Research Article Open Access

Maximilian Ilse*, Patrick Forré, Max Welling, and Joris M. Mooij

Combining Interventional and Observational Data Using Causal Reductions

DOI: DOI, Received ..

Abstract: Unobserved confounding is one of the main challenges when estimating causal effects. We propose a causal reduction method that, given a causal model, replaces an arbitrary number of possibly high-dimensional latent confounders with a single latent confounder that takes values in the same space as the treatment variable, without changing the observational and interventional distributions the causal model entails. This allows us to estimate the causal effect in a principled way from combined data without relying on the common but often unrealistic assumption that all confounders have been observed. We apply our causal reduction in three different settings. In the first setting, we assume the treatment and outcome to be discrete. The causal reduction then implies bounds between the observational and interventional distributions that can be exploited for estimation purposes. In certain cases with highly unbalanced observational samples, the accuracy of the causal effect estimate can be improved by incorporating observational data. Second, for continuous variables and assuming a linear-Gaussian model, we derive equality constraints for the parameters of the observational and interventional distributions. Third, for the general continuous setting (possibly nonlinear or non-Gaussian), we parameterize the reduced causal model using normalizing flows, a flexible class of easily invertible nonlinear transformations. We perform a series of experiments on synthetic data and find that in several cases the number of interventional samples can be reduced when adding observational training samples without sacrificing accuracy.

Keywords: causal inference, causal effect estimation, dose-response curve, latent confounding, combining observational and interventional data

MSC: 62???

1 Introduction

In this work, we propose a novel, principled approach for causal effect estimation that can efficiently combine observational and interventional samples, even in the presence of unobserved confounding in the observational regime. We show that this method can potentially reduce the required Randomized Controlled Trial (RCT) sample size when sufficient observational samples are available (e.g., in the form of electronic health records). Recent real-world examples that could benefit from such an approach are the COVID-19 vaccine trials. Several of the vaccines require multiple dosages. However, due to many factors that are beyond our control, the times between administering the dosages differ. A question that then arises is: What is the effect of the time between, say, the second and the third dosage on the vaccine efficacy? In the absence of any large randomized controlled trials that provide a definite answer to this question, one may hope to estimate this by combining the few available clinical trial data with large quantities of observational data

***Corresponding Author: Maximilian Ilse:** Informatics Institute, University of Amsterdam; E-mail: ilse.maximilian@gmail.com
Patrick Forré: Informatics Institute, University of Amsterdam; E-mail: P.D.Forre@uva.nl
Max Welling: Informatics Institute, University of Amsterdam; E-mail: welling.max@gmail.com
Joris M. Mooij: Korteweg-De Vries Institute of Mathematics, University of Amsterdam; E-mail: j.m.mooij@uva.nl

collected as a part of the different vaccination campaigns performed worldwide. The method we propose here provides a principled approach for such causal inference problems.

A complication when estimating causal effects is the potential presence of observed and—in particular—unobserved *confounders* (common causes of the cause and the effect). Our key technical contribution, which we believe to be a valuable tool on its own, is a construction that typically *reduces* the size of the latent confounder space in a causal model (for example, a causal Bayesian network with latent confounders, a structural causal model, or a potential outcome model).

This *causal reduction* operation shows that without loss of generality, one only needs to model a single latent confounding variable that takes on values in the same space as the treatment variable, even if, in reality, there could be many latent confounders and their joint value space could be much larger. This "reduced confounder" suffices to model exactly the observational distribution of treatment and outcome, as well as the distribution of outcome under any perfect intervention on treatment.

The causal reduction facilitates a parsimonious joint parameterization of the observational and interventional distributions. We apply causal reductions to causal inference from combined observational and interventional data in three different settings: (i) discrete treatment and outcome, (ii) continuous treatment and outcome assuming a linear-Gaussian model, (iii) continuous treatment and outcome assuming a possibly nonlinear, non-Gaussian model.

For discrete treatment and outcome variables, we show in Section 4 that the causal reduction trivially implies bounds between the observational and interventional distributions. For binary variables, these bounds were already known [MN98], and they can also be seen as special cases of the well-known instrumental variable bounds [Pea95, BP97] under perfect compliance. We point out that these bounds are exploited straightforwardly by the maximum likelihood estimator from combined observational and interventional data. In particular, we analyze the special case where the observational samples are heavily disbalanced: a large majority was either treated, or untreated. In such settings, if sufficient observational data is available, one can sometimes substantially reduce the number of interventional samples required to accurately estimate the Average Treatment Effect (ATE) or the Conditional Average Treatment Effect (CATE).

Next, we apply causal reductions to the case of a continuous treatment variable. For the linear-Gaussian case (that is, where all interactions are linear, and all distributions are Gaussian), we prove in Section 5 that our reduced parameterization implies that the observational and interventional distributions are not independent but are related by equality constraints on their parameters. We do not further work out the details of what this entails for estimation here, but instead address the more realistic general (nonlinear, non-Gaussian) setting. In Section 6, we parameterize the reduced causal model using a flexible class of easily invertible nonlinear transformations, so-called normalizing flows [TT13, RM15]. In combination with our reduction technique, normalizing flows enable the use of a simple maximum-likelihood approach to estimate the reduced model parameters. In this way, one can combine observational and interventional training data, while allowing for latent confounding in the observational regime, to obtain estimates of dose-response curves.

We perform a series of experiments on data simulated using nonlinear causal mechanisms. We find that in several cases we can substantially reduce the number of interventional samples required to achieve a certain accuracy when adding sufficient observational training samples. We observe that parameter sharing allows one to learn a more accurate model from a combination of data than from each subset individually. It is not understood at present whether this is due to constraints between the observational and interventional distributions similar to those that arise in the other settings, or due to the likelihood term involving the observational data acting as a regularizer for the interventional data. Nonetheless, we consider the empirical results as encouraging.

In summary, our four main contributions are: (i) A causal reduction method that replaces arbitrary latent confounders with a single latent confounder that takes values in the same space as the treatment variable, without changing the observational and interventional distributions entailed by the causal model; (ii) An analysis of the inequality constraints between the observational and interventional distributions for the discrete case, and a scenario in which these constraints can be successfully exploited for causal effect estimation from combined data; (iii) A derivation of equality constraints between interventional and observational distributions entailed by linear Gaussian causal models; (iv) A flexible parameterization of the general reduced model with continuous treatment and outcome using normalizing flows, enabling joint estimation of the observational and interventional distributions from observational and interventional data without making strong parametric assumptions.

2 Related work

Prior work on causal inference from multiple datasets can be roughly divided into addressing three different tasks: (i) identifying causal effects when the causal structure is known (see, e.g., [LCB20] and references therein), (ii) discovering/learning the causal structure (see, e.g., [MMC20] and references therein), and (iii) estimating causal effects. The present work addresses the latter task.

Randomized Controlled Trials (RCTs) are the gold standard for estimation of a causal effect [Fis21]. The goal of an RCT is to remove all confounding biases by randomization. If we have data from a perfect RCT with no non-compliance, we can directly compute the causal effect. However, in practice it is often costly, risky, unethical, or simply impossible to perform an RCT.

Therefore, a plethora of work on estimating causal effects solely from observational data exists. The vast majority of proposed methods assumes that a set of observed variables can be used to adjust for all confounding factors [HI05, CMC^{+}20]. Unfortunately, one can typically not test this assumption, and the reliability of the conclusions of such observational studies is debated [MSB^{+}14].

Researchers recently started combining those different modalities to address the limitations of causal effect estimation from interventional or observational data alone. Most prior work on this topic still relies on the assumption that all confounders are observed in the observational regime (e.g., [Sil16, ROBB18]). We only found three papers that allow for latent confounding in the observational regime. The method by [RBOB20] attempts to reduce confounding bias rather than completely removing it. For binary treatments, [KPS18] rely on an additional assumption that the hidden confounder has a certain parametric structure that can be modelled effectively (which may also reduce confounding bias). In contrast, [ACI20] depend on observed short-term and long-term outcome variables. In contrast, our approach allows to correct for confounding bias without making such additional assumptions.

Another approach that sidesteps the strong untestable assumption of no unobserved confounding is to *bound* the causal effect in terms of properties of observational data [BP97, MN98, Pea95]. Recently, [WSF19] introduced a technique called inflation that can be used to derive tighter bounds. While these bounds are typically valid in the presence of arbitrary unobserved confounding, we know of no work so far that exploits such bounds to obtain better estimates of the causal effect from combined data.

Furthermore, methods that do not rely on bounds or an adjustment set have to make other untestable assumptions on the causal mechanism. For example, [AIR96, KKS20, Gun20] rely on the existence of instrumental variables that are not affected by unobserved confounders and on restrictions of the model space. [MGTT18] and [LSM^{+}17] assume proxy variables that, while being correlated with unobserved confounders, do not confound the treatment and outcome themselves. Last, the deconfounder of [WB19] builds on the assumptions that there are no unobserved single-cause confounders.

3 Causal Reductions

In this section we introduce our key technical contribution that we refer to as a *causal reduction*. This is a simple procedure to replace any number of confounders by a single confounder that takes on values in the same space as the treatment variable, while preserving the important causal semantics of the model, in particular, the observational distribution and the causal effect of treatment on outcome. While we start by considering a simple setting with only two observed variables, we will extend this in Section 3.3 to account for additional observed confounders. The basic construction procedure can possibly be applied in more general settings as well. For the measure-theoretic justification of our derivation we refer the reader to [For21b]. While we here chose to make use of the formalism of causal Bayesian networks to derive the result, similar results can be obtained easily in other causal modeling frameworks, such as structural causal models and the potential-outcome framework.

3.1 Reduction without observed confounders

Consider a treatment variable $\mathbf{X} \in \mathcal{X}$ and an outcome variable $\mathbf{Y} \in \mathcal{Y}$. The spaces $\mathcal{X}$ and $\mathcal{Y}$ can be any standard measurable space, for example, $\mathcal{X} = \mathbb{R}^M$ and $\mathcal{Y} = \mathbb{R}^N$ in the continuous setting, or $\mathcal{X} = \{1, \dots, m\}$ and $\mathcal{Y} = \{1, \dots, n\}$ in the discrete setting, with binary treatment ($m = 2$) and a real-valued treatment variable ($M = 1$) as frequently occurring special cases. We assume that the outcome does not cause the treatment. Furthermore, let there exist K latent confounders $Z_1, \dots, Z_K$, where each $Z_i \in \mathcal{Z}_i$ also in some standard measurable space, with an arbitrary dependency structure. See Figure 1 (a) for an example of a corresponding Directed Acyclic Graph (DAG) of the causal Bayesian network.[1] Without loss of generality, we can summarize the K latent confounders $Z_1, \dots, Z_K$ with arbitrary dependency structure using a single latent confounder $\mathbf{Z} \in \mathcal{Z} = \prod_{k=1}^{K} \mathcal{Z}_k$:

$$p(\mathbf{x}, \mathbf{y}) = \int\limits_{\mathcal{Z}_1} \cdots \int\limits_{\mathcal{Z}_K} p(\mathbf{x}, \mathbf{y}, z_1, \dots, z_K) dz_1 \dots dz_K = \int\limits_{\mathcal{Z}} p(\mathbf{x}, \mathbf{y}, \mathbf{z}) d\mathbf{z}. \tag{1}$$

The resulting causal Bayesian network is shown in Figure 1 (b), which has the following factorization:

$$p(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p(\mathbf{y}|\mathbf{x}, \mathbf{z}) p(\mathbf{x}|\mathbf{z}) p(\mathbf{z}). \tag{2}$$

We aim to replace the above causal Bayesian network with one that is interventionally equivalent with respect to perfect interventions on any subset of $\{\mathbf{X}, \mathbf{Y}\}$, but where the latent confounder space $\mathcal{Z}$ can be smaller. In particular, this means that we will preserve the observational distribution $p(\mathbf{x}, \mathbf{y})$ and the "causal effect" $p(\mathbf{y}|\operatorname{do}(\mathbf{x}))$, that is, the distribution of $\mathbf{Y}$ for any perfect intervention on $\mathbf{X}$ that sets it to a value $\boldsymbol{x} \in \mathcal{X}$.

First, we generate a copy $\mathbf{W} := \mathbf{X}$ of the treatment variable $\mathbf{X}$. We will interpret $\mathbf{W}$ as a latent variable and $\mathbf{X}$ as an observed deterministic effect of $\mathbf{W}$ via the identity function $\mathbf{X} = \operatorname{id}(\mathbf{W})$. We obtain the Bayesian Network in Figure 1 (c):

$$p(\mathbf{x}, \mathbf{y}, \mathbf{w}, \mathbf{z}) = p(\mathbf{y}|\mathbf{x}, \mathbf{z}) p(\mathbf{x}|\mathbf{w}) p(\mathbf{w}|\mathbf{z}) p(\mathbf{z}), \tag{3}$$

where $p(\mathbf{w}|\mathbf{z}) := p(\mathbf{x}|\mathbf{z})|_{\mathbf{x}=\mathbf{w}}$ is a copy of the Markov kernel that appears as a factor in Equation 2, but evaluated in $\mathbf{w}$ rather than in $\mathbf{x}$. Furthermore, $p(\mathbf{x}|\mathbf{w}) := \delta_{\mathbf{w}}(\mathbf{x})$ is the Dirac measure centered at $\mathbf{w}$, representing the deterministic identity map from $\mathbf{W}$ to $\mathbf{X}$. If we marginalize out $\mathbf{W}$ we arrive at the initial causal Bayesian network in Figure 1 (b) again. Since interventions on observed variables commute with marginalizing over latent variables [BFPM21], the Bayesian networks in Figure 1 (b) and (c) are interventionally equivalent with respect to perfect interventions on any subset of $\{\mathbf{X}, \mathbf{Y}\}$.

Second, we refactorize the latent distribution as shown in Figure 1 (c), (d) and (e):

$$\begin{aligned} p(\mathbf{x}, \mathbf{y}, \mathbf{w}, \mathbf{z}) &= p(\mathbf{y}|\mathbf{x}, \mathbf{z}) p(\mathbf{x}|\mathbf{w}) p(\mathbf{w}|\mathbf{z}) p(\mathbf{z}) & (4) \\ &= p(\mathbf{y}|\mathbf{x}, \mathbf{z}) p(\mathbf{x}|\mathbf{w}) p(\mathbf{w}, \mathbf{z}) & (5) \\ &= p(\mathbf{y}|\mathbf{x}, \mathbf{z}) p(\mathbf{x}|\mathbf{w}) p(\mathbf{z}|\mathbf{w}) p(\mathbf{w}). & (6) \end{aligned}$$

The Bayesian networks representing these three factorizations are interventionally equivalent with respect to perfect interventions on any subset of $\{\mathbf{X}, \mathbf{Y}\}$, as we only factor the latent distributions differently (and do not consider interventions on the latent variables).

[1] Alternatively, one can interpret this as the graph of a Structural Causal Model, see for example [BFPM21].

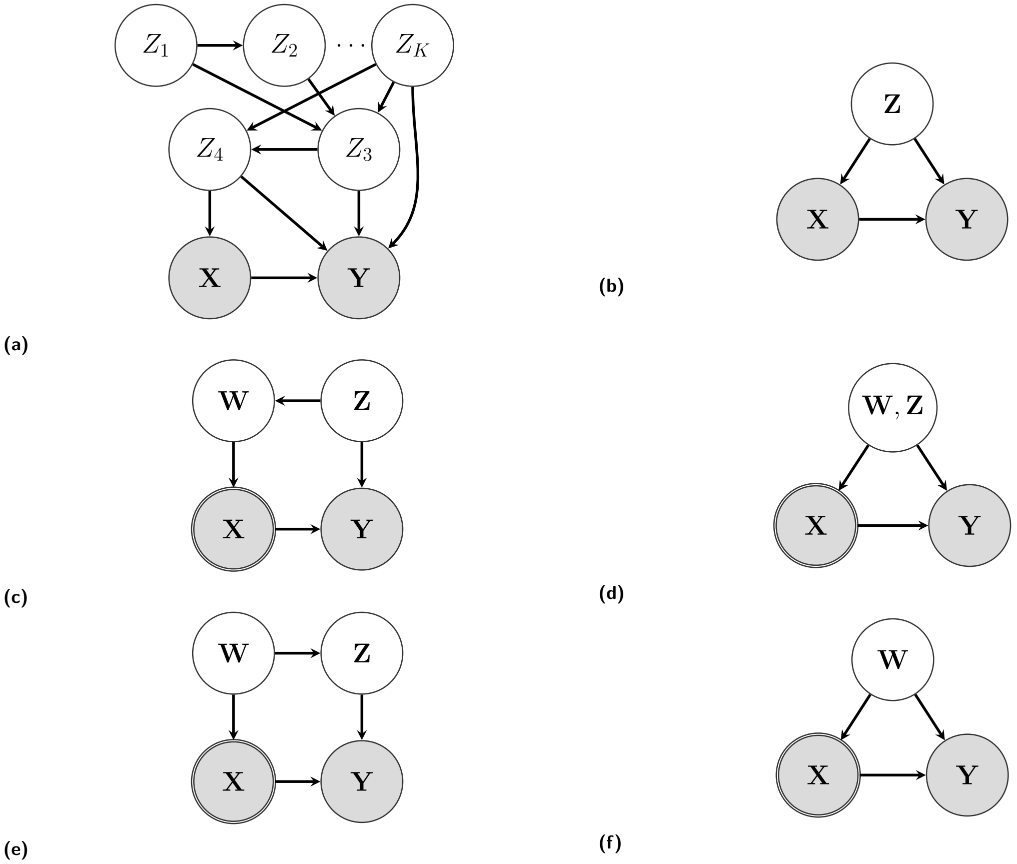

Figure 1. A graphical explanation of our causal reduction technique. (a) We assume a treatment variable $\mathbf{X}$, an outcome variable $\mathbf{Y}$, and K latent confounders $Z_1, \ldots, Z_K$ with an arbitrary dependency structure. (b) We represent the K latent confounders $Z_1, \ldots, Z_K$ by $\mathbf{Z} \in \mathcal{Z}$. (c) We create a copy of $\mathbf{X}$ called $\mathbf{W}$. We use a double circle to indicate that a variable is a deterministic function of its parents. (d, e) Instead of using the factorization from (c), $p(\mathbf{w}, \mathbf{z}) = p(\mathbf{w}|\mathbf{z})p(\mathbf{z})$, we choose $p(\mathbf{w}, \mathbf{z}) = p(\mathbf{z}|\mathbf{w})p(\mathbf{w})$. (f) Last, we marginalize over $\mathbf{Z}$. Note that at every step (a–f) the Bayesian networks are interventionally equivalent with respect to perfect interventions on any subset of $\{\mathbf{X}, \mathbf{Y}\}$. In particular, they all induce the same observational distribution $p(\mathbf{x}, \mathbf{y})$ and interventional distributions $p(\mathbf{y}|\,\mathrm{do}(\mathbf{x}))$.

Last, we can marginalize over $\mathbf{Z}$ and obtain:

$$p(\mathbf{x}, \mathbf{y}, \mathbf{w}) = p(\mathbf{y}|\mathbf{x}, \mathbf{w})p(\mathbf{x}|\mathbf{w})p(\mathbf{w}), \tag{7}$$

where we used the following composed Markov kernel:

$$p(\mathbf{y}|\mathbf{x}, \mathbf{w}) := \int p(\mathbf{y}|\mathbf{x}, \mathbf{z})p(\mathbf{z}|\mathbf{w})\, d\mathbf{z}. \tag{8}$$

Again, since marginalizing over latent variables and interventions on observed variables commute, the final Bayesian network in Figure 1 (f) is interventionally equivalent to the ones in Figure 1 (a–e) with respect to perfect interventions on any subset of $\{\mathbf{X}, \mathbf{Y}\}$.

Since $\mathbf{W}$ is a copy of $\mathbf{X}$, we successfully replaced the latent confounder space $\mathcal{Z} = \prod_{k=1}^{K} \mathcal{Z}_k$ with $\mathcal{X}$. In case $\mathcal{Z} = \mathbb{R}^K$ and $\mathcal{X} = \mathbb{R}^M$ and $M < K$, the dimensionality of the latent space will be reduced. In the common case of a one-dimensional $\mathbf{X}$, we expect $M = 1 \ll K$ and therefore achieve a significant reduction of the latent space. In the discrete case, the cardinality of $\mathcal{X}$ may be much lower than that of $\mathcal{Z}$. For example, in case treatment is binary, a single binary confounder in the model suffices.

