Controlling for discrete unmeasured confounding in nonlinear causal models

Patrick Burauel P.BURAUEL@GOOGLEMAIL.COM
California Institute of Technology Pasadena, CA 91125, USA

Frederick Eberhardt FDE@CALTECH.EDU
California Institute of Technology Pasadena, CA 91125, USA

Michel Besserve MICHEL.BESSERVE@TU-BRAUNSCHWEIG.DE
TU Braunschweig, 38106 Braunschweig, Germany
Max Planck Institute for Intelligent Systems, Tübingen, Germany

Editors: Biwei Huang and Mathias Drton

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 (Pearl, 2009), 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 (Imbens and Rubin, 2015; Chernozhukov et al., 2018). 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 (Imbens and Rubin, 2015), approaches based on parametric assumptions (such as in additive noise models (Tashiro et al., 2014; Hoyer et al., 2008), linear models (Janzing and Schölkopf, 2018a,b) or binary Gaussian mixture models (Gordon et al., 2023)), or settings where observed confounding is assumed to be representative of unobserved confounding (Cinelli and Hazlett, 2020).

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 of Kivva et al. (2022) that assure identification (up to an affine transformation) of the latent Gaussian mixtures under the assumption of

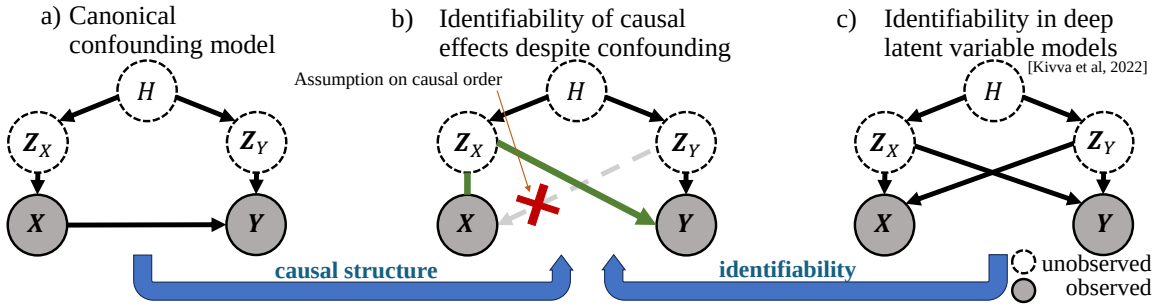

Figure 1: On the left, $\boldsymbol{X}$ causes $\boldsymbol{Y}$ and is confounded by H. On the right, observed variables $\boldsymbol{W} = (\boldsymbol{X}, \boldsymbol{Y})$ are generated by latent variables $\boldsymbol{Z}$, whose identifiability up to affine transformation under model restrictions is shown by Kivva et al. (2022). We combine knowledge of causal structure with identifiability results for latent variable models to estimate causal effects despite unmeasured confounding (middle).

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.

Notation. 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* (Pearl, 2009):

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

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 with 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) (Kingma et al., 2019; Papamakarios et al., 2021) 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 observational 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 Kivva et al. (2022), 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), Kivva et al. (2022) show identifiability of $\boldsymbol{\Psi}$ and $\boldsymbol{Z}$ up to an affine transformation (Kivva et al., 2022, Theorem 3.2). This model is depicted in Figure 1c.

3. Theoretical framework for discrete decounfounding

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 show 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 $\boldsymbol{Z}_X$ 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 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.*

1. We provide a brief description of the formalism of structural causal models in Appendix D.
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.

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

Assumption 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 Kivva et al. (2022). The final key to identifiability is a Gaussian mixture model of the exogenous variables and their confounding induced by H.

Assumption 3 *The exogenous variables are generated according to the following model:*

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

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

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

where $\boldsymbol{\pi}_{L|H}$ and $\boldsymbol{\pi}_{Q|H}$ are conditional probability mass functions. Moreover, we assume there exists at least one mixture component l occurring with non-zero probability has a positive definite Σ_l^X.

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 4 *Under Assumptions 1, 2, and 3 the mixture components and the causal mechanism for the effect $(\boldsymbol{Z}_Y, \boldsymbol{f}_Y)$ in Eq. (3.1) are 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 [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 $\boldsymbol{\Psi}$ 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 proof can be decomposed into three parts:

(1) The identifiability theory in (Kivva et al., 2022, 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 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 Kivva et al. (2022) *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 prevents 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 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 (Pearl, 2009), such that we can estimate the following interventional quantities by the adjustment formula:

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

That is, Theorem 4 provides the basis to deconfound the causal effect:

Proposition 5 *Under the assumptions of Theorem 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 4.

