assumption 3.3).. □

B Structural causal models

Causal dependencies between variables can be described using *Structural Causal Models* (SCM) [27].

Definition B.1 (SCM)**.** An n-variable SCM is a triplet $\mathcal{M} = (\mathcal{G}, \mathbb{S}, P_{\boldsymbol{U}})$ consisting of:

- a directed acyclic graph $\mathcal{G}$ with n vertices,
- a set $\mathbb{S} = \{\boldsymbol{V}_j := \boldsymbol{f}_j(\mathbf{Pa}_j, \boldsymbol{Z}_j), j = 1, \ldots, n\}$ of structural equations, where $\mathbf{Pa}_j$ are the variables indexed by the set of parents of vertex j in $\mathcal{G}$,
- a joint distribution $P_{\boldsymbol{Z}}$ over the exogenous variables $\{\boldsymbol{Z}_j\}_{j \leq n}$.

Due to the directed acyclic structure of $\mathcal{G}$, for each value of the exogenous variables, $\mathbb{S}$ leads to a unique solution for the vector of so-called endogenous variables $\boldsymbol{V} = [\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n]^\top$, such that the distribution $P_{\boldsymbol{Z}}$ entails a well-defined joint distribution over the endogenous variables $P(\boldsymbol{V})$. For the purpose of the present work, we adopt a very general setting by: (1) not enforcing joint independence between the exogenous variables, allowing them to encode hidden confounding, (2) allowing endogenous and exogenous variable to be vector-valued. A given set of random variables, there may be described by different SCMs, e.g. by making different choices of grouping components in vector variables $\boldsymbol{V}_k$, or by choosing which will appear as exogenous or endogenous variables. We may switch between different such choices, provided those choices make a equivalent predictions regarding interventions that we introduce next.

We will consider do-interventions in SCMs involve replacing one or more structural equation by a constant and modifying $\mathcal{G}$ accordingly such that parents of the intervened equations are removed. An intervention transforms the original model $\mathcal{M} = (\mathcal{G}, \mathbb{S}, P_{\boldsymbol{Z}})$ into an intervened model $\mathcal{M}^{do(\boldsymbol{V}_k = \boldsymbol{v}_k)} = \left(\mathcal{G}^{do(\boldsymbol{V}_k = \boldsymbol{v}_k)}, \mathbb{S}^{do(\boldsymbol{V}_k = \boldsymbol{v}_k)}, P_{\boldsymbol{Z}}^{do(\boldsymbol{V}_k = \boldsymbol{v}_k)}\right)$, where $\boldsymbol{v}_k$ is the constant parameterizing the intervention.

B.1 Unmeasured confounding and backdoor criterion

In the standard setting of causal effect estimation, one focuses on a graph comprising a pair of endogenous variables $(\boldsymbol{X}, \boldsymbol{Y})$ such that $\mathcal{G}$ contains the edge $\boldsymbol{X} \to \boldsymbol{Y}$. Hidden counfounding can then be encoded by non-independence of the respective exogenous variables $\boldsymbol{Z}_{\boldsymbol{X}}$ and $\boldsymbol{Z}_{\boldsymbol{Y}}$ of these nodes, which we represent as a dashed bidirectional arrow in Figure 1a. Our framework amounts to constraining the structure of this hidden confounding, which is assumed to be representable as an hidden discrete common cause of two hidden latent variables $\boldsymbol{Z}_X$ and $\boldsymbol{Z}_Y$, as described by the causal diagram of Figure 1b, which does not have any dependence between exogenous variables of the nodes $\boldsymbol{X}$ and $\boldsymbol{Y}$, because confounding is now explicitly represented by a common cause H. The additional variables appearing in this new graph, if they were to be observed, could be used to estimate the interventional probability $P(\boldsymbol{Y}|\mathrm{do}(\boldsymbol{X} = \boldsymbol{x}))$ because they satisfied the so-called backdoor criterion [27]: they block all backdoor paths between $\boldsymbol{X}$ and $\boldsymbol{Y}$, i.e. those going through a parent of $\boldsymbol{X}$. Although latent variable are unobserved, additional assumption may allow to identify them from observational data. In particular, one way is to formulate the observations as a function of the latents, which can be done by introducing an invertible mapping $\boldsymbol{\phi} : \boldsymbol{Z}_X \to \boldsymbol{X}$, leading to the causal diagram of Figure 1c.

We will focus on a case where it can be shown that we can infer and use $\boldsymbol{Z}_Y$ as a backdoor adjustment variable, which leads to the following formula for the interventional distribution

$$P(\boldsymbol{Y}|\mathrm{do}(\boldsymbol{X}))] = \int P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z}_y)p(\boldsymbol{z}_y)d\boldsymbol{z}_y \,.$$

C Twins dataset

The remaining confouding variables are: 'risk factor, Lung', 'risk factor Hemoglobinopathy', 'risk factor, Incompetent cervix', 'mom place of birth', 'race of child', 'total number of births before twins', 'trimester prenatal care begun, 4 is none', 'number of live births before twins', 'married', 'risk factor, Anemia', 'risk factor, Hypertension, chronic', 'risk factor, RH sensitization', 'num of cigarettes /day, quantiled', 'risk factor, tobacco use', 'education category', 'state of occurence FIPB', 'medical person attending birth', 'quintile number of prenatal visits', 'US census region of mplbir', 'dad race', 'place of delivery', 'risk factor, Renal disease', 'mom race', 'risk factor, Cardiac', 'US census region of stoccfipb', 'risk factor, Previous infant 4000+ grams', 'US census region of brstate', 'birth month Jan-Dec', 'risk factor, Eclampsia', 'risk factor, Other Medical Risk Factors', 'octile age of father', 'risk factor, alcohol use', 'dad hispanic', 'num of drinks /week, quantiled', 'risk factor, Herpes', 'mom hispanic', 'risk factor, Hypertension, preqnancy-associated', 'state of residence NCHS', 'risk factor, Uterine bleeding', 'risk factor, Diabetes', 'sex of child', 'risk factor Hvdramnios/Oliqohvdramnios', 'risk factor, Previos pre-term or small', 'adequacy of care'.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, claims are accurate.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Section 7 with a separate 'Limitations' section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Proofs to all results are given in the Appendix A. In addition, a proof sketch for the main result is given in the main text.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 5 contains information about all the parameters used in the simulation results. Code to generate the synthetic data and to implement the method is provided in an anonymized zip file in the Supplementary Material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Yes, link to code will be provided on the first page conditional on acceptance. It is provided with the submission as a zip file in the Supplementary Material to guarantee anonymity.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: These details can be seen in the provided code.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: While no formal error bars are shown, we show results for a number of draws from the data generating process and report the distribution of results in the synthetic data experiments. In the real-world data application, we show results for a number of hyperparameter choices and seeds.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: That information is provided in the relevant Sections of the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics ?

   Answer: [Yes]

   Justification: Yes, the research conducted here conforms the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: While the paper does not discuss negative societal impacts, it emphasizes that the results for the proposed causal inference technique rest on assumptions that need to be fulfilled for the method to work as expected.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: No risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The use of code by other researchers is acknowledged.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Code with documentation is provided.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Not applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Not applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.