We formulate the conclusion as a theorem:[2]

Theorem 3.1 (Causal Reduction)**.** *Let $\mathcal{M}$ be a causal Bayesian network with observed variables $\mathbf{X} \in \mathcal{X}, \mathbf{Y} \in \mathcal{Y}$ and latent variables, $Z_1 \in \mathcal{Z}_1, \dots, Z_K \in \mathcal{Z}_K$ such that $\mathbf{Y}$ is not an ancestor of $\mathbf{X}$.*

Then there exists a causal Bayesian network $\mathcal{M}^*$ with observed variables $\mathbf{X} \in \mathcal{X}$ and $\mathbf{Y} \in \mathcal{Y}$ and a single latent confounder $\mathbf{W} \in \mathcal{X}$ (that takes values in the same space as $\mathbf{X}$) such that $\mathcal{M}^*$ is interventionally equivalent to $\mathcal{M}$ with respect to perfect interventions on the observed variables $\mathbf{X}$ and $\mathbf{Y}$:

$$p_{\mathcal{M}}(\mathbf{x}, \mathbf{y}) = p_{\mathcal{M}^*}(\mathbf{x}, \mathbf{y})$$

$$p_{\mathcal{M}}(\mathbf{x} \mid \mathrm{do}(\mathbf{y})) = p_{\mathcal{M}^*}(\mathbf{x} \mid \mathrm{do}(\mathbf{y}))$$
$$p_{\mathcal{M}}(\mathbf{y} \mid \mathrm{do}(\mathbf{x})) = p_{\mathcal{M}^*}(\mathbf{y} \mid \mathrm{do}(\mathbf{x})).$$

We call the causal Bayesian network $\mathcal{M}^*$ a *causal reduction* of $\mathcal{M}$ since it will typically be the case that the latent space will be reduced, while the causal semantics are preserved by construction. The single latent confounder $\mathbf{Z}$ in $\mathcal{M}^*$ will then more parsimoniously represent the causal influence of *all* latent confounders of $\mathbf{X}$ and $\mathbf{Y}$ in $\mathcal{M}$.

From this result, it also follows that a reduced causal Bayesian network can be constructed from a given acyclic structural causal model. Indeed, this is immediate since any acyclic structural causal model can also be interpreted as a causal Bayesian network with latent variables. Extending the derivation to *simple* structural causal models (a convenient class of structural causal models that can represent causal cycles, such as feedback loops [BFPM21]) is straightforward, as long as $\mathbf{X}$ and $\mathbf{Y}$ are not part of a causal cycle (although the other variables might be involved in cycles). Even more generally, we can also start with given potential outcomes and obtain a reduced causal Bayesian network. This is shown explicitly in Appendix B.

3.2 Replacing conditional distributions by functions

Whereas previously, we made use of causal Bayesian networks to formulate our reduction operation, we now move to Structural Causal Models (SCMs) [Pea09, BFPM21] in order to obtain convenient parameterizations in the non-parametric setting. We make use of the exogenous variables $\mathbf{U}, \mathbf{V}$ to represent the noise in the reduced causal model. This, in turn, allows us to express all causal relationships as deterministic functions. Estimating the model then boils down to estimating these functions, as we will illustrate in Section 6.

[2] All proofs are provided in Appendix A.

Theorem 3.2. *Let $\mathbf{Y}$ be a 'conditional' $\mathbb{R}^N$-valued random variable with Markov kernel $P(\mathbf{Y}|\mathbf{X})$ and input $\mathbf{X}$ that can take values in any measurable space. Then there exists an N-dimensional standard normal random variable $\mathbf{V} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$ independent of $\mathbf{X}$ and a deterministic measurable map F such that:*

$$\mathbf{Y} = F(\mathbf{V}, \mathbf{X}) \quad a.s. \tag{9}$$

Furthermore, the map F is 'well-behaved', in the sense that it is composed out of (inverse) conditional cumulative distribution functions.

Applying Theorem 3.2 twice gives a reduced SCM from the reduced causal Bayesian network in Equation 7 with structural equations

$$\mathbf{X} = F(\mathbf{U}), \tag{10}$$

$$\mathbf{Y} = G(\mathbf{U}, \mathbf{V}, \mathbf{X}), \tag{11}$$

where $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_M), \mathbf{V} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$ and $\mathbf{U} \perp\!\!\!\perp \mathbf{V}$, and F, G are deterministic maps. This SCM encodes the same observational distribution $p(\mathbf{x}, \mathbf{y})$ and interventional distributions $p(\mathbf{y}|\operatorname{do}(\mathbf{x})), p(\mathbf{x}|\operatorname{do}(\mathbf{y}))$ as the causal Bayesian network. This allows us to "parameterize" the reduced causal model in terms of the two functions F and G.

3.3 Reduction with observed confounders

There are many scenarios where we are interested in estimating the conditional causal effect of interventions given additional covariates $\mathbf{C}$ that might confound treatment and outcome, for example when estimating the efficacy of a vaccine depending on age. We again consider a treatment variable $\mathbf{X} \in \mathcal{X}$, an outcome variable $\mathbf{Y} \in \mathcal{Y}$, and a set of K latent confounders $Z_1, \dots, Z_K$ in arbitrary standard measurable spaces $\mathcal{Z}_1, \dots, \mathcal{Z}_K$. In addition, let there be L observed confounders $C_1, \dots, C_L$ of $\mathbf{X}$ and $\mathbf{Y}$, again in arbitrary standard measurable spaces. We allow for arbitrary causal relations and dependencies between the confounders. In the following, we summarize all observed confounders using a single variable $\mathbf{C} = (C_1, \dots, C_L) \in \mathcal{C}$ and all latent confounders as $\mathbf{Z} = (Z_1, \dots, Z_K) \in \mathcal{Z}$. We follow a similar sequence of steps as in Section 3 to derive a reduced causal model of the following form

$$p(\mathbf{x}, \mathbf{y}, \mathbf{w}, \mathbf{c}) = p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \mathbf{c}) p(\mathbf{x}|\mathbf{w}) p(\mathbf{w}|\mathbf{c}) p(\mathbf{c}). \tag{12}$$

as illustrated in Figure 2 (a–d). At every step, the Bayesian network is observationally equivalent to the ones before and also interventionally equivalent for perfect interventions on any subset of $\{\mathbf{X}, \mathbf{Y}\}$.

Specializing now to continuous treatment $\mathbf{X} \in \mathcal{X} = \mathbb{R}^M$ and outcome $\mathbf{Y} \in \mathcal{Y} = \mathbb{R}^N$, we can use a similar approach as in Section 3.2, and in addition marginalize out $\mathbf{W}$ as seen in Figure 2 (g), to convert the causal Bayesian Network into an SCM with structural equations of the form given below

$$\mathbf{X} = F(\mathbf{U}, \mathbf{C}), \tag{13}$$

$$\mathbf{Y} = G(\mathbf{V}, \mathbf{X}, \mathbf{U}, \mathbf{C}), \tag{14}$$

where $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_M)$, $\mathbf{V} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$, $\mathbf{U} \perp\!\!\!\perp \mathbf{V}$ and F and G are two deterministic maps. This is illustrated in Figure 2 (e–h). Again, at every step, the Bayesian network is observationally equivalent to the ones before and also interventionally equivalent for perfect interventions on any subset of $\{\mathbf{X}, \mathbf{Y}\}$.

4 Parameter estimation in the discrete case

In this section we present the first application of our causal reduction technique. We formulate the maximum likelihood estimator of the causal effect from observational and interventional data, considering the case in which all observed variables are discrete. The exact maximum likelihood estimator exploits that the parameters of the two prediction tasks are shared.

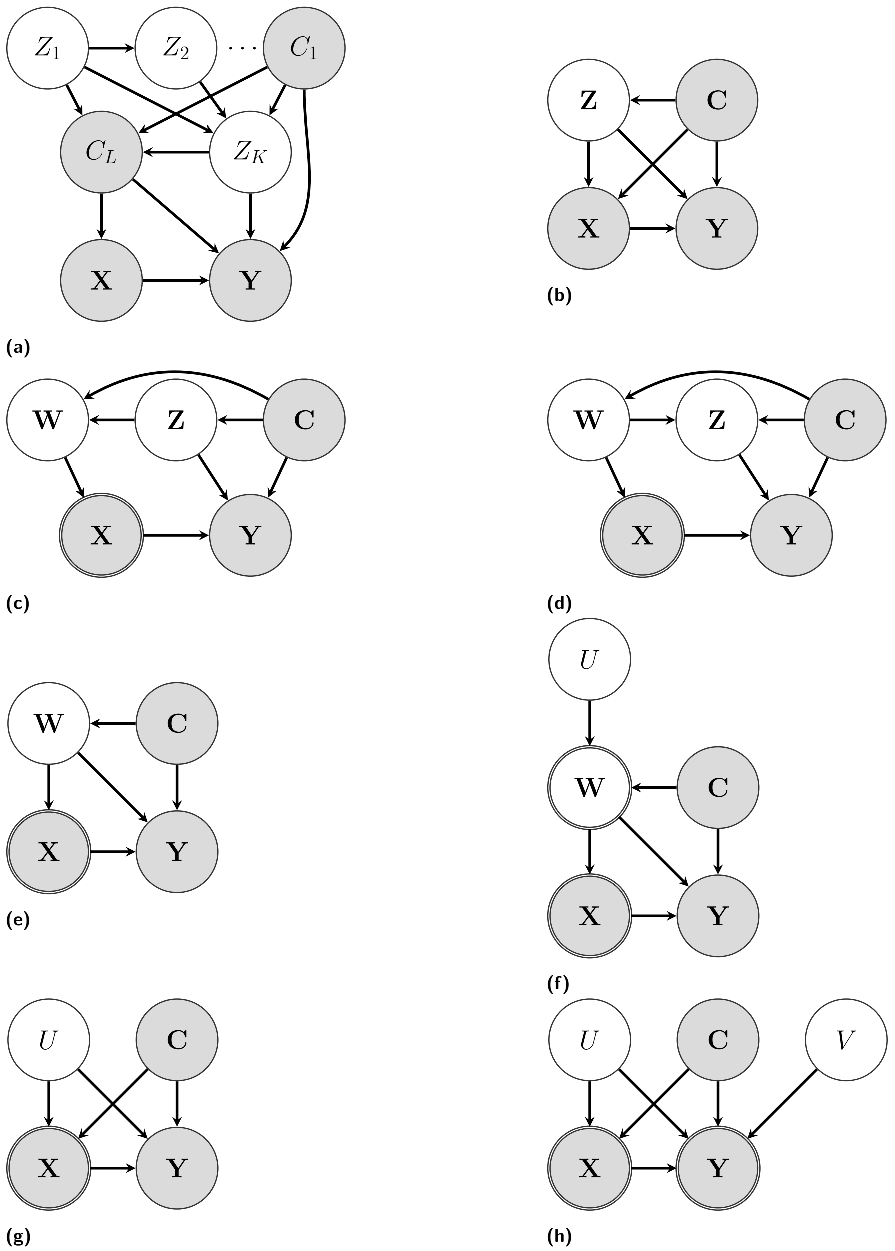

Figure 2. A graphical explanation of our reduction technique in the presence of both observed and latent confounders. (a) We assume a treatment variable $\mathbf{X} \in \mathcal{X}$, an outcome variable $\mathbf{Y} \in \mathcal{Y}$, latent confounders $Z_1, \ldots, Z_K$, and observed confounders $C_1, \ldots, C_L$, with arbitrary causal and probabilistic relations between the confounders. (b) We combine the latent confounders into $\mathbf{Z} \in \mathcal{Z}$ and the observed confounders into $\mathbf{C} \in \mathcal{C}$, and factorize their joint distribution as $p(\mathbf{z} \mid \mathbf{c})p(\mathbf{c})$. (c) We create a copy of $\mathbf{X}$ called $\mathbf{W}$. (d) We refactorize $p(\mathbf{w}, \mathbf{z}, \mathbf{c})$ as $p(\mathbf{z} \mid \mathbf{w}, \mathbf{c})p(\mathbf{w} \mid \mathbf{c})p(\mathbf{c})$. (e) We marginalize over $\mathbf{Z}$. (f) We reparameterize $p(\mathbf{w} \mid \mathbf{c})$ using Theorem 3.2 as a deterministic function, introducing an independent noise variable $\mathbf{U}$. (g) We marginalize over $\mathbf{W}$. (h) We reparameterize $p(\mathbf{y} \mid \mathbf{x}, \mathbf{u}, \mathbf{c})$ with Theorem 3.2 as a deterministic function, introducing an independent noise variable $\mathbf{V}$. Note that at every step (a–h) the models remain observationally equivalent (i.e., $p(\mathbf{c}, \mathbf{x}, \mathbf{y})$ is invariant), and also interventionally equivalent with respect to perfect interventions on $\mathbf{X}$ and $\mathbf{Y}$ (i.e., $p(\mathbf{y}, \mathbf{c} \mid \mathrm{do}(\mathbf{x}))$ and $p(\mathbf{x}, \mathbf{c} \mid \mathrm{do}(\mathbf{y}))$ are invariant).

4.1 Bounds on the causal effect

Let us consider treatment variable $X \in \mathcal{X}$, outcome variable $Y \in \mathcal{Y}$, observed confounder $\mathbf{C} \in \mathcal{C}$ and our surrogate latent confounder $W \in \mathcal{X}$ that was obtained by the causal reduction operation. Assume that $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{C}$ are finite or countably infinite. For simplicity, we start with the case without observed covariates $\boldsymbol{C}$. In the reduced causal model, we obtain the following expression for the causal effect of X on Y (i.e., the interventional distributions) from Equation 7:

$$p(y \mid \mathrm{do}(x)) = \sum_{w \in \mathcal{X}} p(w) p(y \mid x, w), \tag{15}$$

while for the observational distribution we obtain:

$$\begin{aligned} p(x, y) &= \sum_{w \in \mathcal{X}} p(w) p(x \mid w) p(y \mid x, w) \\ &= \sum_{w \in \mathcal{X}} p(w) \delta_w(x) p(y \mid x, w) \\ &= p(W = x) p(y \mid X = x, W = x). \end{aligned} \tag{16}$$

From this we immediately obtain bounds relating the observational and interventional distributions. Splitting the sum in (15) and substituting (16), we get:

$$\begin{aligned} p(y \mid \mathrm{do}(x)) &= \sum_{w \in \mathcal{X}} p(w) p(y \mid x, w) \\ &= p(W = x) p(y \mid X = x, W = x) + \sum_{\substack{w \in \mathcal{X} \\ w \neq x}} p(w) p(y \mid x, w) \\ &= p(x, y) + \sum_{\substack{w \in \mathcal{X} \\ w \neq x}} p(w) p(y \mid x, w). \end{aligned}$$

Since $0 \leq p(w) \leq 1$ and $0 \leq p(y \mid x, w) \leq 1$, we can bound

$$0 \leq \sum_{\substack{w \in \mathcal{X} \\ w \neq x}} p(w) p(y \mid x, w) \leq \sum_{\substack{w \in \mathcal{X} \\ w \neq x}} p(w) = 1 - p(W = x) = 1 - p(x),$$

and we conclude

$$p(x, y) \leq p(y \mid \mathrm{do}(x)) \leq p(x, y) + 1 - p(x) \tag{17}$$

for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$. This means that the interventional distributions corresponding to the causal effect of X on Y are bounded by the observational distribution of X and Y. For binary treatment and outcome, this bound is not novel, as it was already derived in [MN98] and can also be obtained from the instrumental inequality of [BP97] under the assumption of perfect compliance.

In the presence of covariates $\boldsymbol{C}$, one similarly derives the bound

$$p(x, y|\boldsymbol{c}) \leq p(y \mid \mathrm{do}(x), \boldsymbol{c}) \leq p(x, y|\boldsymbol{c}) + 1 - p(x|\boldsymbol{c}) \tag{18}$$

for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$, $\boldsymbol{c} \in \mathcal{C}$.

4.2 Maximum likelihood estimation

To formulate the corresponding maximum likelihood estimation problem, we introduce real-valued parameters

$$\begin{cases} \phi_w := p(w) & \text{for } w \in \mathcal{X} \\ \vartheta_{y|x} := p(Y = y \mid X = x) & \text{for } x \in \mathcal{X}, y \in \mathcal{Y} \\ \psi_{y|x} := p(y \mid \mathrm{do}(x)) & \text{for } x \in \mathcal{X}, y \in \mathcal{Y}. \end{cases}$$

The parameters ϕ and ϑ parameterize the observational distribution, while ψ directly parameterizes the interventional distributions. These parameters are subject to the following constraints:

$$\forall x \in \mathcal{X} : 0 \leq \phi_x \leq 1, \qquad \sum_{x \in \mathcal{X}} \phi_x = 1 \tag{19}$$

$$\forall x \in \mathcal{X}\ \forall y \in \mathcal{Y} : 0 \leq \vartheta_{y|x} \leq 1, \qquad \forall x \in \mathcal{X} : \sum_{y \in \mathcal{Y}} \vartheta_{y|x} = 1 \tag{20}$$

$$\forall x \in \mathcal{X}\ \forall y \in \mathcal{Y} : \phi_x \vartheta_{y|x} \leq \psi_{y,x} \leq \phi_x \vartheta_{y|x} + 1 - \phi_x, \qquad \forall x \in \mathcal{X} : \sum_{y \in \mathcal{Y}} \psi_{y|x} = 1 \tag{21}$$

Apart from the obvious nonnegativity and normalization constraints, these also include the bounds of Equation 17.

Given an observational dataset $(x_1^O, y_1^O), \dots, (x_{N_O}^O, y_{N_O}^O)$ and an interventional dataset $(x_1^I, y_1^I), \dots, (x_{N_I}^I, y_{N_I}^I)$, we can write down the total log-likelihood in terms of the parameters:

$$\begin{aligned} \ell(\phi, \vartheta, \psi; x^I, y^I, x^O, y^O) &= \sum_{j=1}^{N_I} \log p(y_j^I \mid \mathrm{do}(x_j^I)) + \sum_{j=1}^{N_O} \log p(x_j^O, y_j^O) \\ &= \sum_{j=1}^{N_I} \log \psi_{y_j^I, x_j^I} + \sum_{j=1}^{N_O} \big(\log \phi_{x_j^O} + \log \vartheta_{y_j^O | x_j^O} \big) \end{aligned} \tag{22}$$

This decomposes as a sum of two terms, the first term only involving the interventional data and the parameters ψ, the second term only involving the observational data and the parameters ϑ, ϕ. However, importantly, the two problems are not independent, as the parameters have to satisfy the constraints in Equation 21.

In principle, one can solve for the maximum likelihood estimator by convex programming and using the technique of Lagrange multipliers to take into account the (equality and inequality) constraints on the parameters. We will not do so here as the calculation seems to become rather cumbersome, and it does not seem possible to obtain an analytical closed-form expression for the maximum likelihood estimator. However, standard numerical optimization techniques can be applied to compute the maximum likelihood estimator numerically.