See Appendix A for the proof. Importantly, we cannot rely on $\boldsymbol{Z}_X$ 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 D'Amour (2019). Positivity of $p(\boldsymbol{x}, \boldsymbol{z}_Y)$ is achieved under mild assumptions: it only requires the occurrence of one non-degenerate mixture component of $\boldsymbol{Z}$ in the observational setting.

Proposition 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 5 is satisfied.*

See Appendix A for the proof. Overall, the positive definite assumptions required on covariance matrices in Theorem 4 and Proposition 6 emphasize the importance for identification of having, for at least one value of the confounder, mutually independent exogenous variations injected in both mechanisms $\boldsymbol{f}_X$ and $\boldsymbol{f}_Y$.

Identification of counterfactuals. In addition to interventional quantities, it is also possible to identify counterfactual quantities for $\boldsymbol{Y}$ in the same setting. More precisely, consider the setting where we observe "factual" values $(\boldsymbol{X}(\boldsymbol{z}), \boldsymbol{Y}(\boldsymbol{z}))$, where $\boldsymbol{z}$ is a particular value of the exogenous variable, corresponding to the characteristics of the so-called individual "unit" (e.g. a patient) (see (Pearl, 2009, Chapter 4). Then, as shown in Proposition 9 of Appendix B, we can estimate the counterfactual value $\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{z})$ of $\boldsymbol{Y}$, when intervention $\mathrm{do}(\boldsymbol{X}=\boldsymbol{x})$ is performed.

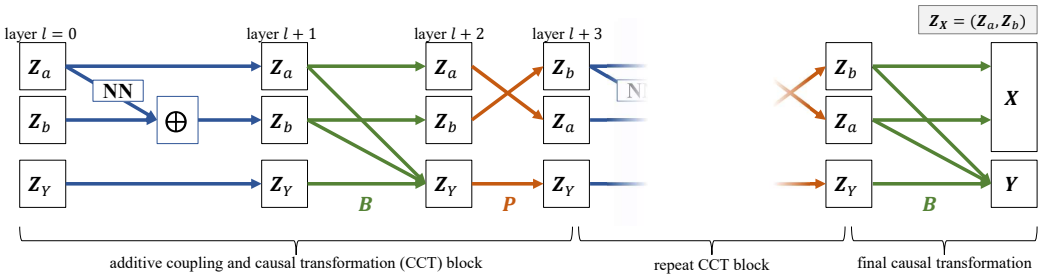

Figure 2: (Flow model implementation) The sequence of transformations that make up one block are composed of an additive coupling bijection from layer l to $l + 1$, see lines 5 and 6, a causal transformation with a block-triangular structure ($\boldsymbol{Z}_Y$ node does not influence other nodes), see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

4. Flow-based implementation

We use flow-based models (Papamakarios et al., 2021) to estimate the discrete confounding model.[4] Flow-based 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 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 (VAE), flow-based models allow optimizing the exact likelihood of the data, which seems to be critical for their use to estimate causal quantities precisely. Indeed, VAEs trained with a Gaussian mixture prior (Jiang et al., 2017), as used in the experimental section of Kivva et al. (2022), have proven not to perform as well as flow-based models for the use of deconfounding.[5]

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 transformation: $\boldsymbol{z}_x \to \boldsymbol{z}_y$)
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 flow-based models, observed variables $\boldsymbol{w} := (\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^{m+n}$ are expressed as a transformation T of $\boldsymbol{z}$, $\boldsymbol{w} = T(\boldsymbol{z})$, where $\boldsymbol{z}$ follows 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

4. Code is available at https://github.com/pburauel/DeconFlow/.
5. We have implemented VAEs with appropriate architectural restrictions in experiments (not reported here) that did not recover the true causal effects well even in the simple $m = n = 1$ linear case.

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.[6] 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 $\mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_k, \Sigma_k)$ with diagonal covariance matrix denotes the Gaussian distribution for component k. All $\boldsymbol{\mu}_k$, Σ_k, π_k as well as the parameters of the transformation T are optimized.

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. The subsequent section introduces a more expressive multi-layered model.

4.1. One-layer linear flow

In the simplest proof-of-concept model, we assume we observe 2D Gaussian mixtures in $\boldsymbol{w}$ generated by linear transformations of the latent variables. In order to satisfy the constraints of our causal model, transformation T is then a block lower triangular matrix,

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

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 spell out the relation between $\boldsymbol{z}$ and $\boldsymbol{w}$ in detail to draw attention to how the causal restriction is implemented through a lower triangular matrix as in Eq. (4.3),

$$\boldsymbol{w} = \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} = \boldsymbol{A}\boldsymbol{z}\,, \tag{4.4}$$

$$= \begin{bmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{z}_x \\ \boldsymbol{z}_y \end{bmatrix}. \tag{4.5}$$

Note that $\boldsymbol{x}$ is only influenced by $\boldsymbol{z}_x$ while $\boldsymbol{y}$ is influenced by both $\boldsymbol{z}_x$ and $\boldsymbol{z}_y$, which reflects the causal structure of our model. We apply this simple model to simulated data with a one-dimensional cause and discuss the results below.

6. A GMM base distibution in flow-based models has previously been used by e.g. Stimper et al. (2022).

4.2. Additive coupling bijection

To model more complex distributions of $\boldsymbol{w}$, we propose a flow-based model applying successive invertible transformation blocks to the latent variables. One transformation block is composed of an additive coupling layer (Dinh et al., 2014) and a "causal transformation" akin to a masked autoregressive layer (Papamakarios et al., 2017). 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,

$$\boldsymbol{B} = \begin{bmatrix} \text{diag}(\boldsymbol{a_{1,1}}) & \boldsymbol{0} \\ \boldsymbol{a_{2,1}} & a_{d,d} \end{bmatrix}$$

with $\boldsymbol{a_{1,1}} = \begin{bmatrix} a_{1,1} & \cdots & a_{d-1,d-1} \end{bmatrix}$ and $\boldsymbol{a_{2,1}} = \begin{bmatrix} a_{d,1} & \cdots & a_{d,d-1} \end{bmatrix}$, 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 by Dinh et al. (2016)) 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.

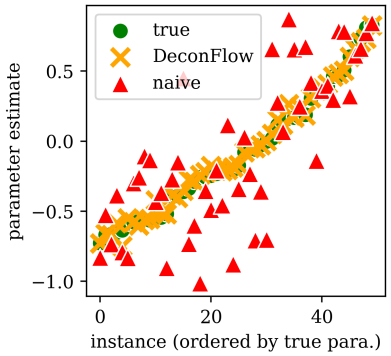

Figure 3: With one-dimensional cause and effect ($m = n = 1$), performance can be evaluated by comparing the `DeconFlow`-adjusted slope parameter estimates (orange crosses) to the ground truth (green circles). Red triangles are naive estimates, obtained without addressing confounding.

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

$$\log p_{\boldsymbol{w}}(\boldsymbol{w}) = \log p_{\boldsymbol{z}}(\boldsymbol{z}^{(0)}) + \sum_{l=1}^{L} \sum_{i=1}^{d} \log |a_{ii}^{(l)}| \tag{4.6}$$

where $\boldsymbol{z}^{(0)} = \overline{T}^{-1}\boldsymbol{w}$ with $\overline{T} = T_{(l=0)} \circ \ldots \circ T_{(l=L)}$ denoting the composition of the transformations performed by each layer, as described above (similarly for its inverse, $\overline{T}^{-1}$) and $p_{\boldsymbol{z}}$ being the density of a Gaussian mixture model with diagonal covariances, as in Eq. (4.2). The transformation in line 6, as well as the partial permutation of lines 8–9, have a unit Jacobian determinant. Therefore, they do not appear in the above log likelihood. We then optimize the log-likelihood in Eq. (4.6) using backpropagation.

4.3. Estimation of interventional quantites

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 $\mathbb{E}[Y|\text{do}(\boldsymbol{X} = \boldsymbol{x})]$ 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

$$\overline{\boldsymbol{w}} = (\boldsymbol{x}, \overline{\boldsymbol{y}})^\top = \frac{1}{N_p} \sum_{\tilde{\boldsymbol{z}}_y \sim P(\tilde{\boldsymbol{Z}}_Y)}^{N_p} \overline{T}(\tilde{\boldsymbol{z}}_{\boldsymbol{x}}, \tilde{\boldsymbol{z}}_y)\,, \tag{4.7}$$

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

$$\mathbb{E}[Y|\mathrm{do}(\boldsymbol{X} = \boldsymbol{x})] \approx \overline{\boldsymbol{y}} =: \hat{\theta}(\boldsymbol{x}). \tag{4.8}$$

5. Simulation Study

5.1. Data Generation

Given the generative model, we simulate data from a Generalized Additive Model (GAM, Hastie et al. (2009)) as follows, focusing on the case $m = 1$, a scalar effect. 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 mixture parameters for $\boldsymbol{Z}_X|\{L = l\} \sim \mathcal{N}(\boldsymbol{\mu}_l^X, \Sigma_l^X)$ and $Z_Y|\{Q = q\} \sim \mathcal{N}\left(\mu_q^Y, \left(\sigma_q^Y\right)^2\right)$ where $\boldsymbol{\mu}_l^X \sim \mathcal{U}(1, 4)$ and $\mu_q^Y \sim \mathcal{U}(0, 1)$, $\Sigma_l^X = \boldsymbol{I} \times 0.01$ and $\left(\sigma_q^Y\right)^2 = 0.01$.