4.3 Infinite observational data

To gain some intuition, we will instead consider the limit that the interventional data size N_I is fixed, whereas the observational data size $N_O \to \infty$. In that limit, the observational part of the log-likelihood dominates, and we can easily optimize this with respect to ϑ and ϕ to obtain:

$$\hat{\phi}_x = \frac{N_{x+}^O}{N_{++}^O}, \qquad \hat{\vartheta}_{y|x} = \frac{N_{xy}^O}{N_{x+}^O}$$

where $N_{++}^O := \sum_{x \in \mathcal{X}} N_{x+}^O$, with $N_{x+}^O := \sum_{y \in \mathcal{Y}} N_{xy}^O$, with $N_{xy}^O := \sum_{j=1}^{N_O} \mathbb{1}_x(x_j^O) \mathbb{1}_y(y_j^O)$. Given these (asymptotically) optimal $\hat{\vartheta}, \hat{\phi}$ we can now maximize the interventional part of the log-likelihood with respect to ψ, where we have to bear in mind the constraints in Equation 21. The optimum for $\psi_{y,x}$ will either be at the lower or the upper bound, or in between. Ignoring the bounds, the optimum would be taken at $\frac{N_{y|x}^I}{N_{+|x}^I}$. If this falls below the lower bound, or above the upper bound, the value will be clipped at the respective bound. Hence,

$$\hat{\psi}_{y,x} = \begin{cases} \hat{\phi}_x \hat{\vartheta}_{y|x} & \frac{N_{y|x}^I}{N_{+|x}^I} < \hat{\phi}_x \hat{\vartheta}_{y|x} \\ \frac{N_{y|x}^I}{N_{+|x}^I} & \hat{\phi}_x \hat{\vartheta}_{y|x} \leq \frac{N_{y|x}^I}{N_{+|x}^I} \leq \hat{\phi}_x \hat{\vartheta}_{y|x} + 1 - \hat{\phi}_x \\ \hat{\phi}_x \hat{\vartheta}_{y|x} + 1 - \hat{\phi}_x & \hat{\phi}_x \hat{\vartheta}_{y|x} + 1 - \hat{\phi}_x < \frac{N_{y|x}^I}{N_{+|x}^I} \end{cases} \tag{23}$$

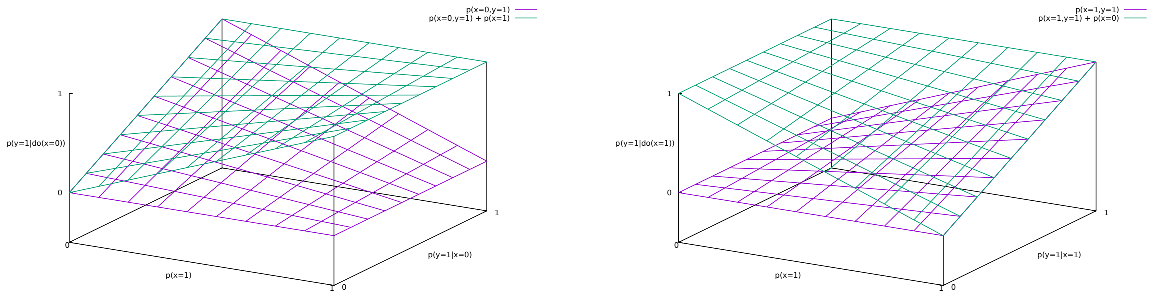

Figure 3. Illustration of the bounds in Equation 24 (left) and Equation 25 (right) of the interventional distributions in terms of the observational distribution.

Since N^O is large, the observational parameters ϕ, ϑ are estimated with high accuracy. This can help improve the accuracy in the estimated interventional parameters ψ by means of the bound, especially in cases where only few interventional data points are available (i.e., if some $N^I_{+|x} \ll N^O_{+|x}$). For larger values of N^I, the estimator $\hat{\psi}_{y,x}$ will ignore the observational data, and reduce to the standard causal effect estimator from interventional data, $\frac{N^I_{y|x}}{N^I_{+|x}}$.

4.4 Case study: mostly untreated observational patient population

We now consider one specific example to get an impression of when this estimator may yield a benefit, and how large the benefit can be. We assume treatment and outcome to be binary. We can then focus on estimating the *average treatment effect (ATE)*,

$$\tau := p(Y = 1 \mid \mathrm{do}(X = 1)) - p(Y = 1 \mid \mathrm{do}(X = 0)).$$

The maximum likelihood (plug-in) estimator for this is just

$$\hat{\tau}^{IO} := \hat{\psi}_{1,1} - \hat{\psi}_{1,0}$$

where we used the superscript "IO" to indicate that this estimator uses interventional and observational data. The classical estimator of the ATE based on interventional data only is

$$\hat{\tau}^I := \frac{N^I_{1|1}}{N^I_{+|1}} - \frac{N^I_{1|0}}{N^I_{+|0}}.$$

Our statistical model has a 5-dimensional parameterization. The bounds in Equation 17 read explicitly:

$$p(x = 0, y = 1) \leq p(y = 1 \mid \mathrm{do}(x = 0)) \leq 1 - p(x = 0, y = 0) \tag{24}$$
$$p(x = 1, y = 1) \leq p(y = 1 \mid \mathrm{do}(x = 1)) \leq 1 - p(x = 1, y = 0) \tag{25}$$

See Figure 3 for an illustration of these bounds.

Let us consider the following scenario. We have a new drug and we are setting up an RCT. We have historical observational data where all individuals are untreated, i.e., $p(x = 1) = 0$. The bound in Equation 24 then becomes tight, yielding

$$p(y = 1) = p(y = 1 \mid \mathrm{do}(x = 0)).$$

On the other hand, the bound in Equation 25 becomes non-informative, and can be ignored. Thus, we can identify $p(y = 1 \mid \mathrm{do}(x = 0))$ from the observational data as $p(y = 1)$. When setting up an RCT, we therefore only need a treatment group with $\mathrm{do}(x = 1)$ in order to identify $p(y = 1 \mid \mathrm{do}(x = 1))$, and there is no need to include a group in which individuals are not treated. In this way, we can reduce the number of participants of the RCT by a factor of two without loosing accuracy (comparing with a standard RCT with a 50%-50% split in treatment and control group). We might even gain in accuracy if N_O is large, because that allows to estimate $p(y = 1)$, and hence $p(y = 1 \mid \mathrm{do}(x = 0))$, and therefore τ, more accurately.

An important remark, though, is that in this way we do not control for a placebo effect. If one is specifically interested in comparing the effect of treatment with the drug with treatment with a placebo, it seems that the observational data (in which individuals were not even treated with a placebo) have nothing to offer in this scenario. On the other hand, if the aim is to compare the effect of treatment with the drug to no treatment at all, one can save out on the control group in the RCT. Indeed, in this scenario there is simply no room for potential unobserved confounding in the observational data to bias the estimate. In practice, when applying this idea, one should make sure that the observational data stem from a similar population as the treatment group.

More generally, if the drug is not new, and not everyone in the observational data is untreated, but only a small fraction is treated, the observational data may still allow for a more accurate estimate of $p(y = 1 \mid \mathrm{do}(x = 0))$ than the interventional data if the number of interventional samples with $x = 0$ is small. A similar situation can arise if only a small fraction of the observational population were *un*treated. In general, the approximate ML estimator in Equation 23 can be used as long as the observational data size is large enough. Even more generally, one can use the exact ML estimator, which can be computed numerically. In that case, it is not even necessary to assume that the observational data set is much larger than the interventional data set for obtaining an estimate of the causal effect based on a non-trivial combination of observational and interventional data.

5 Parameter constraints in the linear Gaussian case

For the second application of our causal reduction, we consider the case where all causal relationships in Figure 1 (a) are linear, and all distributions are Gaussian. We can then guarantee that the reduced causal model is linear Gaussian as well.

Corollary 5.1 (Reduced linear Gaussian model)**.** *Consider a linear Gaussian SCM (or causal Bayesian network with possible latent variables) with observed variables* $\mathbf{X}$ *and* $\mathbf{Y}$ *such that* $\mathbf{Y}$ *is not ancestor of* $\mathbf{X}$*. Then this causal model is interventionally equivalent to a reduced linear Gaussian causal model with the following structural equations:*

$$\mathbf{X} = \mathbf{a} + B\mathbf{U}, \tag{26}$$

$$\mathbf{Y} = \mathbf{c} + D\mathbf{X} + E\mathbf{U} + F\mathbf{V}, \tag{27}$$

with vectors $\mathbf{a}$*,* $\mathbf{c}$ *and matrices* B*,* D*,* E*,* F*, where* B *and* F *can be chosen to be lower-triangular with non-negative diagonal entries, and where* $\mathbf{U}$ *is a standard Gaussian latent variable of the same dimension as* $\mathbf{X}$ *and where* $\mathbf{V}$ *is a standard Gaussian latent variable of the same dimension as* $\mathbf{Y}$ *that is independent of* $\mathbf{U}$*.*

We use the reduced linear Gaussian model from Corollary 5.1 to prove that the parameters of the interventional distribution constrain the parameters of the observational distribution.

Theorem 5.2 (Linear Gaussian parameter constraints)**.** *Consider a linear-Gaussian SCM (or causal Bayesian network with possible latent variables) with two observed variables* $\mathbf{X}$ *and* $\mathbf{Y}$ *such that* $\mathbf{Y}$ *is not ancestor of* $\mathbf{X}$*. The entailed observational and interventional distributions are Gaussian. Modeling*

$p(\mathbf{x}), p(\mathbf{y}|\mathbf{x})$ and $p(\mathbf{y}|\operatorname{do}(\mathbf{x}))$ independently from each other could be done with the following parameterization:

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\alpha}, \Sigma\right), \tag{28}$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}\left(\mathbf{y}|\boldsymbol{\gamma} + \Delta\mathbf{x}, \Pi\right), \tag{29}$$

$$p(\mathbf{y}|\operatorname{do}(\mathbf{x})) = \mathcal{N}\left(\mathbf{y}|\widetilde{\boldsymbol{\gamma}} + \widetilde{\Delta}\mathbf{x}, \widetilde{\Pi}\right), \tag{30}$$

with covariance matrices Σ, Π, $\widetilde{\Pi}$. However, using the reduced causal model from Corollary 5.1 we find that these parameters are constrained by the following relations:

$$(\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) + \left(\widetilde{\Delta} - \Delta\right)\boldsymbol{\alpha} = 0, \tag{31}$$

$$\left(\widetilde{\Delta} - \Delta\right)\Sigma\left(\widetilde{\Delta} - \Delta\right)^{\top} + \Pi = \widetilde{\Pi}. \tag{32}$$

From Equation 32 we can easily see that $\widetilde{\Pi} - \Pi$ is positive semidefinite. Furthermore, we see that these constraints lead to a reduced parameter count, N parameters for Equation 31 and $N(N+1)/2$ parameters for Equation 32, assuming $\mathbf{y}$ to be N-dimensional. In total, we have reduced the parameter count by $N(N+3)/2$ by modeling the parameters of the observational and interventional distributions jointly.

In the linear Gaussian case, the reduced causal model tells us exactly how many parameters we need to model the observational and interventional distribution and which parameters are shared. Indeed, the parameters $\mathbf{c}, D, E$ and, F are shared between the observational and interventional distribution. We can estimate them jointly using observational and interventional data, effectively reducing sample complexity when trying to model the interventional distribution. This can be beneficial for causal effect estimation when we assume that we only have access to a small number of interventional samples and a large number of observational samples.

6 Estimation in the nonlinear, non-Gaussian case

In the final application of our causal reduction, we derive a flexible parameterization of the reduced model, which enables us to estimate the observational and interventional distributions by jointly learning from observational and interventional data without making strong parametric assumptions, where we parameterize the functions F and G in Equations 13 and 14 to learn the model from data. Intuitively, the reduced SCM derived in Section 3.2 tells us that the parameters of G are shared among observational and interventional samples for an intervention on $\mathbf{X}$, whereas the parameters of F are not.

6.1 Parameterization using normalizing flows

While parameterizing the functions F and G can be done in many different ways, we here use diffeomorphisms, i.e., differentiable mappings with a differentiable inverse. Using the change-of-variables formula, we can derive a maximum-likelihood estimator for the mappings' parameters that can be efficiently optimized through backpropagation. In the deep learning community, those invertible and differentiable mappings are called *normalizing flows*, and much recent research went into finding flexible and easily invertible mappings. See for example [PCdCG20] and [KMLH20] for other current applications of normalizing flows to approximate nonlinear causal mechanisms.

Our flow model consists of two flows, where the first flow corresponds to F and is trained using observational data, while the second flow corresponds to G and is trained using observational and interventional data. In the following, we derive the loss function for observational and interventional data separately. For the remaining part of this sectoin we focus on one-dimensional treatment outcome pairs, i.e. $X \in \mathcal{X} = \mathbb{R}$ and $Y \in \mathcal{Y} = \mathbb{R}$, and (optionally) an L-dimensional observed confounder $\mathbf{C} \in \mathcal{C} = \mathbb{R}^L$.

6.1.1 Observational data

To keep the notation simple, we will henceforth suppress the dependence on $\mathbf{c}$. According to the SCM in Equations 13 and 14, the joint likelihood $p(x, y)$ can be factorized as $\log p(x, y) = \log p(y|x) + \log p(x)$. We now use the following bijective transformations between observed variables x, y and latent variables u, v

$$u = f_{\boldsymbol{\phi}}(x), \tag{33}$$

$$v = g_{x,u;\boldsymbol{\theta}}(y), \tag{34}$$

where the functions $f_{\boldsymbol{\phi}}$ and $g_{x,u;\boldsymbol{\theta}}$ are invertible for all $\boldsymbol{\phi}, x, u, \boldsymbol{\theta}$. Here, $f_{\boldsymbol{\phi}} = F^{-1}$ from Equation 10 and $g_{x,u;\boldsymbol{\theta}}$ is the inverse of $v \mapsto G(u, v, x)$ from Equation 11 (for fixed $u, x, \boldsymbol{\theta}$).[3]

The SCM also specifies that $u \sim \mathcal{N}(0, 1)$, $v \sim \mathcal{N}(0, 1)$ and $u \perp\!\!\!\perp v$. The transformations defined above allow us to rewrite the joint likelihood using the change of variable formula

$$\begin{aligned}\log p(x, y) &= \log p_V(g_{x,u;\boldsymbol{\theta}}(y)) + \log \left|\frac{\partial g_{x,u;\boldsymbol{\theta}}(y)}{\partial y}\right| + \log p_U(f_{\boldsymbol{\phi}}(x)) + \log \left|\frac{\partial f_{\boldsymbol{\phi}}(x)}{\partial x}\right|. \\ &= \log p_V(g_{x,f_{\boldsymbol{\phi}}(x);\boldsymbol{\theta}}(y)) + \log \left|\frac{\partial g_{x,f_{\boldsymbol{\phi}}(x);\boldsymbol{\theta}}(y)}{\partial y}\right| + \log p_U(f_{\boldsymbol{\phi}}(x)) + \log \left|\frac{\partial f_{\boldsymbol{\phi}}(x)}{\partial x}\right|. \end{aligned} \tag{35}$$

where in the last step, we substituted $u = f_{\boldsymbol{\phi}}(x)$ into $g_{x,u;\boldsymbol{\theta}}(y)$. The parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are jointly updated by minimizing $\sum_{o=1}^{N_O} - \log p(x_o^O, y_o^O)$ given N_O observational training samples $(x_1^O, y_1^O), \ldots, (x_{N_O}^O, y_{N_O}^O)$.

6.1.2 Interventional data

In contrast to the observational setting, we only have to consider the conditional likelihood $p(y| \operatorname{do}(x))$ in the interventional case. Since we cannot use $f_{\boldsymbol{\phi}}(x)$ to impute u, we instead marginalize over u

$$\log p(y \mid \operatorname{do}(x)) = \log \int p(y| \operatorname{do}(x), u) p(u) du. \tag{36}$$

Substituting the bijective mapping $v = g_{x,u;\boldsymbol{\theta}}(y)$ into Equation 36, we obtain

$$\log p(y \mid \operatorname{do}(x)) = \log \int p_V(g_{x,u;\boldsymbol{\theta}}(y)) \left|\frac{\partial g_{x,u;\boldsymbol{\theta}}(y)}{\partial y}\right| p(u) du. \tag{37}$$

Since this is a one-dimensional integral, we can approximate it accurately numerically by means of the trapezoidal rule. The parameter $\boldsymbol{\theta}$ can be updated by minimizing $\sum_{i=1}^{N_I} - \log p(y_i^I | \operatorname{do}(x_i^I))$ given N_I interventional training samples $(x_1^I, y_1^I), \ldots, (x_{N_I}^I, y_{N_I}^I)$.

6.1.3 Joint optimization

Assuming we have N_O observational samples and N_I interventional samples, we define the full loss as given by

$$\ell = \frac{1}{N_O} \sum_{o=1}^{N_O} - \log p(x_o^O, y_o^O) + \frac{1}{N_I} \sum_{i=1}^{N_I} - \log p(y_i^I | \operatorname{do}(x_i^I)). \tag{38}$$

The parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ of the transformation f and g are learned by minimizing the loss using gradient descent. In Equation 38, we scale each loss term by the number of samples used for training to balance their contribution during optimization.

[3] In the presence of observed confounding, we simply have to replace the functions $f_{\boldsymbol{\phi}}$ and g by functions $f_{\mathbf{c};\boldsymbol{\phi}}$ and $g_{x,u,\mathbf{c};\boldsymbol{\theta}}$.

6.1.4 Sampling from the model

After training we can easily generate observational and interventional samples from the trained model. The sampling procedure from the observational conditional distribution $p(y \mid x)$ consists of the following steps:

$$\begin{aligned} v &\sim \mathcal{N}(0,1), \\ u &\leftarrow f_{\boldsymbol{\phi}}(x^O), \\ y^O &\leftarrow g^{-1}_{x^O,u;\boldsymbol{\theta}}(v), \end{aligned}$$

where we assume $x^O \in \mathbb{R}$ to be given. If we instead want to generate an interventional sample from $p(y \mid \mathrm{do}(x))$, the sampling procedure is as follows:

$$\begin{aligned} v &\sim \mathcal{N}(0,1), \\ u &\sim \mathcal{N}(0,1), \\ y^I &\leftarrow g^{-1}_{x^I,u;\boldsymbol{\theta}}(v), \end{aligned}$$

where we assume $x^I \in \mathbb{R}$ to be observed.

6.2 Experiments

We perform a series of experiments on simulated data, where the causal relationships between all variables are nonlinear, showing that we can significantly reduce the number of interventional samples required to estimate the interventional distribution $p(y|\,\mathrm{do}(x))$ by training jointly with (possibly confounded) observational and interventional samples. Throughout this section, we are using the parameterization described in Section 6, where we use linear rational spline flows [DEL20]. For a detailed description of this choice, see Appendix C. We perform two sets of experiments: (1) We consider K latent confounders $Z_1, \ldots, Z_K \in \mathbb{R}$ with an arbitrary dependency structure. (2) We consider L additional, observed confounders $C_1, \ldots, C_L \in \mathbb{R}$ with an arbitrary dependency structure. All flow models are implemented with the automatic differentiation packages Pytorch [PGM+19] and Pyro [BCJ+19]. All code is available under https://github.com/max-ilse/CausalReduction.

6.2.1 Without observed confounders

We simulate cause and effect pairs from the SCM with structural equations: $X = F(E_X, \mathbf{Z})$, $Y = G(X, E_Y, \mathbf{Z})$. A single dataset consists of observational and interventional samples. All causal relationships are simulated using fully connected neural networks with a single hidden layer, where the weights are randomly initialized. The activation functions are REctified Linear Units (ReLUs). As a result, the simulated causal mechanisms are nonlinear. The values of $E_X, E_Y, \mathbf{Z}$ and $\mathrm{do}(X)$ are sampled from a random

Table 1. Comparison of a flow model trained with interventional samples only and a flow model trained with interventional and observational samples. We calculate the ratio N_I^*/N_I, where N_I^* is the number of interventional samples necessary to match the interventional test log-likelihood of a flow model trained with N_I interventional and $N_O = 1000$ observational samples. E.g. in the case of dataset 3 and $N_I = 100$, if we were to use only interventional samples, we would require twice as many interventional samples compared to using 100 interventional and 1000 observational samples. For dataset 11 to 15, we simulate an additional observed confounder $\mathbf{C}$. Note that if a large number of interventional samples ($250 < N_I \leq 1000$) are available the improvements become smaller as shown in Appendix D.3.