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 τ_1, τ_2, τ_3:

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

where β is the true causal effect, which is drawn from $\mathcal{U}[-1, 1]$, and $\varepsilon \sim \mathcal{N}(0, 0.01)$. Inspired by He et al. (2016), 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 does not cover 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.

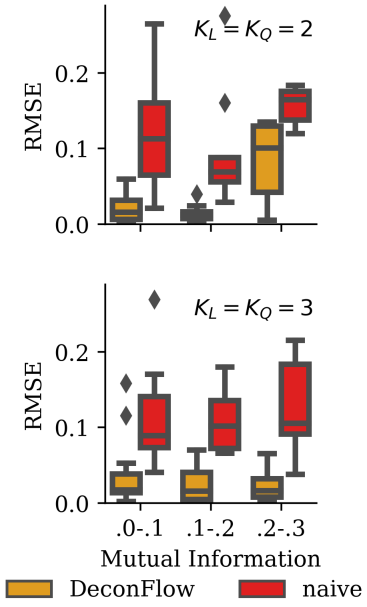

Figure 4: `DeconFlow` controls unmeasured confounding, see Section 5.2 for details.

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, in particular we use the (known since simulated)

ground truth τ_3 and Z_Y to average out the confounding effect. We compare $\theta^*(\boldsymbol{x})$ with the estimate defined in Eq. (4.8):

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

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

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

where for $\mathbb{E}(Y|\boldsymbol{x})$ we plug in the observed y associated with $\boldsymbol{x}$.

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 $\mathbb{E}[Y|\text{do}(X = x)]$, which we regress on x to compute the deconfounded parameter estimates that almost perfectly match the ground truth.[7]

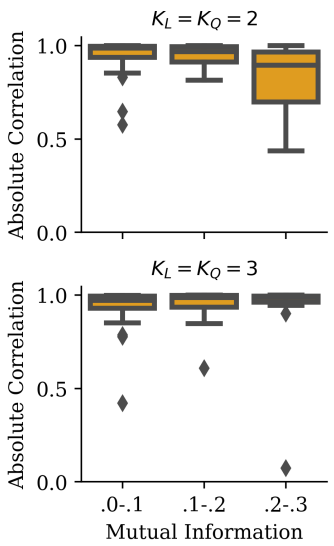

Figure 5: Recovered unmeasured confounder correlates with true confounder, see Section 5.2, part 2 for details.

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 $\text{RMSE}_{\text{naive}}$, 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` substantially decreases the error incurred when estimating $\mathbb{E}[Y|\text{do}(\boldsymbol{X} = \boldsymbol{x})]$ without observing the discrete confounder. What we achieve here is the estimation of a nonlinear causal quantity, $\mathbb{E}[Y|\text{do}(\boldsymbol{X} = \boldsymbol{x})]$, without observing the latent quantity that induces the discrepancy between it and $\mathbb{E}[Y|\boldsymbol{x}]$.[8]

To corroborate these empirical results, we compute the (absolute) correlation between the ground truth $\boldsymbol{Z}_Y$ (which `DeconFlow` never has access to) and its recovered version $\tilde{\boldsymbol{Z}}_Y$ in Figure 5. Note

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

that $\boldsymbol{Z}_Y$ is the critical variable whose recovered distribution we sample from to deconfound. The correlation between recovered and true $\boldsymbol{Z}_Y$ is large and almost always very close to 1, which lends support to the efficacy of our method. The instances depicted in Figure 5 correspond to those depicted in Figure 4.

In Appendix C, we provide additional simulation results that show the good performance of the method for smaller sample sizes (Figure 8) and higher dimensions (Figure 10), as well as robustness of the method to misspecified number of clusters (Figure 9).

In Appendix B we provide simulation results for the estimation of counterfactual quantities, based on Proposition 9. Specifically Figure 7 in Appendix B shows a much better estimation of individual effects using `DeconFlow` than when using a naive approach that just uses the value of Y matching $\boldsymbol{x}$ in the empirical distribution of the observations.

6. Application

We use data on twin births in the USA collected around 1990, which has been used before by Louizos et al. (2017) 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 E 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 that we can toggle between being observed or not. We treat the only three ordinal variables in the dataset as causes of interest, since we can approximate them with continuous variables by adding uniformly distributed noise. In this way we satisfy our model requirements of continuous cause variables and discrete confounding variables.

From the set of covariates $\{X_1, \dots, X_K\}$ we select the three ordinal variables that are directly related to the mother as observed causes:[9] *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 $\boldsymbol{V} = \{X_4, \dots, X_K\}$. 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 $\tilde{\boldsymbol{\beta}}$.

We run `DeconFlow` for multiple seeds and hyperparameters. In Figure 6, 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

9. We do not use 'dtotord_min' (total number of births before twins) and 'dlivord_min' (number of live births before twins), technically also ordinal variables, as causes because these variables have a massively skewed empirical distribution.

observe that our method yields mean parameter estimates that are closer to β^* than $\hat{\beta}$. For *gestation type*, we find $\tilde{\beta}$ to be lower than both β^* and $\hat{\beta}$.