N_I	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
50	1.4	1.8	2.2	1.2	0.2	2.2	2.1	1.7	1.9	1.6	3.2	3.2	2.2	2.7	3.2
100	0.8	2.6	2.0	1.5	0.3	2.1	2.0	2.5	2.0	2.1	3.2	2.9	2.5	3.0	2.5
250	1.0	1.5	1.8	1.6	0.5	1.7	1.1	1.5	1.2	1.7	2.4	2.3	2.3	2.1	1.7

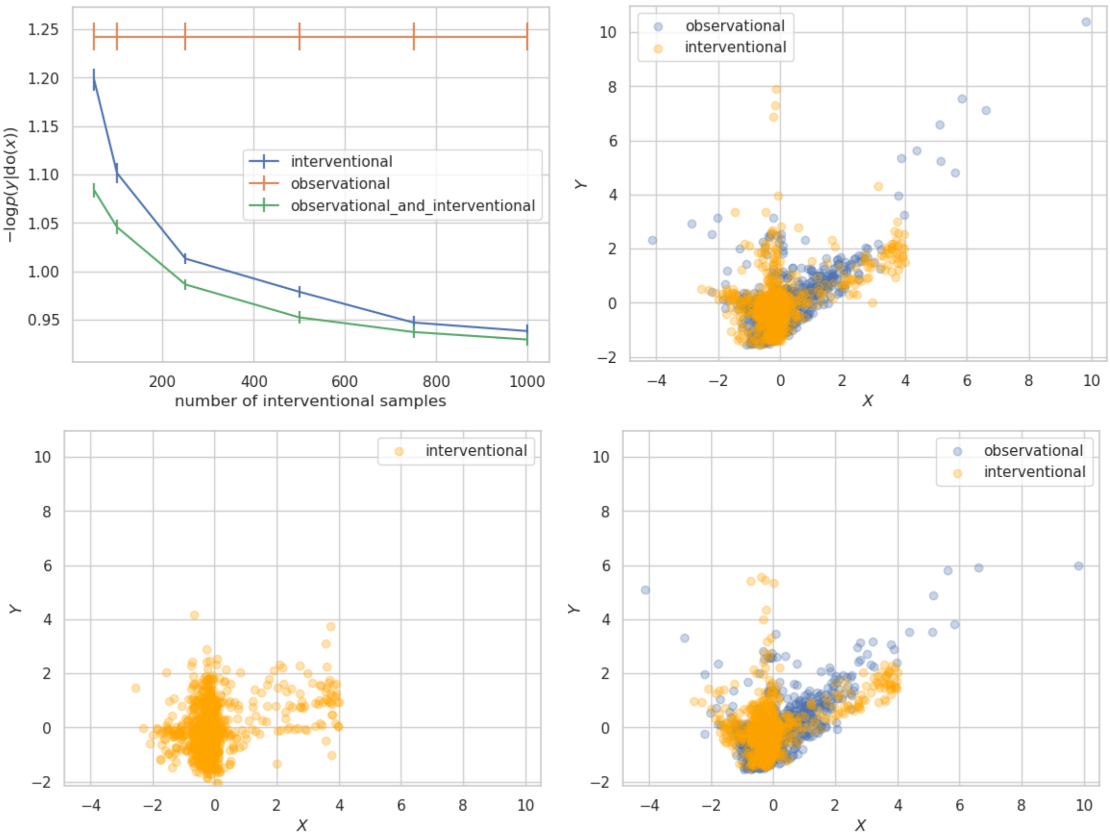

Figure 4. Results for an example dataset: (Top left) Comparison of a flow model trained with 1000 observational samples, a flow model trained with 50, 100, 250, 500, 750, 1000 interventional samples, and a flow model trained with both 1000 observational samples and 50, 100, 250, 500, 750, 1000 interventional samples. All flow models are evaluated on 1000 interventional samples from the test set. (Top right) Observational and interventional training samples. (Bottom left) Interventional samples from a flow model trained with 50 interventional samples. (Bottom right) Observational and interventional samples from a flow model trained with 50 interventional samples and 1000 observational samples.

distribution, as seen in [MPJ$^+$16]. A detailed step-by-step description of the simulation procedure is given in Appendix D.

Following the process described above, we simulate 100 datasets while varying the number of dimensions K of the unobserved confounder $\mathbf{Z}$ and the random seed that is, among others, controlling the initialization of the neural networks used to model the causal mechanisms. We choose K between 1 and 10 since for $K > 10$, the joint distribution $p(x, y)$ becomes increasingly Gaussian due to the central limit theorem. Next, we manually select ten datasets with the smallest overlap of observational and interventional samples to select cases with “strong” confounding. Note that we choose these ten datasets before training a single flow model. A scatter plot of 1000 observational and 1000 interventional samples for each of the ten datasets can be found in Appendix D.3.

In this experiment we are interested in estimating the interventional distribution $p(y|\operatorname{do}(x))$. For each dataset, we train three variants of our reduced causal model parameterized with normalizing flows. The first flow model is trained using only observational data, see Section 6.1.1. The second flow model is trained using only interventional data, see Section 6.1.2. The third flow model is trained using observational and interventional data jointly, see Section 6.1.3. For each of the ten datasets, we keep the number of observational samples constant at 1000 and use an increasing number of interventional samples 50, 100, 250, 500, 750, 1000, resulting in six experiments per dataset. For example, in the case of 50 interventional and 1000 observational samples, the first flow model is trained with 1000 observational samples, the second flow model is trained with 50 interventional samples, and the third flow model is jointly trained with 1000 observational and 50 interventional samples, see Figure 4 for an example.

Motivated by the work of [OOR$^+$18] on the realistic evaluation of semi-supervised learning algorithms, we use the same number of samples for training and validation. In every case, we use 1000 interventional samples for testing. To compare the performance of the three flow models, we calculate the negative log-likelihood averaged over the test set, $-\frac{1}{N_I}\sum_{i=1}^{N_T} \log p(y_i^T \mid \mathrm{do}(x_i^T))$, where the test set consists of N_T samples $(x_1^T, y_1^T), \ldots, (x_{N_T}^T, y_{N_T}^T)$ from $p(y \mid \mathrm{do}(x))$. To have a fair comparison, the same training procedure, architecture, optimizer, and hyperparameters are used for all flow models in all experiments. We use Adam [KB15] with a learning rate of 0.001 and the default values for β_1, β_2. We train for 10,000 epochs. The training is terminated early when the validation loss did not improve for 1,000 epochs. We perform full batch gradient descent, where we alternate between batches of observational and interventional samples for the third flow model. For the linear rational spline flows, we use 32 bins and set the bound $B = 6$. We use a fully connected neural network with two hidden layers and ReLU activations for the conditional version of the linear rational spline flows.

In Appendix D.3, we provide extensive visualizations of the results of all experiments, including scatter plots of training data, samples from the trained flow models, negative log-likelihood values for all flow models on the interventional and observational test sets. To summarize our findings, we calculate the ratio of samples required to reach the same performance, measured in averaged negative log-likelihood when only using interventional samples.

In Table 1 we see that in the case of dataset 3 and $N_I = 100$, we need two times the number of interventional samples (in the absence of observational training samples) to achieve the same performance as a flow model that is jointly trained with 100 interventional and 1000 observational samples. We can substantially reduce the number of interventional samples required when using an additional 1000 observational samples in eight of the ten datasets. Only in the case of dataset 5, we find that we need substantially more interventional samples to train our flow model jointly with observational and interventional data. We observe that in dataset 5, the interventional distribution resembles a standard Gaussian distribution that can easily be estimated from very few interventional samples. Last, the results in Table 1, dataset 1 to 10, are in agreement with qualitative results in Appendix D.3, where we find that samples from the flow model trained with interventional and observational data better resemble the training data compared to samples from a flow model trained with interventional data only.

6.2.2 With observed confounders

We now consider the case of an additional L-dimensional observed confounder $\mathbf{C}$. We use the same setup as in Section 6.2.1 to simulate triples $(x, y, \mathbf{c})$. We use the following nonlinear causal mechanisms to generate treatment X and outcome Y: $X = f(E_X, \mathbf{Z}, \mathbf{C})$ and $Y = g(X, E_Y, \mathbf{Z}, \mathbf{C})$, a detailed description of the simulation procedure is given in Appendix D. Again, we generate 100 datasets by varying K, L between 1 and 5 and the random seed. We select five datasets following the same criteria as described in Section 6.2.1. Furthermore, we use the implementation described in Section 6.1.4 to estimate the SCM in Figure 3.3. For each of the five datasets, we keep the number of observational samples constant at 1000 and use an increasing number of interventional samples: 50, 100, 250, 500, 750, 1000, resulting in six experiments per dataset. We compare three flow models trained with observational, interventional, and observational plus interventional data, respectively. The training details are the same as in Section 6.2.1. An extensive comparison of the three flow models, as well as visualizations for each dataset, can be found in Appendix D.3.11. The main result of the experiments with additional, observed confounders is the following: For each of the five datasets, we can substantially reduce the required number of interventional samples with our flow model trained with observational and interventional data, see Table 1, dataset 11 to 15. We find that we can reduce the number of required samples by a factor of two to three when training with 1000 additional observational samples.

7 Conclusion

We showed that without loss of generality, when modeling an unobserved confounder of a treatment and an outcome variable, we may assume that this confounder takes values in the same space as the treatment variable. Applying this insight to the setting of discrete treatment and outcome, we can easily derive bounds between observational and interventional distributions that can be exploited for estimation purposes. We pointed out that in certain cases with highly unbalanced observational samples, the accuracy of the causal effect estimate can be improved by incorporating observational data. Furthermore, in the linear-Gaussian setting, we derived equality constraints between the parameters of the observational and interventional distributions, showing that these distributions are not independent. Finally, for the general continuous (possibly nonlinear and non-Gaussian) setting, we proposed a flexible parameterization of the reduced causal model using normalizing flows. This parameterization allows to train a single flow model from combined observational and interventional data. In simulations, for 13 out of 15 simulated datasets, we could substantially reduce the number of interventional samples if sufficient observational samples are available without sacrificing accuracy.

Together, our results suggest that there is still untapped potential to obtaining more accurate estimates of causal effects by combining observational and interventional data, while allowing for latent confounding in the observational regime. Our work opens up practical applications and further theoretical questions regarding the precise nature of the relationship between observational and interventional distributions in parametric and non-parametric settings.

Possible future work includes (i) investigating the potential for improved causal effect estimation from combined data in clinical applications, (ii) extending the reduction operation to more than two observed variables, and (iii) applying the flow model to high-dimensional outcome variables, e.g., medical images.

References

[ACI20] Susan Athey, Raj Chetty, and Guido Imbens. Combining Experimental and Observational Data to Estimate Treatment Effects on Long Term Outcomes. *arXiv:2006.09676 [cs, stat]*, 2020.

[AIR96] Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

[BCJ$^+$19] Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep Universal Probabilistic Programming. *Journal of Machine Learning Research*, 20:1–6, 2019.

[BFPM21] Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M. Mooij. Foundations of structural causal models with cycles and latent variables. *Annals of Statistics*, 49(5):2885–2915, 2021.

[BP97] Alexander Balke and Judea Pearl. Bounds on Treatment Effects From Studies With Imperfect Compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997.

[CMC$^+$20] Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: a review. *arXiv:2011.08047*, 2020.

[DBMP19] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, 2019.

[DEL20] Hadi Mohaghegh Dolatabadi, Sarah M. Erfani, and Christopher Leckie. Invertible generative modeling using linear rational splines. In Silvia Chiappa and Roberto Calandra, editors, *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020*, 2020.

[DSB17] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

[Fis21] F. A. Fisher. Statistical methods for research workers. *Oliver and Boyd, Edinburgh, UK*, 2021.

[FK92] Richard D. Fuhr and Michael Kallay. Monotone linear rational spline interpolation. *Computer Aided Geometric Design*, 9(4):313–319, 1992.

[For21a] Patrick Forré. Quasi-Measurable Spaces. *arxiv:2109.11631*, 2021.

[For21b] Patrick Forré. Transitional Conditional Independence. *arxiv:2104.11547*, 2021.

[Gun20] Florian Gunsilius. A path-sampling method to partially identify causal effects in instrumental variable models. *arxiv:1910.09502*, 2020.

[HI05] Keisuke Hirano and Guido W. Imbens. The Propensity Score with Continuous Treatments. In *Wiley Series in Probability and Statistics*, pages 73–84. Wiley, 2005.

[KB15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

[KKS20] Niki Kilbertus, Matt J Kusner, and Ricardo Silva. A class of algorithms for general instrumental variable models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, pages 20108–20119, 2020.

[KMLH20] Ilyes Khemakhem, Ricardo Pio Monti, Robert Leech, and Aapo Hyvärinen. Causal Autoregressive Flows. In *arXiv:2011.02268*, 2020.

[KPB20] I. Kobyzev, S. Prince, and M. Brubaker. Normalizing Flows: An Introduction and Review of Current Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020.

[KPS18] Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, 2018.

[LCB20] Sanghack Lee, Juan D. Correa, and Elias Bareinboim. General identifiability with arbitrary surrogate experiments. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence (UAI-19)*, volume 115, pages 389–398. PMLR, 2020.

[LSM$^+$17] Christos Louizos, Uri Shalit, Joris M. Mooij, David A. Sontag, Richard S. Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, NeurIPS 2017*, 2017.

[MGTT18] Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

[MMC20] Joris M. Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020.

[MN98] Charles F. Manski and Daniel S. Nagin. Bounding disagreements about treatment effects. *Sociological Methodology*, 28(1):99–137, 1998.

[MPJ$^+$16] Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing Cause from Effect Using Observational Data: Methods and Benchmarks. *Journal of Machine Learning Research*, 17:1–102, 2016.

[MSB$^+$14] David Madigan, Paul E. Stang, Jesse A. Berlin, Martijn Schuemie, J. Marc Overhage, Marc A. Suchard, Bill Dumouchel, Abraham G. Hartzema, and Patrick B. Ryan. A systematic statistical approach to evaluating evidence from observational studies. *Annual Review of Statistics and Its Application*, 1(1):11–39, 2014.

[OOR$^+$18] Avital Oliver, Augustus Odena, Colin Raffel, Ekin Dogus Cubuk, and Ian J. Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, 2018.

[PCdCG20] Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020.

[Pea95] Judea Pearl. On the testability of causal models with latent and instrumental variables. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, 1995.

[Pea09] Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.

[PGM$^+$19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, 2019.

[PMP17] George Papamakarios, Iain Murray, and Theo Pavlakou. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, 2017.

[PNR$^+$21] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing Flows for Probabilistic Modeling and Inference. In *Journal of Machine Learning Research*, volume 22, pages 1–64, 2021.

[RBOB20] Evan Rosenman, Guillaume Basse, Art Owen, and Michael Baiocchi. Combining Observational and Experimental Datasets Using Shrinkage Estimators. *arXiv:2002.06708*, 2020.

[RM15] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, 2015.

[ROBB18] Evan Rosenman, Art B. Owen, Michael Baiocchi, and Hailey Banack. Propensity Score Methods for Merging Observational and Experimental Datasets. *arXiv:1804.07863*, 2018.

[Sil16] Ricardo Silva. Observational-interventional priors for dose-response learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, NeurIPS 2016*, 2016.

[TT13] E. G. Tabak and Cristina V. Turner. A Family of Nonparametric Density Estimation Algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013.

[WB19] Yixin Wang and David M. Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019.

[WSF19] Elie Wolfe, Robert W. Spekkens, and Tobias Fritz. The inflation technique for causal inference with latent variables. *Journal of Causal Inference*, 7(2), 2019.

A Proofs

Theorem A.1. *Let Y be a 'conditional' random variable with Markov kernel $P(Y|\mathbf{X})$ that takes values in $\mathbb{R}$ (or $[-\infty, \infty]$) and whose input $\mathbf{X}$ has values in any measurable space (e.g. $\mathbb{R}^M$). Then there exists a uniformly distributed variable $E \sim U[0,1]$ that is independent of $\mathbf{X}$ and a deterministic function F, namely the conditional quantile function of Y given $\mathbf{X}$, such that:*

$$Y = F(E|\mathbf{X}) \quad a.s. \tag{39}$$

Proof. Consider the interpolated conditional cumulative distribution function of Y given $\mathbf{X}$ with $u \in [0,1]$:

$$G(y; u|\mathbf{x}) := P(Y < y|\mathbf{x}) + u \cdot P(Y = y|\mathbf{x}). \tag{40}$$

Furthermore, consider the conditional quantile function (cqf) of Y given $\mathbf{X}$ with $e \in [0,1]$:

$$F(e|\mathbf{x}) := \inf\{\tilde{y} \in \mathbb{R} \,|\, G(\tilde{y}; 1|\mathbf{x}) \geq e\}. \tag{41}$$

Then take any uniformly distributed random variable $U \sim U[0,1]$ independent of $(Y, \mathbf{X})$ and define:

$$E := G(Y; U|\mathbf{X}), \tag{42}$$

where we plugged Y, U and $\mathbf{X}$ into G. Then one can check using standard arguments for cdf and cqf that E is uniformly distributed, $E \sim U[0,1]$, which is independent of the value $\mathbf{x}$ of $\mathbf{X}$. Furthermore, one can show that:

$$Y = F(E|\mathbf{X}) \quad \text{a.s.} \tag{43}$$

A detailed proof can be found in [For21b] in Appendix G. □

Theorem 3.1 (Causal Reduction)**.** *Let $\mathcal{M}$ be a causal Bayesian network with observed variables $\mathbf{X} \in \mathcal{X}, \mathbf{Y} \in \mathcal{Y}$ and latent variables, $Z_1 \in \mathcal{Z}_1, \ldots, Z_K \in \mathcal{Z}_K$ such that $\mathbf{Y}$ is not an ancestor of $\mathbf{X}$.*

Then there exists a causal Bayesian network $\mathcal{M}^*$ with observed variables $\mathbf{X} \in \mathcal{X}$ and $\mathbf{Y} \in \mathcal{Y}$ and a single latent confounder $\mathbf{W} \in \mathcal{X}$ (that takes values in the same space as $\mathbf{X}$) such that $\mathcal{M}^*$ is interventionally equivalent to $\mathcal{M}$ with respect to perfect interventions on the observed variables $\mathbf{X}$ and $\mathbf{Y}$:

$$p_{\mathcal{M}}(\mathbf{x}, \mathbf{y}) = p_{\mathcal{M}^*}(\mathbf{x}, \mathbf{y})$$

$$p_{\mathcal{M}}(\mathbf{x} \mid \mathrm{do}(\mathbf{y})) = p_{\mathcal{M}^*}(\mathbf{x} \mid \mathrm{do}(\mathbf{y}))$$

$$p_{\mathcal{M}}(\mathbf{y} \mid \mathrm{do}(\mathbf{x})) = p_{\mathcal{M}^*}(\mathbf{y} \mid \mathrm{do}(\mathbf{x})).$$

Proof of Theorem 3.1. While the main text preceding Theorem 3.1 already provides a proof, we show here explicitly that the reduction operation commutes with a perfect intervention on $\mathbf{X}$.

$$\begin{aligned}
p(\mathbf{y}|\operatorname{do}(\mathbf{x})) &= \int\limits_{\mathcal{Z}_1} \cdots \int\limits_{\mathcal{Z}_K} p(\mathbf{y}, z_1, \ldots, z_K|\operatorname{do}(\mathbf{x}))dz_1 \ldots dz_K \\
&\overset{(a)}{=} \int\limits_{\mathcal{Z}} p(\mathbf{y}, \mathbf{z}|\operatorname{do}(\mathbf{x}))d\mathbf{z} \\
&\overset{(b)}{=} \int\limits_{\mathcal{Z}} p(\mathbf{y}|\mathbf{x}, \mathbf{z})p(\mathbf{z})d\mathbf{z}, \\
&\overset{(c)}{=} \int\limits_{\mathcal{Z}} \left(\int\limits_{\mathcal{X}} p(\mathbf{y}|\mathbf{x}, \mathbf{z})p(\mathbf{z})p(\mathbf{w}|\mathbf{z})d\mathbf{w} \right) d\mathbf{z}, \\
&\overset{(d,e)}{=} \int\limits_{\mathcal{X}} \left(\int\limits_{\mathcal{Z}} p(\mathbf{y}|\mathbf{x}, \mathbf{z})p(\mathbf{w})p(\mathbf{z}|\mathbf{w})d\mathbf{z} \right) d\mathbf{w}, \\
&\overset{(8)}{=} \int\limits_{\mathcal{X}} p(\mathbf{y}|\mathbf{x}, \mathbf{w})p(\mathbf{w})d\mathbf{w}\,.
\end{aligned}$$

□

Theorem 3.2. *Let* $\mathbf{Y}$ *be a 'conditional'* $\mathbb{R}^N$*-valued random variable with Markov kernel* $P(\mathbf{Y}|\mathbf{X})$ *and input* $\mathbf{X}$ *that can take values in any measurable space. Then there exists an* N*-dimensional standard normal random variable* $\mathbf{V} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$ *independent of* $\mathbf{X}$ *and a deterministic measurable map* F *such that:*

$$\mathbf{Y} = F(\mathbf{V}, \mathbf{X}) \quad a.s. \tag{9}$$

Furthermore, the map F *is 'well-behaved', in the sense that it is composed out of (inverse) conditional cumulative distribution functions.*

Proof of Theorem 3.2. We use Theorem A.1 inductively.

1. Consider the cqf F_1 of $P(Y_1|\mathbf{X})$. Then by A.1 there is a random variable $E_1 \sim U[0,1]$ independent of $\mathbf{X}$ such that $Y_1 = F_1(E_1|\mathbf{X})$ a.s.
2. Now consider the cqf F_2 of $P(Y_2|E_1, \mathbf{X})$. Then by A.1 there is a random variable $E_2 \sim U[0,1]$ independent of E_1, $\mathbf{X}$ such that $Y_2 = F_2(E_2|E_1, \mathbf{X})$ a.s.
3. Now consider the cqf F_3 of $P(Y_3|E_2, E_1, \mathbf{X})$. Then by A.1 there is a random variable $E_3 \sim U[0,1]$ independent of E_2, E_1, $\mathbf{X}$ such that $Y_3 = F_3(E_3|E_2, E_1, \mathbf{X})$ a.s.
4. and so on .... until:
5. $Y_N = F_N(E_N|E_{N-1}, \ldots, E_1, \mathbf{X})$ a.s. with $E_N \sim U[0,1]$ independent of $E_{N-1}, \ldots, E_1$, $\mathbf{X}$.

Now we put $Z_d := \Phi^{-1}(E_d)$, where Φ is the cdf of $\mathcal{N}(0,1)$. Then $E_d = \Phi(Z_d)$ and the Z_d are $\mathcal{N}(0,1)$-distributed and $\mathbf{Z} = (Z_1, \ldots, Z_N)$ is independent of $\mathbf{X}$). So $\mathbf{Z} = (Z_1, \ldots, Z_N) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$ and independent of $\mathbf{X}$. Furthermore, we have almost surely the equations:

$$\begin{aligned}
Y_1 &= F_1(\Phi(Z_1)|\mathbf{X}), && (44) \\
Y_2 &= F_2(\Phi(Z_2)|\Phi(Z_1), \mathbf{X}), && (45) \\
&\vdots \qquad\qquad \ddots && (46) \\
Y_N &= F_M(\Phi(Z_N)|\Phi_{(Z_{N-1})}, \ldots, \Phi(Z_1), \mathbf{X}). && (47)
\end{aligned}$$

This shows the claim. □

Corollary 5.1 (Reduced linear Gaussian model)**.** *Consider a linear Gaussian SCM (or causal Bayesian network with possible latent variables) with observed variables* $\mathbf{X}$ *and* $\mathbf{Y}$ *such that* $\mathbf{Y}$ *is not ancestor of* $\mathbf{X}$*. Then this causal model is interventionally equivalent to a reduced linear Gaussian causal model with the following structural equations:*

$$\mathbf{X} = \mathbf{a} + B\mathbf{U}, \tag{26}$$

$$\mathbf{Y} = \mathbf{c} + D\mathbf{X} + E\mathbf{U} + F\mathbf{V}, \tag{27}$$

with vectors $\mathbf{a}$, $\mathbf{c}$ *and matrices* B, D, E, F*, where* B *and* F *can be chosen to be lower-triangular with non-negative diagonal entries, and where* $\mathbf{U}$ *is a standard Gaussian latent variable of the same dimension as* $\mathbf{X}$ *and where* $\mathbf{V}$ *is a standard Gaussian latent variable of the same dimension as* $\mathbf{Y}$ *that is independent of* $\mathbf{U}$*.*

Proof of Corollary 5.1. This follows the same steps as the general construction in Equations 2, 3, 4, 5, where $p(\mathbf{x}|\mathbf{w}) = \delta_w(\mathbf{x})$ reflects the identity map. In Equation 6, note that $p(\mathbf{z}|\mathbf{w})$ is linear Gaussian by the well-known conditioning formula for jointly Gaussian distributions. We then arrive at Equation 7, where it can be checked that in Equation 8 both parts, $p(\mathbf{z}|\mathbf{w})$ and $p(\mathbf{y}|\mathbf{x},\mathbf{z})$, are linear Gaussian, thus makes $p(\mathbf{y}|\mathbf{x},\mathbf{w})$ linear Gaussian. Finally, we use the reparameterization trick together with a Cholesky decomposition, as seen in Section 3.2, to turn $p(\mathbf{w})$ into a standard Gaussian $p(\mathbf{u})$, which makes $p(\mathbf{x}|\mathbf{u})$, as a composition of identity map and linear Gaussian also a linear Gaussian. Note that $p(\mathbf{y}|\mathbf{x},\mathbf{u})$ again is linear Gaussian by similar arguments. Last we use the reparameterization trick again to obtain $p(\mathbf{y}|\mathbf{x},\mathbf{u},\mathbf{v})$ where $\mathbf{V} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$. □

Theorem 5.2 (Linear Gaussian parameter constraints)**.** *Consider a linear-Gaussian SCM (or causal Bayesian network with possible latent variables) with two observed variables* $\mathbf{X}$ *and* $\mathbf{Y}$ *such that* $\mathbf{Y}$ *is not ancestor of* $\mathbf{X}$*. The entailed observational and interventional distributions are Gaussian. Modeling* $p(\mathbf{x}), p(\mathbf{y}|\mathbf{x})$ *and* $p(\mathbf{y}|\operatorname{do}(\mathbf{x}))$ *independently from each other could be done with the following parameterization:*

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\alpha}, \Sigma\right), \tag{28}$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}\left(\mathbf{y}|\gamma + \Delta\mathbf{x}, \Pi\right), \tag{29}$$

$$p(\mathbf{y}|\operatorname{do}(\mathbf{x})) = \mathcal{N}\left(\mathbf{y}|\widetilde{\gamma} + \widetilde{\Delta}\mathbf{x}, \widetilde{\Pi}\right), \tag{30}$$

with covariance matrices Σ, Π, $\widetilde{\Pi}$*. However, using the reduced causal model from Corollary 5.1 we find that these parameters are constrained by the following relations:*

$$(\widetilde{\gamma} - \gamma) + \left(\widetilde{\Delta} - \Delta\right)\boldsymbol{\alpha} = 0, \tag{31}$$

$$\left(\widetilde{\Delta} - \Delta\right)\Sigma\left(\widetilde{\Delta} - \Delta\right)^{\top} + \Pi = \widetilde{\Pi}. \tag{32}$$

Proof of Theorem 5.2. The linear version of the reduced SCM in Equation 27 entails the following distributions over $\mathbf{x}$ and $\mathbf{y}$

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}|\mathbf{a}, BB^{\top}\right), \tag{48}$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}\left(\mathbf{y}|\mathbf{c} + D\mathbf{x} + EB^{-1}(\mathbf{x} - \mathbf{a})\right., FF^{\top}), \tag{49}$$

$$p(\mathbf{y}|\operatorname{do}(\mathbf{x})) = \mathcal{N}\left(\mathbf{y}|\mathbf{c} + D\mathbf{x}, EE^{\top} + FF^{\top}\right), \tag{50}$$

Comparing Equations 28, 29, 30 with 48, 49, 50 we immediately get the equations for the parameters:

$$\boldsymbol{\alpha} = \mathbf{a}, \tag{51}$$

$$\Sigma = BB^\top, \tag{52}$$

$$\boldsymbol{\gamma} + \Delta\mathbf{x} = \mathbf{c} + \left(D + EB^{-1}\right)\mathbf{x} - EB^{-1}\mathbf{a}, \tag{53}$$

$$\boldsymbol{\gamma} \overset{\mathbf{x}=\mathbf{0}}{=} \mathbf{c} - EB^{-1}\mathbf{a}, \tag{54}$$

$$\Pi = FF^\top, \tag{55}$$

$$\widetilde{\boldsymbol{\gamma}} = \mathbf{c}, \tag{56}$$

$$\widetilde{\Delta} = D, \tag{57}$$

$$\widetilde{\Pi} = EE^\top + FF^\top. \tag{58}$$

Substituting $\mathbf{a}, \mathbf{c}, D, FF^\top$ and then subtracting Equation 54 from 53 and solving for all $\mathbf{x}$ we get the constraints:

$$\Delta = \widetilde{\Delta} + EB^{-1}, \tag{59}$$

$$\boldsymbol{\gamma} = \widetilde{\boldsymbol{\gamma}} - EB^{-1}\boldsymbol{\alpha}, \tag{60}$$

$$\widetilde{\Pi} = \Pi + EE^\top. \tag{61}$$

With Equation 59 we see that $E = (\Delta - \widetilde{\Delta})B$, which we can just plug into Equations 60 and 61. Finally using Equation 52 to replace $BB^\top$ with Σ in Equation 61 will give the claim. □

B Reduction starting from potential outcomes

Theorem 3.1 states that one can construct a reduced causal Bayesian network from a given causal Bayesian network of a certain form. Here, we show that one can also obtain such a reduced causal Bayesian network when starting instead from random variables $\xi : \Omega \to \mathcal{X}$ (the observational treatment) and the potential outcomes $\eta : \Omega \to \mathcal{Y}^{\mathcal{X}}$. Here, we consider ξ to represent the treatment in the observational regime, while η is a random function, with $\eta(\mathbf{x}) \in \mathcal{Y}$ the potential outcome in the regime under treatment $\mathbf{x} \in \mathcal{X}$. Note that we do not need a random variable corresponding to the outcome in the observational regime, since this is just $\eta(\xi)$ by the consistency assumption. We will here limit ourselves to considering a finite space $\mathcal{X}$ (that is, finitely many possible treatments) for simplicity of exposition.[4] In that case, one can also think of η as a tuple $\eta = (\eta(\mathbf{x}_1), \ldots, \eta(\mathbf{x}_n))$ for $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$.

Corollary B.1 (Causal Reduction From Potential Outcomes)**.** *Let ξ be a random variable taking values in $\mathcal{X}$ and η a random function $\mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ is a finite space and $\mathcal{Y}$ a standard measurable space. Then there exists a causal Bayesian network $\mathcal{M}^*$ with observed variables $\mathbf{X} \in \mathcal{X}$ and $\mathbf{Y} \in \mathcal{Y}$ and a single latent confounder $\mathbf{W} \in \mathcal{X}$ (that takes values in the same space as $\mathbf{X}$), with graph as in Figure 1 (b), and such that $\mathcal{M}^*$ entails the same observational and interventional distributions as encoded by ξ and η, that is,*

$$p_{\mathcal{M}^*}(\mathbf{x}, \mathbf{y}) = p_{\xi,\eta(\xi)}(\mathbf{x}, \mathbf{y})$$

$$p_{\mathcal{M}^*}(\mathbf{y} \mid \mathrm{do}(\mathbf{x})) = p_{\eta(\mathbf{x})}(\mathbf{y})$$

where $p_{\xi,\eta(\xi)}$ denotes the joint distribution of observational treatment ξ and outcome $\eta(\xi)$, and $p_{\eta(\mathbf{x})}$ the distribution of the potential outcome $\eta(\mathbf{x})$.

[4] Extending this to infinite spaces $\mathcal{X}$ is possible, but more mathematical machinery is needed in order to deal with the measure-theoretic subtleties, see for example [For21a].

Proof. We will construct a causal Bayesian network $\mathcal{M}$ with the graph depicted in Figure 1 (b), that is, with a treatment variable $\mathbf{X}$, an outcome variable $\mathbf{Y}$, and a latent confounder $\mathbf{Z}$ such that it encodes the same distributions as ξ and η, that is,

$$\begin{cases} p_{\mathcal{M}}(\mathbf{x}, \mathbf{y}) = p_{\xi, \eta(\xi)}(\mathbf{x}, \mathbf{y}) \\ p_{\mathcal{M}}(\mathbf{y} \mid \mathrm{do}(\mathbf{x})) = p_{\eta(\mathbf{x})}(\mathbf{y}). \end{cases} \tag{62}$$

We take $Z := (\xi, \eta) \in \mathcal{Z} := \mathcal{X} \times \mathcal{Y}^{\mathcal{X}}$ as the latent confounder. We can then define $p(\mathbf{x} \mid \mathbf{z}) := \delta_{\mathbf{z}_1}$, the delta measure centered on the first component of $\mathbf{z}$, and $p(\mathbf{y} \mid \mathbf{z}, \mathbf{x}) := \delta_{\mathbf{z}_2(\mathbf{x})}$, the delta measure centered on the second component of $\mathbf{z}$, evaluated in $\mathbf{x}$. It is straightforward to check that this causal Bayesian network entails Equation 62.

We now apply Theorem 3.1 to $\mathcal{M}$ to obtain a causal Bayesian network $\mathcal{M}^*$ with a single latent confounder taking values in $\mathcal{X}$ that does the job. The surrogate confounder $\mathbf{W}$ constructed in the reduced causal Bayesian network has distribution $p_{\mathcal{M}^*}(\mathbf{w}) = p_{\xi}(\mathbf{w})$, while the Markov kernels for $\mathbf{X}$ and $\mathbf{Y}$ are respectively $p_{\mathcal{M}^*}(\mathbf{x} \mid \mathbf{w}) = \delta_{\mathbf{w}}(\mathbf{x})$ and $p_{\mathcal{M}^*}(\mathbf{y} \mid \mathbf{x}, \mathbf{w}) = p_{\eta(\mathbf{x})|\xi}(\mathbf{y} \,|\, \mathbf{w})$. □

This reduction offers a more parsimonious parameterization of the observational and interventional distributions, since $p(\mathbf{w})$ and $p(\mathbf{y}|\mathbf{w}, \mathbf{x})$ together are $(|\mathcal{X}| - 1) + (|\mathcal{Y}| - 1)|\mathcal{X}|^2$ dimensional, while $p(\xi, \eta)$ is $|\mathcal{X}|\,|\mathcal{Y}|^{\mathcal{X}} - 1$ dimensional. The dimensionality can be further reduced to $(|\mathcal{X}| - 1) + 2(|\mathcal{Y}| - 1)|\mathcal{X}|$ at the cost of introducing constraints between the parameters (see Section 4).

C Background: Normalizing Flows

Normalizing flows are based on the idea of transforming samples from a simple distribution into samples from a complex distribution using the change of variable formula [RM15, TT13]:

$$p(\mathbf{x}) = p_{\mathbf{Z}}(f(\mathbf{x})) \left| \det \left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \right) \right|, \tag{63}$$

where $\mathbf{z} = f(\mathbf{x})$ is a bijective map $f : \mathcal{X} \to \mathcal{Z}$, $p_{\mathbf{Z}}(\mathbf{z})$ a simple prior distribution, and $\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}$ the Jacobian with respect to $\mathbf{x}$. The transformation $f(\mathbf{x})$ is commonly composed of K transformations $f(\mathbf{x}) = f_K \circ \cdots \circ f_1(\mathbf{x})$ to increase the overall expressivity of $f(\mathbf{x})$. The choice of $f(\mathbf{x})$ is restricted by the computational complexity of calculating the Jacobian $\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}$. In recent years, a multitude of transformations with easy to compute Jacobians have been developed, for an overview see [KPB20, PNR$^+$21].

In this paper we will use neural spline flows [DBMP19, DEL20]. Neural spline flows have two major advantages: 1. A better functional flexibility than affine transformations ($\mathbf{y} = \mathbf{s}\mathbf{x} + \mathbf{t}$), 2. A numerically stable, analytic inverse that has the same computational and space complexities as the forward operation. While [DBMP19] use quadratic, cubic, and rational quadratic functions whose inversion is done after solving polynomial equations, [DEL20] show that piecewise linear rational splines can perform competitively with these methods without requiring a polynomial equation to be solved in the inversion. Because of its reduced computational cost, we will use linear rational splines throughout this paper.

Consider a set of monotonically increasing points $\{(x^{(k)}, y^{(k)})\}_{k=0}^{K}$ called knots and a set of derivatives at each of the points $\{d^{(k)}\}_{k=0}^{K}$. For each bin $[x^{(k)}, x^{(k+1)}]$ we want to find a linear rational function of the form $\frac{ax+b}{cx+d}$ that fit the given points and derivatives.

The values returned by Algorithm 1 are subsequentely used to express the following linear rational spline function

$$f(\boldsymbol{\phi}) = \begin{cases} \frac{w^{(k)} y^{(k)} (\lambda^{(k)} - \boldsymbol{\phi}) + w^{(m)} y^{(m)} \boldsymbol{\phi}}{w^{(k)} (\lambda^{(k)} - \boldsymbol{\phi}) + w^{(m)} \boldsymbol{\phi}} & 0 \leq \boldsymbol{\phi} \leq \lambda^{(k)} \\ \frac{w^{(m)} y^{(m)} (1 - \boldsymbol{\phi}) + w^{(k+1)} y^{(k+1)} (\boldsymbol{\phi} - \lambda^{(k)})}{w^{(m)} (1 - \boldsymbol{\phi}) + w^{(k+1)} (\boldsymbol{\phi} - \lambda^{(k)})} & \lambda^{(k)} \leq \boldsymbol{\phi} \leq 1 \end{cases} \tag{64}$$

where $\boldsymbol{\phi} = (x - x^{(k)})/(x^{(k+1)} - x^{(k)})$.

Algorithm 1 [FK92] Linear Rational Spline Interpolation of Monotonic data in the interval $\left[x^{(k)}, x^{(k+1)}\right]$.

```
Input: $x^{(k)} < x^{(k+1)}$, $y^{(k)} < y^{(k+1)}$, $d^{(k)} > 0$, $d^{(k+1)} > 0$
set $w^{(k)} > 0$
set $0 < \lambda^{(k)} < 1$
$w^{(k)} = \sqrt{\frac{d^{(k)}}{d^{(k+1)}}} w^{(k)}$
$y^m = \frac{w^{(k)} y^{(k)} \left(1-\lambda^{(k)}\right) + w^{(k+1)} y^{(k+1)} \lambda^{(k)}}{w^{(k)} \left(1-\lambda^{(k)}\right) + w^{(k+1)} \lambda^{(k)}}$
$w^{(m)} = \left(\lambda^{(k)} w^{(k)} d^{(k)} + \left(1 - \lambda^{(k)}\right) w^{(k+1)} d^{(k+1)}\right) \frac{x^{(k+1)} - x^{(k)}}{y^{(k+1)} - y^{(k)}}$
Return: $\lambda^{(k)}, w^{(k)}, w^{(m)}, w^{(k+1)}, y^{(m)}$
```

Spline flows have two hyperparameters, the boundary B of the interval $[-B, B]$ and the number of bins K. Outside of the interval $[-B, B]$, the identity function is used. Using Equation 63 we can update the parameters of the neural spline flow using maximum-likelihood estimation in combination with gradient descent. In the case where $\mathbf{x}$ has two or more dimensions, either coupling layers [DSB17] or autoregressive layers [PMP17] can be used.

At multiple points in this paper we are required to estimate conditional distributions, e.g. $p(\mathbf{y}|\mathbf{x})$, where we will use conditional normalizing flows to estimate conditional probabilities. We consider the mapping $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$, which is bijective in $\mathcal{Y}$ and $\mathcal{Z}$, and a simple prior distribution $p_{\mathbf{Z}}(\mathbf{z})$. Again, using the change of variable formula we can express the conditional distributions $p(\mathbf{y}|\mathbf{x})$ as follows

$$p(\mathbf{y}|\mathbf{x}) = p_{\mathbf{Z}}(f_x(\mathbf{y})) \left| \det \left(\frac{\partial f_x(\mathbf{y})}{\partial \mathbf{y}} \right) \right| . \tag{65}$$

The conditional version of the linear rational spline transformation uses a neural network to predict the derivatives $\mathbf{d}$, width $\mathbf{w}$, height $\mathbf{h}$, and $\boldsymbol{\lambda}$ from $\mathbf{x}$: $\mathbf{w}, \mathbf{h}, \mathbf{d}, \boldsymbol{\lambda} = NN_{\boldsymbol{\theta}}(\mathbf{x})$.

D Simulation details

The generation of observational and interventional samples follows [MPJ$^+$16]. Instead of using Gaussian processes to model the causal mechanisms, we use two randomly initialized neural networks, NN_1 and NN_2.