While we consider similar β^* and $\tilde{\beta}$ as evidence that our method accounts for $\boldsymbol{V}$ without observing it, we stress that β^* 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 β^* and $\tilde{\beta}$ might indicate the existence of additional confounders unmeasured in the dataset, rather than a shortcoming of our method. For instance, the discrepancy between $\tilde{\beta}$ and β^* for *gestation type* could be due to additional unmeasured confounders.

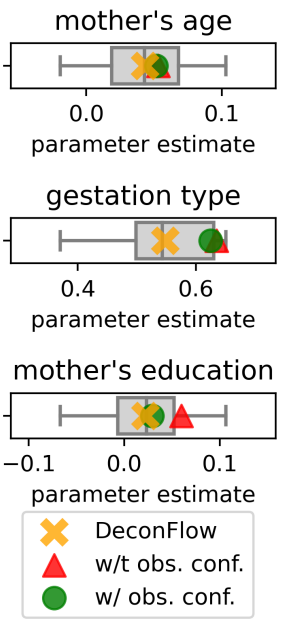

Figure 6: Estimated confounder adjustment in empirical application, see Section 6 for details.

7. Discussion

While there is a large literature on using measured confounders to deconfound causal effect estimates (see e.g. Chernozhukov et al. (2018)), or to gauge the sensitivity to unmeasured confounders by benchmarking against *measured* confounders in treatment effect estimation (Cinelli and Hazlett, 2020) or policy learning (Kallus and Zhou, 2021; Marmarelis et al., 2024), 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) (Peters et al., 2017; Janzing and Schölkopf, 2010). For instance, Janzing and Schölkopf (2018a,b) formalize ICM in multivariate linear models to estimate a degree of confounding. ICM can also be seen as motivating additive noise models as used by Janzing et al. (2012), which is similar to our approach in the sense that a latent confounder is learned from observed variables. However, that method did not allow for both a causal *and* a confounding effect between the two variables.

Even without implicit or explicit motivation through ICM, restricting 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 (Shimizu et al., 2006). In this model class, Hoyer et al. (2008) 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. Khemakhem et al. (2020) and Hyvarinen and Morioka (2017) but these require observed auxiliary information (such as environment variables) or assumptions like ICM (Gresele et al., 2021). None of these methods are specifically designed to address unmeasured confounding in a principled 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 (Leek et al., 2010)), 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. In addition, one practical limitation to the methodology is the fact that GMM parameters are increasingly challenging to estimate as the variance of the mixture components increases relative to the squared distance between components' means. The method is thus likely to work best when the clusters formed by the discrete confounding variable have limited overlap.

8. Conclusion

We propose a method to address unmeasured discrete confounding in (non-)linear 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 prove that under a specific set of 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. Given the success of deep latent variable models in a variety of applications, there has been much interest in understanding their identifiability properties. Our results contribute to this effort by building a bridge to techniques of handling unmeasured confounding in causal inference.

References

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.

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.

Alexander D'Amour. On multi-cause causal inference with unobserved confounding: Counterexamples, impossibility, and alternatives. *arXiv preprint arXiv:1902.10286*, 2019.

Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

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.

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.

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.

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.

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.

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

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

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

Dominik Janzing and Bernhard Schölkopf. Detecting confounding in multivariate linear models via spectral analysis. *Journal of Causal Inference*, 6(1):20170013, 2018a.

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

Dominik Janzing, Jonas Peters, Joris Mooij, and Bernhard Schölkopf. Identifying confounders using additive noise models. *arXiv preprint arXiv:1205.2640*, 2012.

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.

Nathan Kallus and Angela Zhou. Minimax-optimal policy learning under unobserved confounding. *Management Science*, 67(5):2870–2890, 2021.

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.

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

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.

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.

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.

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.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, 30, 2017.

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.

Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

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.

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.

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.

Appendices

Appendix A. Proof of main text results

Theorem 4 *Under Assumptions 1, 2, and 3 the mixture components and the causal mechanism for the effect* $(\boldsymbol{Z}_Y, \boldsymbol{f}_Y)$ *in Eq. (3.1) are 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 We follow the steps of the sketch from main test in more detail.

Step 1a: *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 (Kivva et al., 2022, Theorem 3.2).