D.1 Nonlinear experiments without observed confounders

D.1.1 Sampling from a random distribution

We use the following steps to generate samples from a random distribution

1. $X \sim \mathcal{N}(0, 1)$
2. sort X in ascending order $= \overrightarrow{X}$
3. sample from Gaussian Process: $F \sim \mathcal{N}(0, K_{\boldsymbol{\theta}}(\overrightarrow{X}) + \sigma^2 I)$, where for the kernel $K_{\boldsymbol{\theta}}$ we use the squared exponential covariance function with automatic relevance determination kernel
4. use the trapezoidal rule to calculate the cumulative integral of $\exp(F)$, we obtain a vector G where each element G_i corresponds to $G_i = \int_{\overrightarrow{X}_1}^{\overrightarrow{X}_i} \exp(F(x)) dx$

We will denote this whole sampling procedure by $G \sim \mathcal{RD}(\boldsymbol{\theta}, \sigma)$, where we sample $\boldsymbol{\theta}$ from a Gamma distribution $\Gamma(a, b)$ and set $\sigma = 0.0001$.

D.1.2 Generate observational and interventional data

1. Sample from latent variables

$$\begin{aligned} \boldsymbol{\theta}_{E_X} &\sim \Gamma(a_{E_X}, b_{E_X}), & (66)\\ \boldsymbol{\theta}_{E_Y} &\sim \Gamma(a_{E_Y}, b_{E_Y}), & (67)\\ \boldsymbol{\theta}_{\mathbf{Z}} &\sim \Gamma(a_{\mathbf{Z}}, b_{\mathbf{Z}}), & (68)\\ E_X &\sim \mathcal{RD}(\boldsymbol{\theta}_{E_X}, \sigma), & (69)\\ E_Y &\sim \mathcal{RD}(\boldsymbol{\theta}_{E_Y}, \sigma), & (70)\\ \mathbf{Z} &\sim \mathcal{RD}(\boldsymbol{\theta}_{\mathbf{Z}}, \sigma). & (71) \end{aligned}$$

2. Generate $X_{\text{observational}}$

$$X_{\text{observational}} = NN_1(E_X, \mathbf{Z}). \tag{72}$$

3. Normalize $X_{\text{observational}}$

$$X_{\text{observational}} = \frac{X_{\text{observational}} - \mathbb{E}\left[X_{\text{observational}}\right]}{\sqrt{\mathbb{V}[X_{\text{observational}}]}}. \tag{73}$$

4. Generate $Y_{\text{observational}}$

$$Y_{\text{observational}} = NN_2(X_{\text{observational}}, E_Y, \mathbf{Z}). \tag{74}$$

5. Sample from latent variables

$$\begin{aligned} E_Y &\sim \mathcal{RD}(\boldsymbol{\theta}_{E_Y}, \sigma) & (75)\\ \mathbf{Z} &\sim \mathcal{RD}(\boldsymbol{\theta}_{\mathbf{Z}}, \sigma) & (76) \end{aligned}$$

6. Generate $X_{\text{interventional}}$

$$\begin{aligned} \boldsymbol{\theta}_X &\sim \Gamma(a_X, b_X), & (77)\\ X_{\text{interventional}} &\sim \mathcal{RD}(\boldsymbol{\theta}_X, \sigma). & (78) \end{aligned}$$

7. Normalize $X_{\text{interventional}}$

$$X_{\text{interventional}} = \frac{X_{\text{interventional}} - \mathbb{E}\left[X_{\text{interventional}}\right]}{\sqrt{\mathbb{V}[X_{\text{interventional}}]}}. \tag{79}$$

8. Generate $Y_{\text{interventional}}$

$$Y_{\text{inter}} = NN_2(X_{\text{inter}}, E_Y, \mathbf{Z}). \tag{80}$$

9. Generate noise

$$\begin{aligned} \epsilon_{x,\text{observational}} &\sim \mathcal{N}(0,1), & (81)\\ \epsilon_{x,\text{interventional}} &\sim \mathcal{N}(0,1), & (82)\\ \boldsymbol{\theta}_{\epsilon_x} &\sim \Gamma(a_{\epsilon_x}, b_{\epsilon_x}), & (83)\\ \epsilon_{y,\text{observational}} &\sim \mathcal{N}(0,1), & (84)\\ \epsilon_{y,\text{interventional}} &\sim \mathcal{N}(0,1), & (85)\\ \boldsymbol{\theta}_{\epsilon_y} &\sim \Gamma(a_{\epsilon_y}, b_{\epsilon_y}). & (86) \end{aligned}$$

10. Add noise

$$\begin{aligned} X'_{\text{observational}} &= X_{\text{observational}} + \boldsymbol{\theta}_{\epsilon_x}\epsilon_{x,\text{observational}}, & (87)\\ X'_{\text{interventional}} &= X_{\text{interventional}} + \boldsymbol{\theta}_{\epsilon_x}\epsilon_{x,\text{interventional}}, & (88)\\ Y'_{\text{observational}} &= Y_{\text{observational}} + \boldsymbol{\theta}_{\epsilon_y}\epsilon_{y,\text{observational}}, & (89)\\ Y'_{\text{interventional}} &= Y_{\text{interventional}} + \boldsymbol{\theta}_{\epsilon_y}\epsilon_{y,\text{interventional}}. & (90) \end{aligned}$$

11. Normalize Y jointly

$$Y' = [Y'_{\text{observational}}, Y'_{\text{interventional}}], \tag{91}$$

$$Y'_{\text{observational}} = \frac{Y'_{\text{observational}} - \mathbb{E}\left[Y'\right]}{\sqrt{\mathbb{V}[Y']}}, \tag{92}$$

$$Y'_{\text{interventional}} = \frac{Y'_{\text{interventional}} - \mathbb{E}\left[Y'\right]}{\sqrt{\mathbb{V}[Y']}}. \tag{93}$$

The two neural networks NN_1 and NN_2 are Multi-layer perceptrons with a single hidden layer. The hidden layer contains 1024 units. The input layer and the hidden layer use a ReLU activation function. The weights and biases for both neural networks are uniformly sampled from the interval $[-1, 1]$. We choose the other simulation parameters as follows: $a_{E_X} = a_{E_Y} = a_Z = a_X = 5$, $a_{\epsilon_x} = a_{\epsilon_y} = 2$, $b_{E_X} = b_{E_Y} = b_Z = b_X = b_{\epsilon_x} = b_{\epsilon_y} = 0.1$, $\sigma = 0.0001$

D.2 Simulation details: Nonlinear experiments with observed confounders

In order to simulate data with additional observed confounders, we first generate $\mathbf{C}$

$$\boldsymbol{\theta}_C = \Gamma(a_C, b_C), \tag{94}$$

$$\mathbf{C} \sim \mathcal{RD}(\boldsymbol{\theta}_C, \sigma), \tag{95}$$

where $a_C = 10$ and $b_C = 1$. In addition, we modify steps 2,4 and 8 as follows

$$X_{\text{observational}} = NN_1(E_X, \mathbf{Z}, \mathbf{C}), \tag{96}$$

$$Y_{\text{observational}} = NN_2(X_{\text{observational}}, E_Y, \mathbf{Z}, \mathbf{C}), \tag{97}$$

$$Y_{\text{inter}} = NN_2(X_{\text{inter}}, E_Y, \mathbf{Z}, \mathbf{C}). \tag{98}$$

D.3 Nonlinear experiment results without observed confounders

In the following, for each of the 15 datasets we present the original training data (top), as well as interventional samples from a flow model trained with 50 interventional samples (center), and samples from a flow model trained with 50 interventional samples and 1000 observational samples (bottom). The samples are generated as described in Section 6.1.4.

In addition we show performances measured in terms of negative log-likelihood on the observational and the interventional test sets, respectively. We compare a flow model trained with 1000 observational samples, a flow model trained with 50, 100, 250, 500, 750, 1000 interventional samples, and a flow model trained with both 1000 observational samples and 50, 100, 250, 500, 750, 1000 interventional samples. All flow models are evaluated on 1000 interventional samples from the test set. Last, we compare a flow model trained with 1000 observational samples, a flow model trained with 50, 100, 250, 500, 750, 1000 interventional samples, and a flow model trained with both 1000 observational samples and 50, 100, 250, 500, 750, 1000 interventional samples. All flow models are evaluated on 1000 observational samples from the test set. We report the mean and standard error for ten runs of each experiment.

D.3.1 Dataset 1: # of confounders = 1, random seed = 6

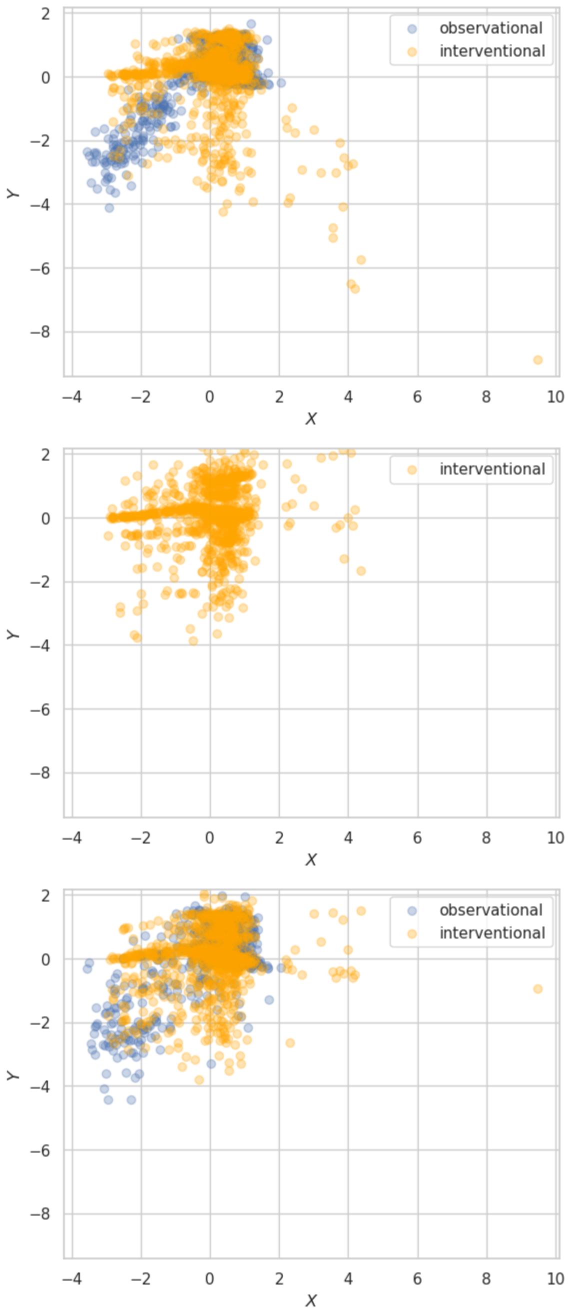

Figure 5. Dataset 1: Interventional and observational samples. (Top) training data. (Center) samples from flow model trained with only interventional data. (Bottom) samples from flow model trained with observational and interventional data.

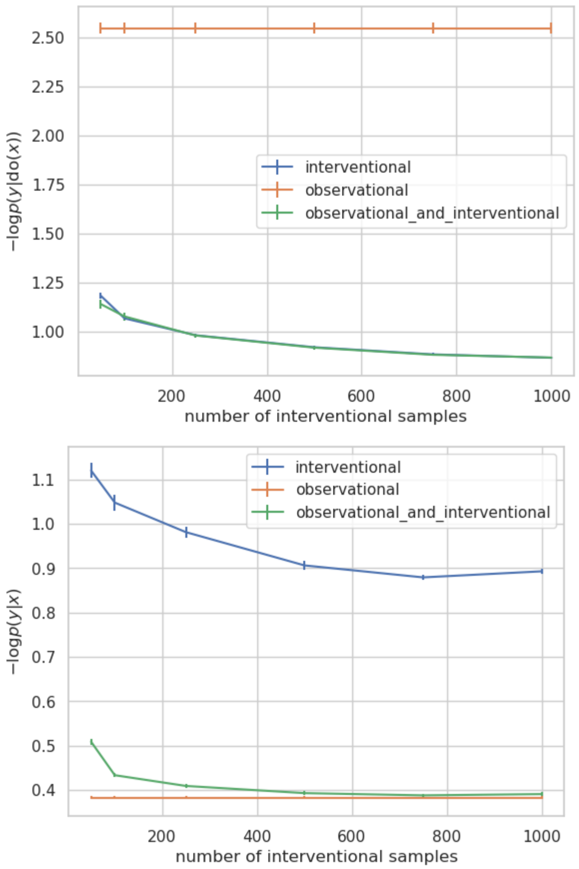

Figure 6. Dataset 1: Performances measured in terms of negative log-likelihood on the interventional (top) and the observational (bottom) test sets.

D.3.2 Dataset 2: # of confounders = 1, random seed = 8

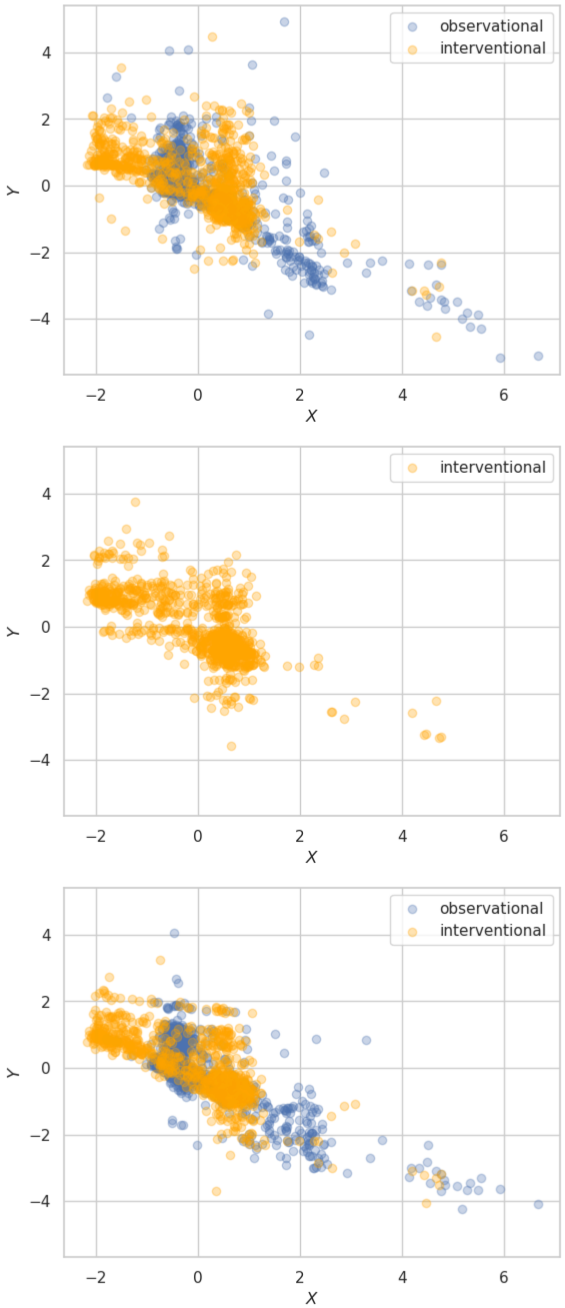

Figure 7. Dataset 2: Interventional and observational samples. (Top) training data. (Center) samples from flow model trained with only interventional data. (Bottom) samples from flow model trained with observational and interventional data.

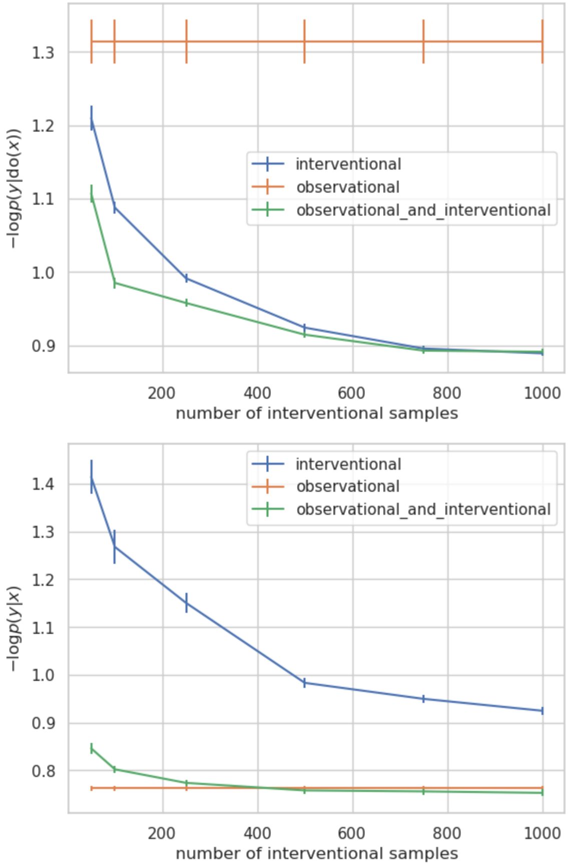

Figure 8. Dataset 2: Performances measured in terms of negative log-likelihood on the interventional (top) and the observational (bottom) test sets.

D.3.3 Dataset 3: $\#$ of confounders $= 2$, random seed $= 7$

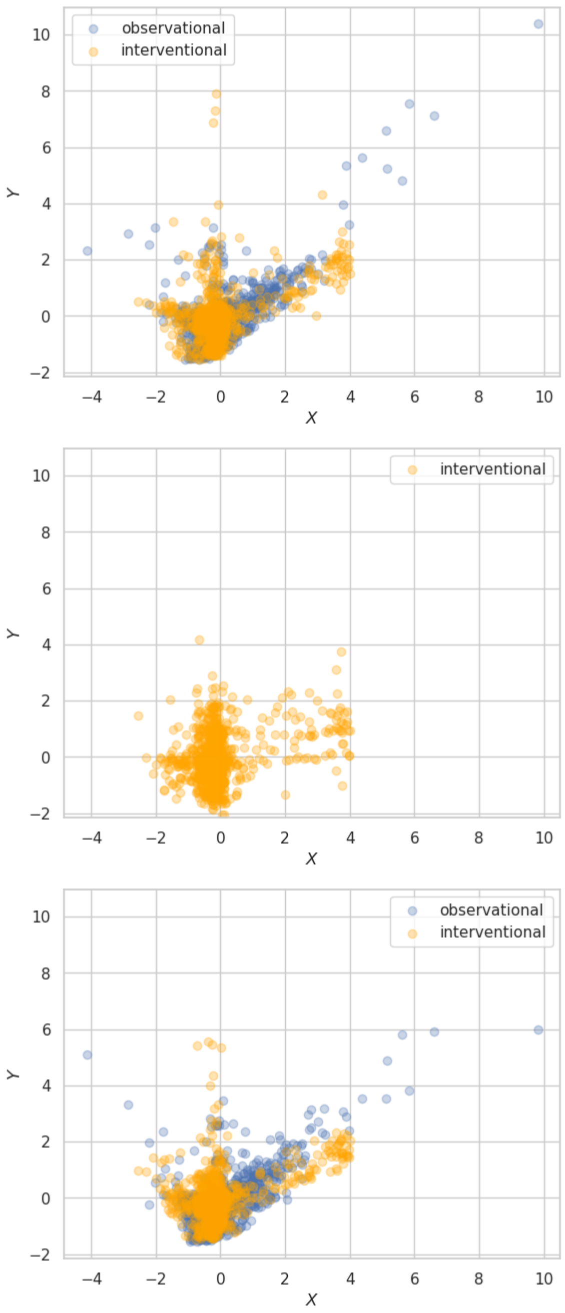

Figure 9. Dataset 3: Interventional and observational samples. (Top) training data. (Center) samples from flow model trained with only interventional data. (Bottom) samples from flow model trained with observational and interventional data.

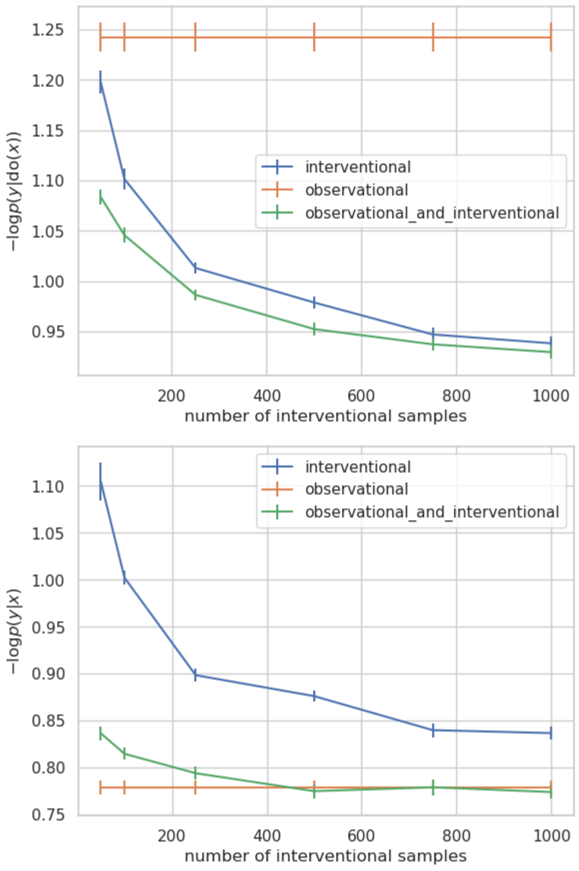

Figure 10. Dataset 3: Performances measured in terms of negative log-likelihood on the interventional (top) and the observational (bottom) test sets.

D.3.4 Dataset 4: 3 confounders, random seed = 1

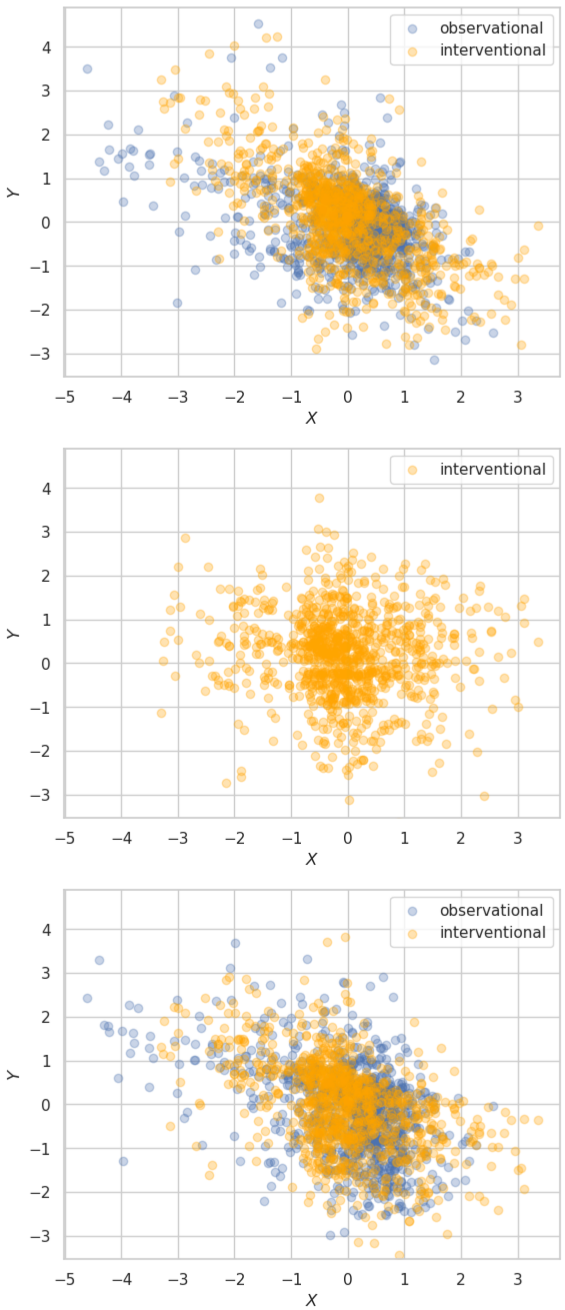

Figure 11. Dataset 4: Interventional and observational samples. (Top) training data. (Center) samples from flow model trained with only interventional data. (Bottom) samples from flow model trained with observational and interventional data.

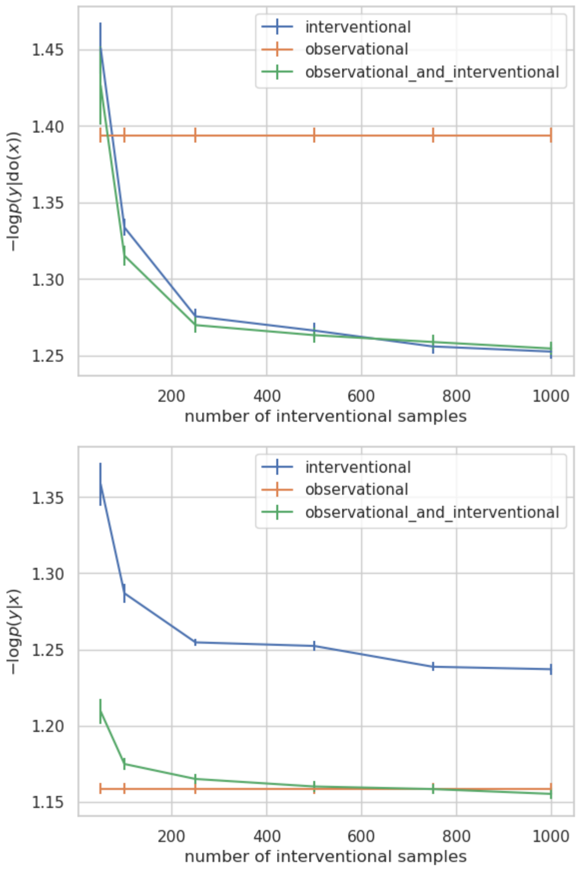

Figure 12. Dataset 4: Performances measured in terms of negative log-likelihood on the interventional (top) and the observational (bottom) test sets.

D.3.5 Dataset 5: # of confounders = 4, random seed = 0

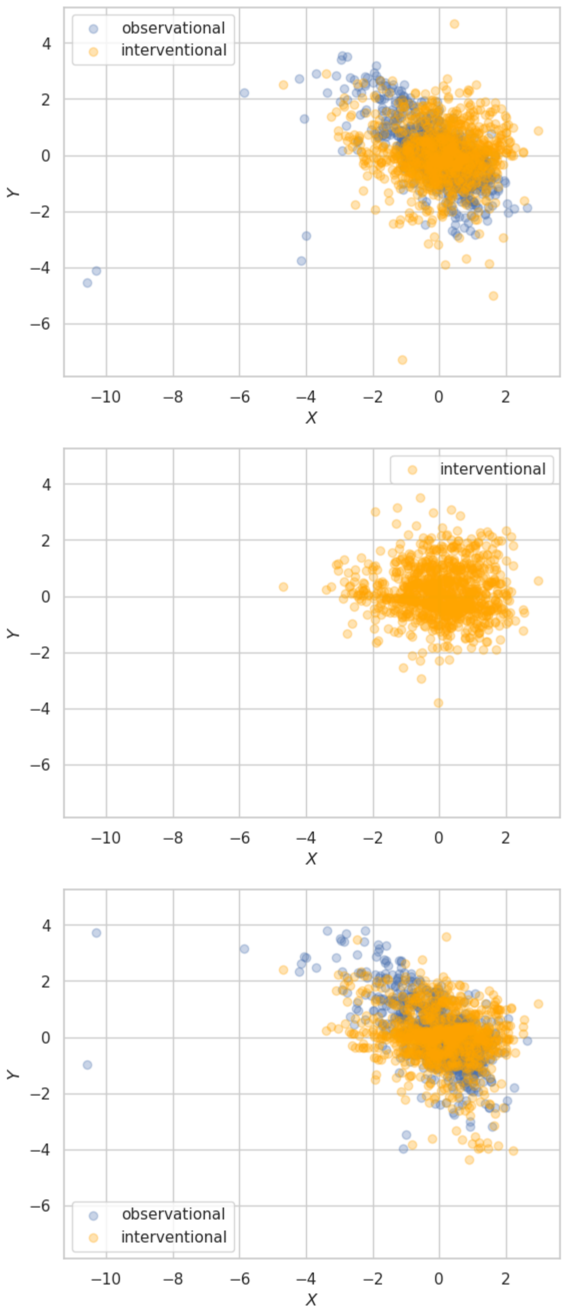

Figure 13. Dataset 5: Interventional and observational samples. (Top) training data. (Center) samples from flow model trained with only interventional data. (Bottom) samples from flow model trained with observational and interventional data.

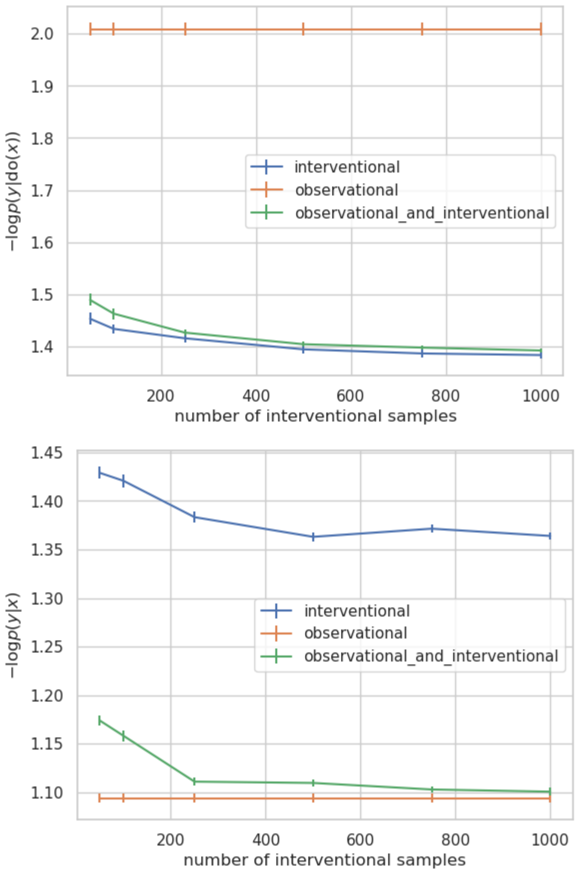

Figure 14. Dataset 5: Performances measured in terms of negative log-likelihood on the interventional (top) and the observational (bottom) test sets.

D.3.6 Dataset 6: # of confounders = 4, random seed = 7

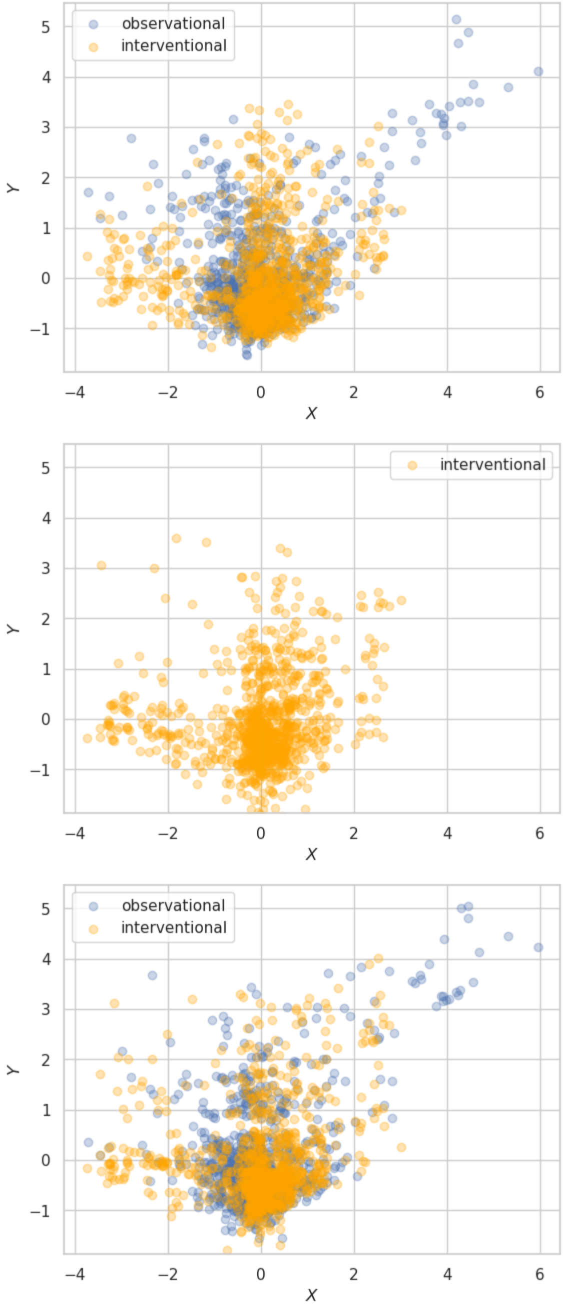

Figure 15. Dataset 6: Interventional and observational samples. (Top) training data. (Center) samples from flow model trained with only interventional data. (Bottom) samples from flow model trained with observational and interventional data.

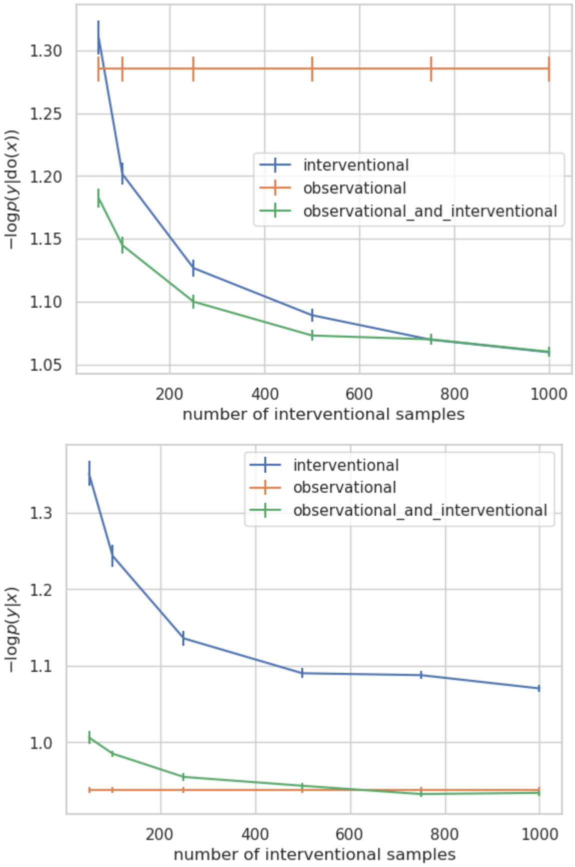

Figure 16. Dataset 6: Performances measured in terms of negative log-likelihood on the interventional (top) and the observational (bottom) test sets.

D.3.7 Dataset 7: # of confounders = 5, random seed = 5

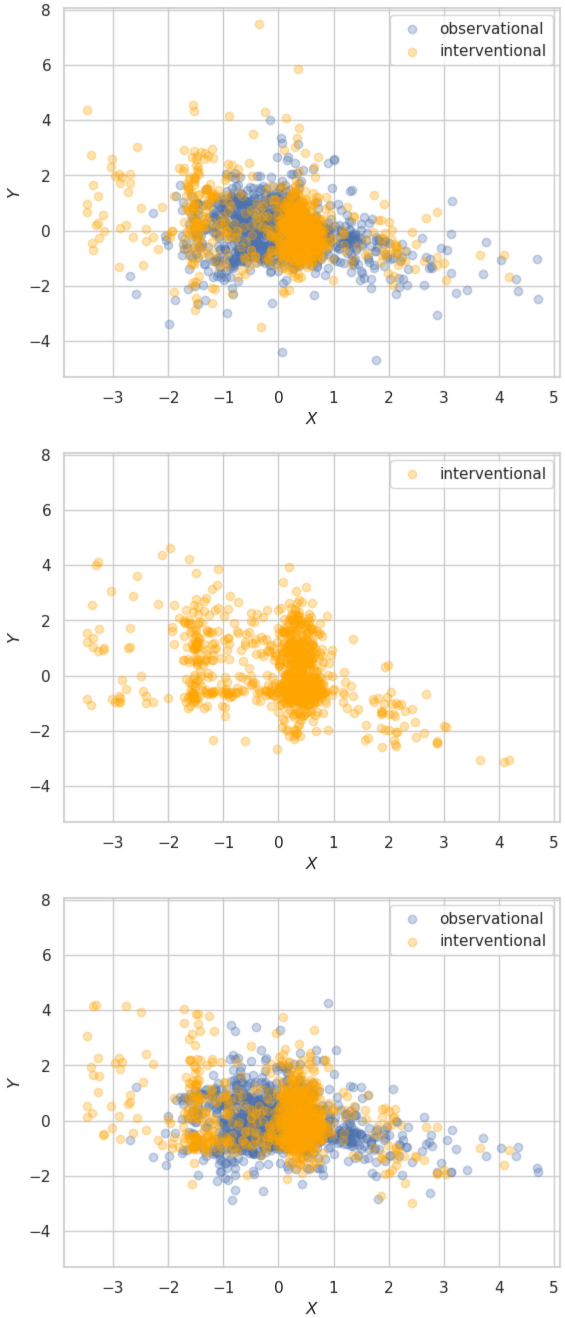

Figure 17. Dataset 7: Interventional and observational samples. (Top) training data. (Center) samples from flow model trained with only interventional data. (Bottom) samples from flow model trained with observational and interventional data.

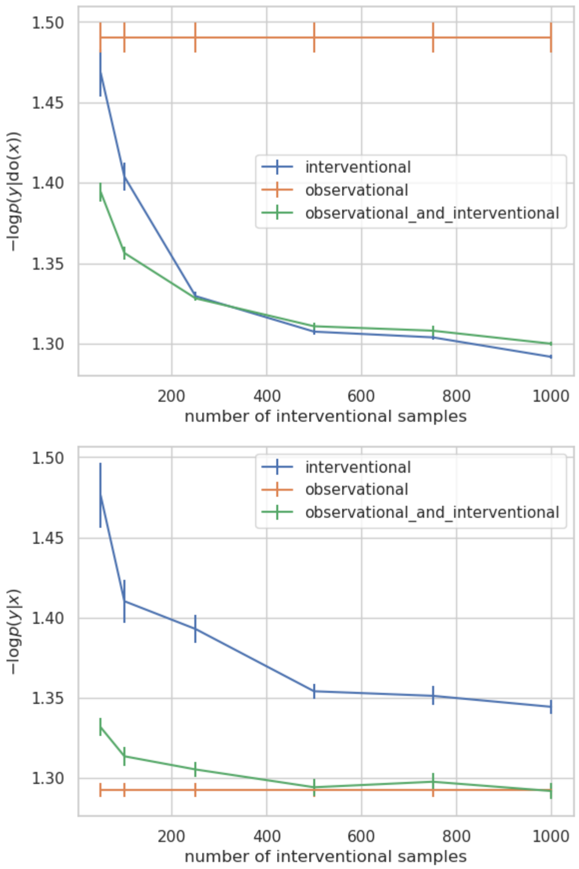

Figure 18. Dataset 7: Performances measured in terms of negative log-likelihood on the interventional (top) and the observational (bottom) test sets.

D.3.8 Dataset 8: # of confounders = 5, random seed = 9

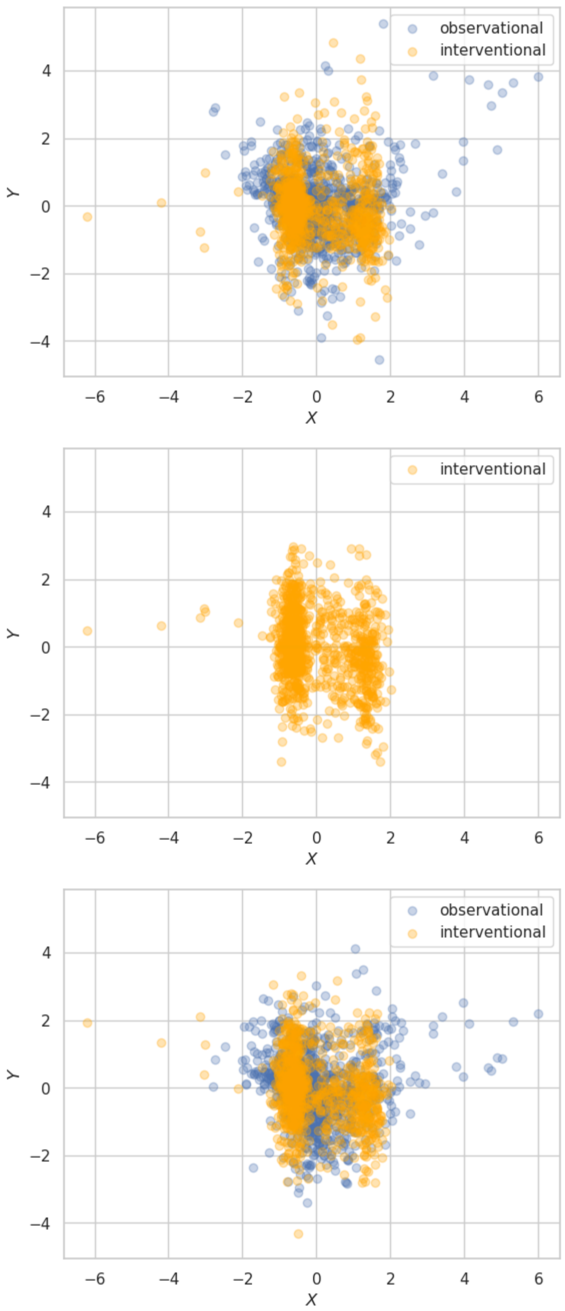

Figure 19. Dataset 8: Interventional and observational samples. (Top) training data. (Center) samples from flow model trained with only interventional data. (Bottom) samples from flow model trained with observational and interventional data.