Step 1b: *Form restriction on the affine transformation due to causal structure.*[10] 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 components' dependency structure of these maps, we deduce that[11]

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

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

Step 2: *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 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

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

10. We only recover $\boldsymbol{\Psi}$ up to an ambiguity, namely up to an affine transformation. The point here is that it is a special ambiguity: one where the linear component of the transformation is block-lower triangular.
11. This is because $\boldsymbol{\Psi}$ and $\tilde{\boldsymbol{\Psi}}$ are block-lower triangular by assumption (in the sense that the matrix indicating dependency between input and output variables is block-lower triangular, which also results in a lower triangular Jacobian matrices wherever defined), therefore $\tilde{\boldsymbol{\Psi}}^{-1}$ is also block-lower triangular, and therefore $\tilde{\boldsymbol{\Psi}}^{-1} \circ \boldsymbol{\Psi}$ is block-lower triangular.

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 Kivva et al. (2022), stating that the mixture components are identified up to a permutation and affine transformation). As a consequence, for any index (l, q), the corresponding mixture component covariance matrix $\widetilde{\Sigma}_{\sigma(l,q)}$ corresponds to $\Sigma_{l,q}$ up to linear transformation via the blocks of matrix A, i.e.

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

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

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

where the off diagonal blocks must again be equal to zero by Assumption 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.

Step 3: *Detailed ambiguity relation.* 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 7 *Under the assumptions of Theorem 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 4.

Proof Consider a given $\boldsymbol{x}$ in the support of $p(\boldsymbol{X})$, the above backdoor adjustment requires $p(\boldsymbol{y}|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{z}_Y)$ 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 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 8 *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 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 4, it follows that $p(\boldsymbol{x}, \boldsymbol{z}_Y)$ 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). ■

Appendix B. Identification of counterfactual quantities

Proposition 9 *Under the assumptions of Theorem 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 $P(\boldsymbol{X})$, the counterfactual value $\boldsymbol{Y}_{\boldsymbol{z}}(\boldsymbol{z})$ for an individual unit with exogenous values $\boldsymbol{z}$ is identifiable from the observation of $P(\boldsymbol{X}, \boldsymbol{Y})$ and the factual values $(\boldsymbol{X}(\boldsymbol{z}), \boldsymbol{Y}(\boldsymbol{z}))$ with adjustment formula*

$$\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{z}) = \tilde{\boldsymbol{f}}_Y(\boldsymbol{x}, S\boldsymbol{z}_Y + \boldsymbol{b}_2) = \tilde{\boldsymbol{f}}_Y(\boldsymbol{x}, \tilde{\boldsymbol{z}}_Y)\,, \tag{B.1}$$

where $(\tilde{\boldsymbol{z}}_X, \tilde{\boldsymbol{z}}_Y) = \tilde{\boldsymbol{\Psi}}^{-1}(\boldsymbol{X}(\boldsymbol{z}), \boldsymbol{Y}(\boldsymbol{z}))$ are the latents corresponding to the factual observation under the identified latent variable model in Theorem 4, and $\tilde{\boldsymbol{f}}_Y$ is the corresponding component function of this model.

Proof Given the assumptions, the results of Theorem 4 hold and we can use the following intermediate result of its proof to establish the link between the realizations of ground truth latent variables $\boldsymbol{z} = (\boldsymbol{z}_X, \boldsymbol{z}_Y)$ and realization of latent variables of the identified model, which we apply to the realization of the variables corresponding to factual observations:

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

where $\tilde{\boldsymbol{\Psi}}$ and $\boldsymbol{\Psi}$ follow the definition of Eq. (A.4) and Eq. (A.2), respectively.

Given $\boldsymbol{\Psi}(\boldsymbol{z}_X, \boldsymbol{z}_Y)$ correspond to the factual observations, this leads to

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

Additionally, given our SCM, the counterfactual for this unit writes

$$\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{z}) = \boldsymbol{f}_Y(\boldsymbol{x}, \boldsymbol{z}_Y)$$

Using the result of Theorem 4 this leads to

$$\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{z}) = \tilde{\boldsymbol{f}}_Y(\boldsymbol{x}, S\boldsymbol{z}_Y + \boldsymbol{b}_2)$$

On replacing by the above expression for the latent we get

$$\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{z}) = \tilde{\boldsymbol{f}}_Y(\boldsymbol{x}, \tilde{\boldsymbol{z}}_Y)$$

■

Appendix C. Additional simulation results

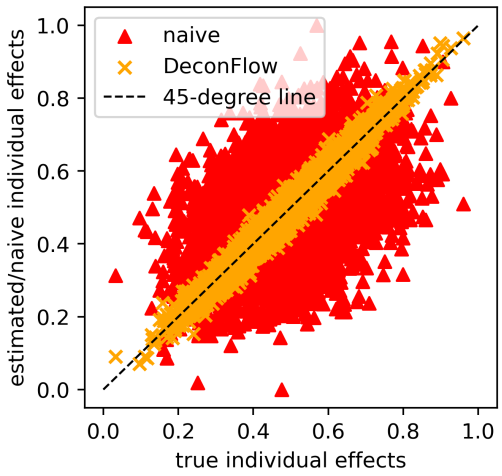

Figure 7: **Estimation of individual treatment effects in synthetic data:** We show the empirical performance of `DeconFlow` for estimating counterfactual individual effects. For given factual values $(\boldsymbol{X}(\boldsymbol{z}), \boldsymbol{Y}(\boldsymbol{z}))$ for a unit with associated exogenous value $\boldsymbol{z}$, we choose a random counterfactual $\check{\boldsymbol{x}}$ from the marginal distribution of $\boldsymbol{X}(\boldsymbol{z})$. As naive estimate for the counterfactual, we use the prediction of the estimated model for $\check{\boldsymbol{x}}$ without any adjustment for confounding. We plot the naive and estimated individual counterfactual effects against the true individual effects. $m = 4$, $n = 1$, $K_L = K_Q = 2$, no of layers of `DeconFlow` $= 50$.

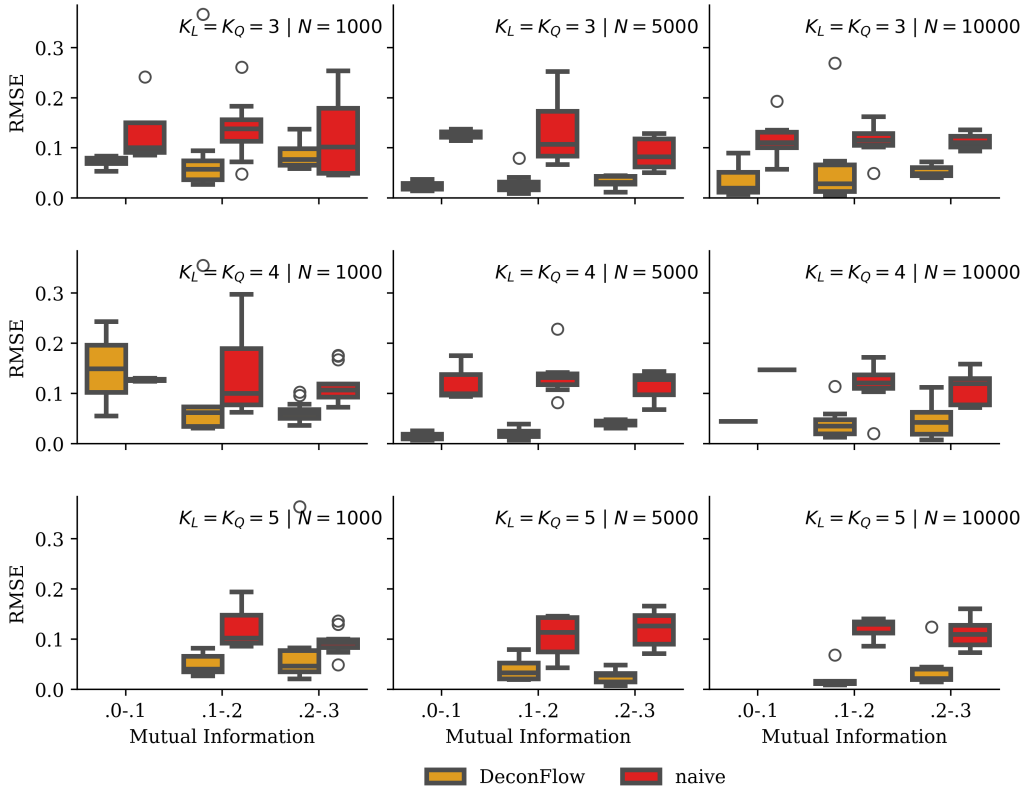

Figure 8: **Finite sample performance:** This figure shows the performance of our algorithm for smaller sample sizes than used in the experimental results discussed in the main paper. The sample size increases from left ($N = 1,000$) to right ($N = 10,000$, as in the main paper). The performance deteriorates slightly when the number of samples is $1,000$, but the method generally performs well also with lower sample sizes. Note that the column '.0-.1' is empty in the bottom row because the degree of mutual information between discrete variable L and Q (which is what is plotted here on the x-axis) is chosen at random and happens to lie above 0.1 in these instances. $m = 5$, $n = 1$.

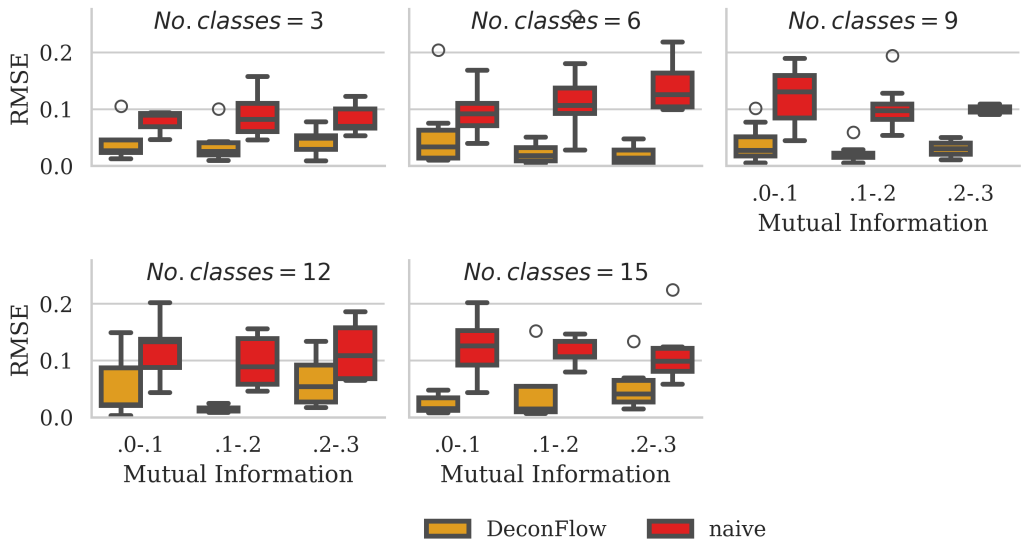

Figure 9: **Robustness to misspecified number of clusters:** Results for experimental setup with true $K_L = K_Q = 3$, implying a ground truth number of classes equal to 9. One can see that the performance of the algorithm does not degrade substantially when the number of components in the Gaussian mixture prior in the latent space (indicated by individual figure titles) is misspecified. $m = 5, n = 1$.

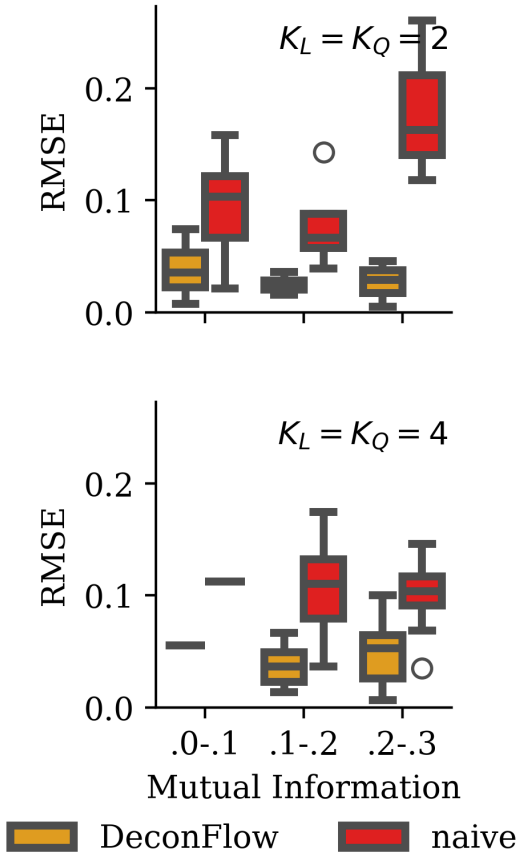

Figure 10: **Robust performance in high-dimensional space:** We run the algorithm in a higher-dimensional setting, specifically for $m = 10$, $n = 1$, all other parameters as in Figure 4 in the main text. Performance does not degrade relative to lower-dimensional problem.

Appendix D. Structural causal models

Causal dependencies between variables can be described using *Structural Causal Models* (SCM) (Pearl, 2009).

Definition 10 (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 $\mathbb{S} = \{\boldsymbol{V}_j \coloneqq \boldsymbol{f}_j(\boldsymbol{Pa}_j, \boldsymbol{Z}_j), j = 1, \dots, n\}$ of structural equations, where $\boldsymbol{Pa}_j$ 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 $\boldsymbol{V} = [\boldsymbol{V}_1, \dots, \boldsymbol{V}_n]^\top$, 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 variables to be vector-valued.

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.

D.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 by the unobserved common cause H 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 $\boldsymbol{Z}_Y$. If these variables were to be observed, they could be used to estimate the interventional probability $P(\boldsymbol{Y}|\mathrm{do}(\boldsymbol{X} = \boldsymbol{x}))$ because they satisfy the so-called backdoor criterion (Pearl, 2009): they block all backdoor paths between $\boldsymbol{X}$ and $\boldsymbol{Y}$, i.e. those going through a parent of $\boldsymbol{X}$. Although the latent variables are unobserved, additional assumptions permit their identification 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 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

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

Appendix E. Twins dataset

cause variables (ordinal): 'mother's age', 'gestation type', and 'mother's education'
effect variable of interest (continuous): 'birth weight of the first-born twin'
remaining confounding variables (discrete): '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'.