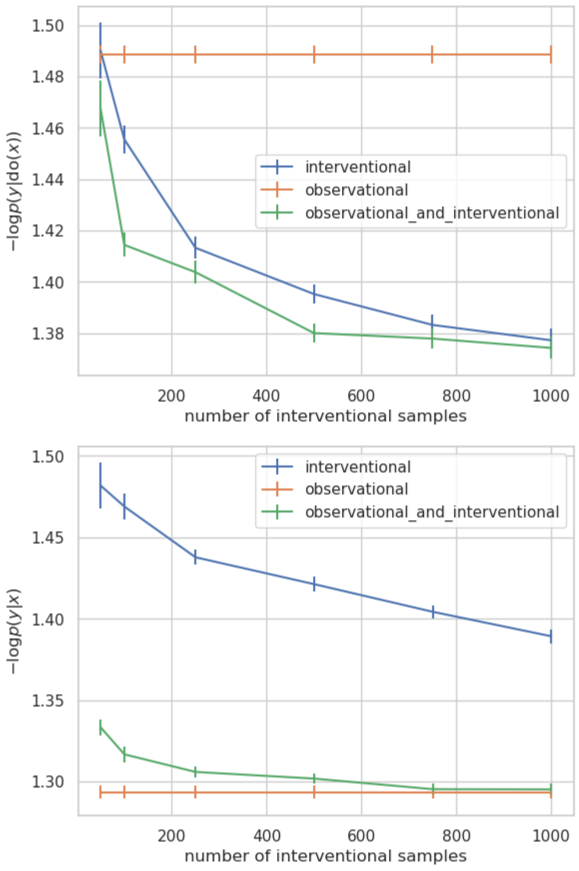

Figure 20. Dataset 8: Performances measured in terms of negative log-likelihood on the interventional (top) and the observational (bottom) test sets.

D.3.9 Dataset 9: # of confounders = 7, random seed = 0

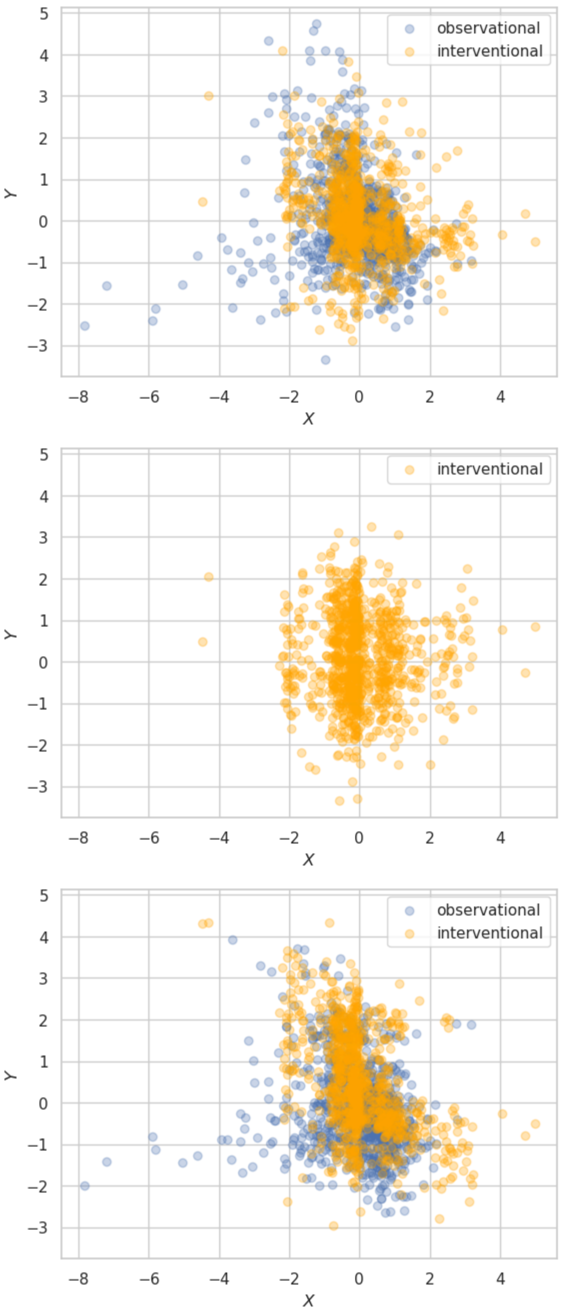

Figure 21. Dataset 9: Interventional and observational samples. (Top) training data. (Center) samples from flow model trained with only interventional data. (Bottom) samples from flow model trained with observational and interventional data.

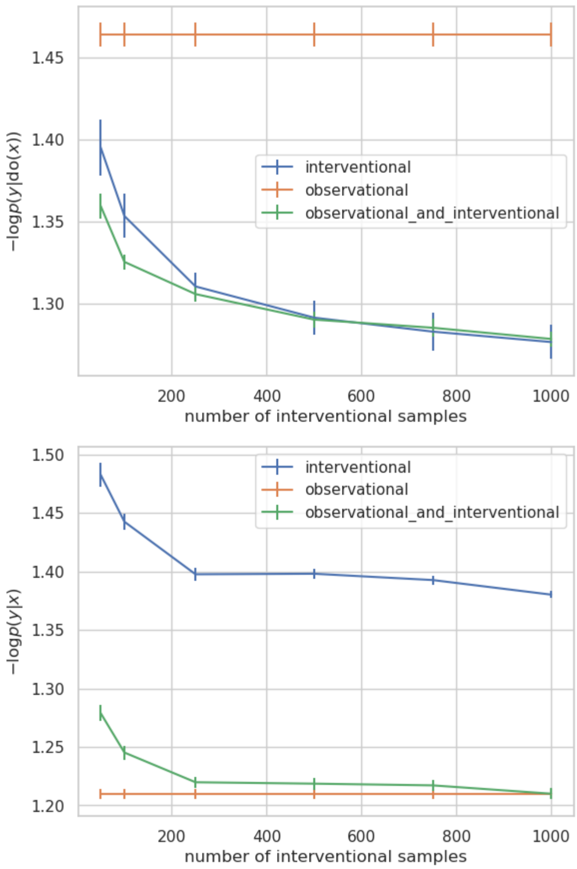

Figure 22. Dataset 9: Performances measured in terms of negative log-likelihood on the interventional (top) and the observational (bottom) test sets.

D.3.10 Dataset 10: # of confounders = 7, random seed = 5

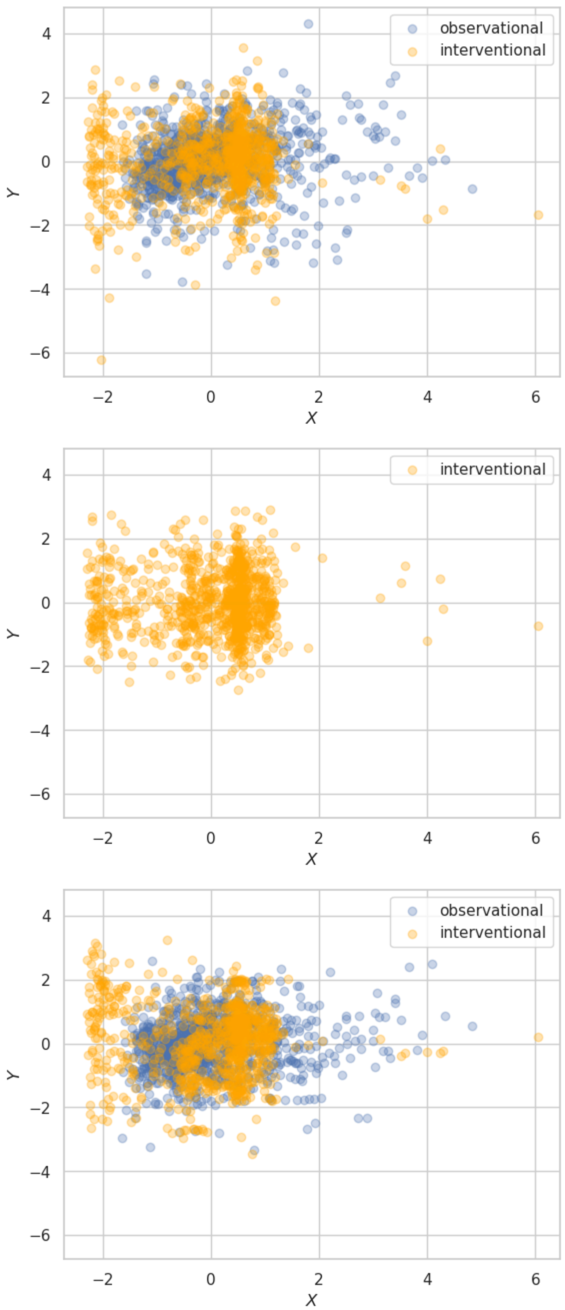

Figure 23. Dataset 10: Interventional and observational samples. (Top) training data. (Center) samples from flow model trained with only interventional data. (Bottom) samples from flow model trained with observational and interventional data.

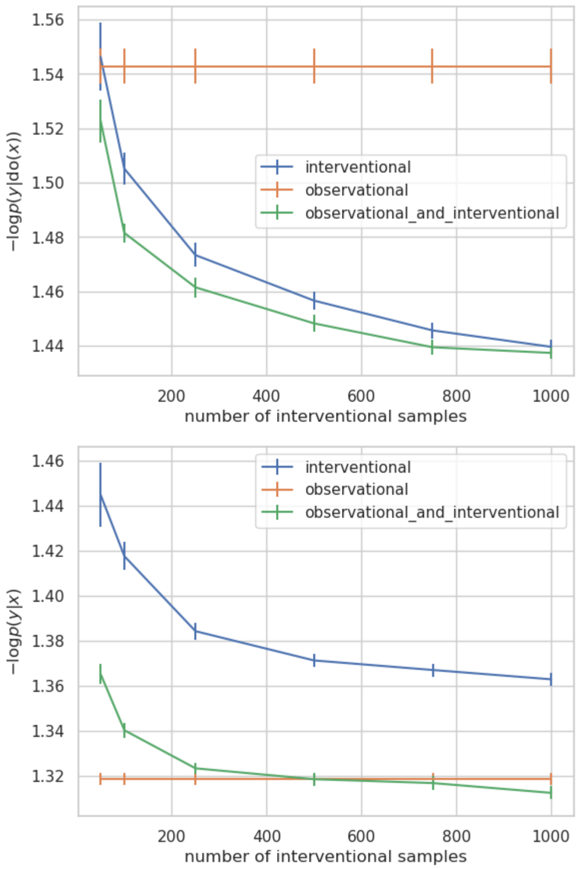

Figure 24. Dataset 10: Performances measured in terms of negative log-likelihood on the interventional (top) and the observational (bottom) test sets.

D.3.11 Dataset 11: # of latent confounders = 1, # of observed confounders = 3, random seed = 7

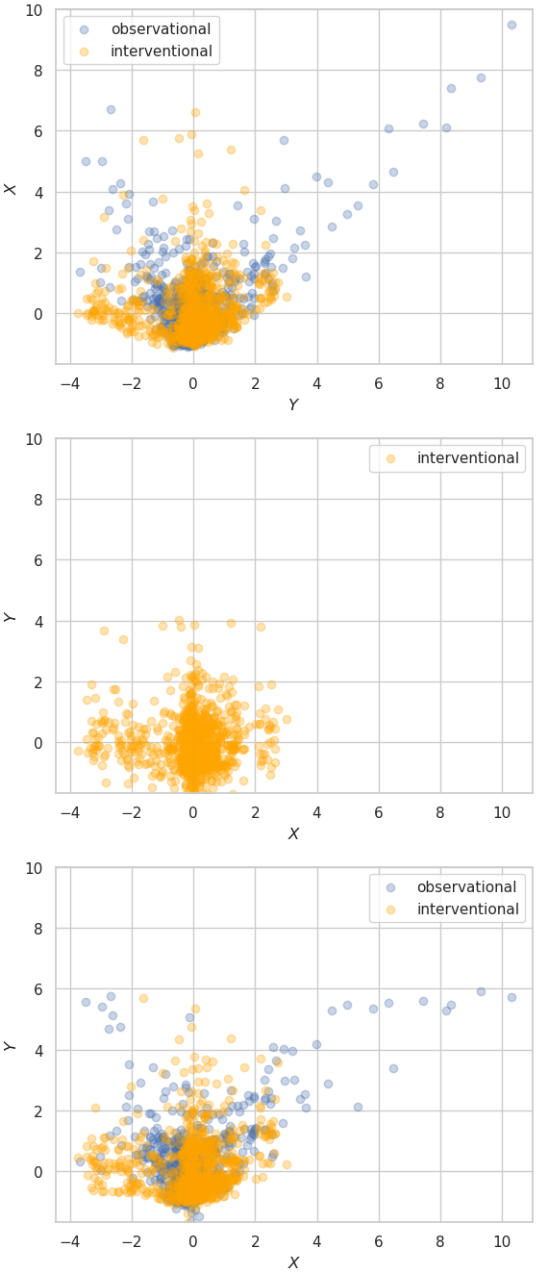

Figure 25. Dataset 11: Interventional and observational samples. Top: Observational and interventional training samples. Center: Interventional samples from a model trained with 50 interventional samples. Bottom: Observational and interventional samples from a model trained with 50 interventional samples and 1000 observational samples. The samples are generated as described in Section 6.1.4.

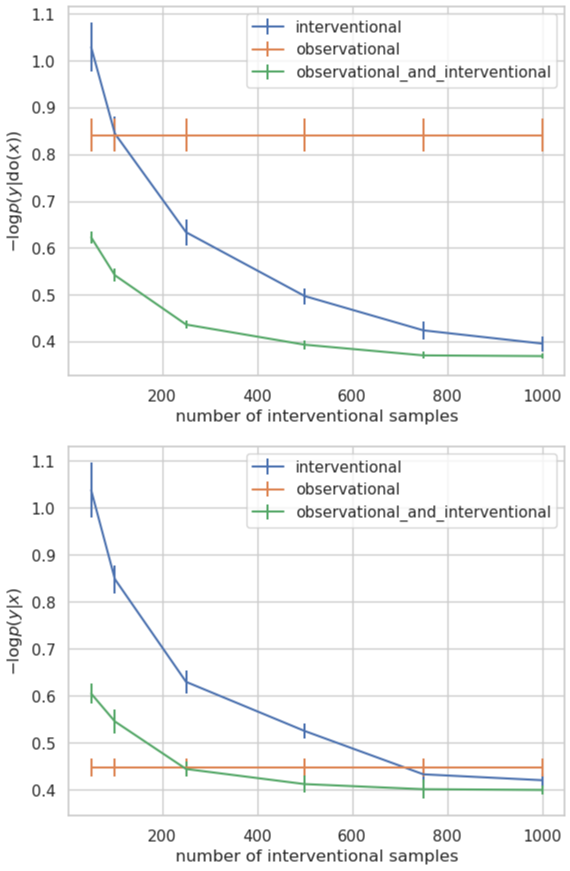

Figure 26. Dataset 11: Performances measured in terms of negative log-likelihood on the interventional (top) and the observational (bottom) test sets.

D.3.12 Dataset 12: # of latent confounders = 1, # of observed confounders = 3, random seed = 9

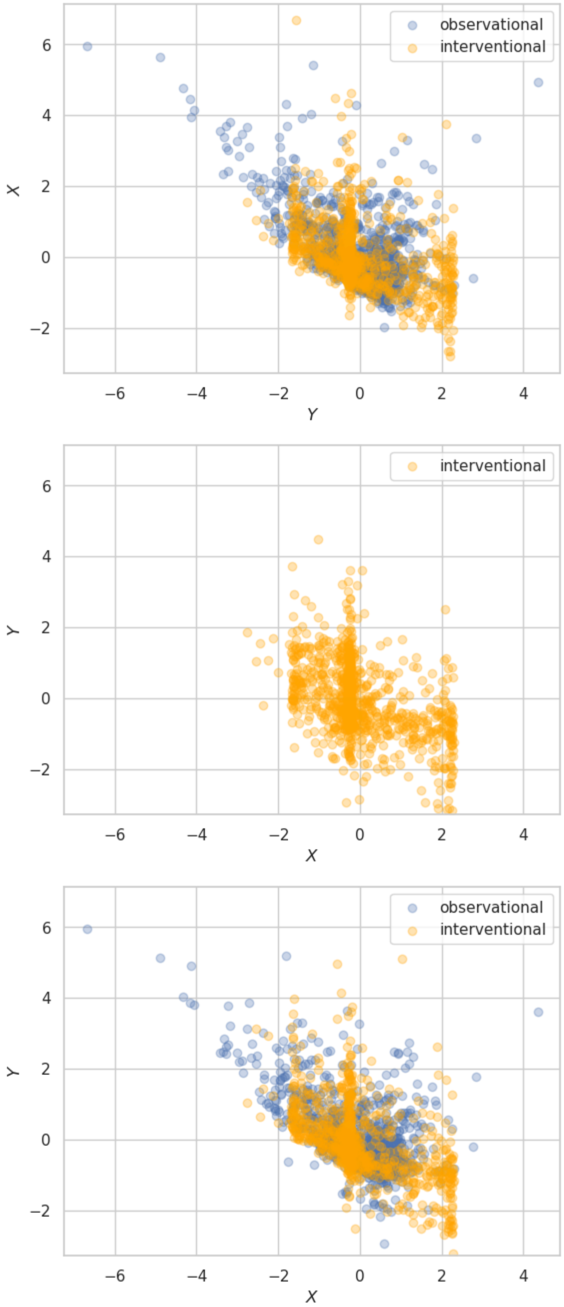

Figure 27. Dataset 12: Interventional and observational samples. (Top) training data. (Center) samples from flow model trained with only interventional data. (Bottom) samples from flow model trained with observational and interventional data.

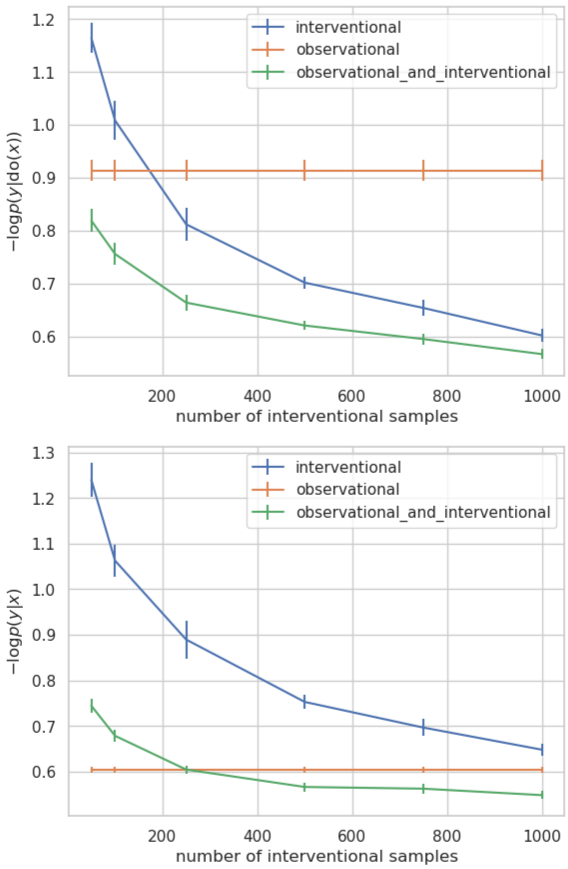

Figure 28. Dataset 12: Performances measured in terms of negative log-likelihood on the interventional (top) and the observational (bottom) test sets.

D.3.13 Dataset 13: # of latent confounders = 2, # of observed confounders = 1, random seed = 0

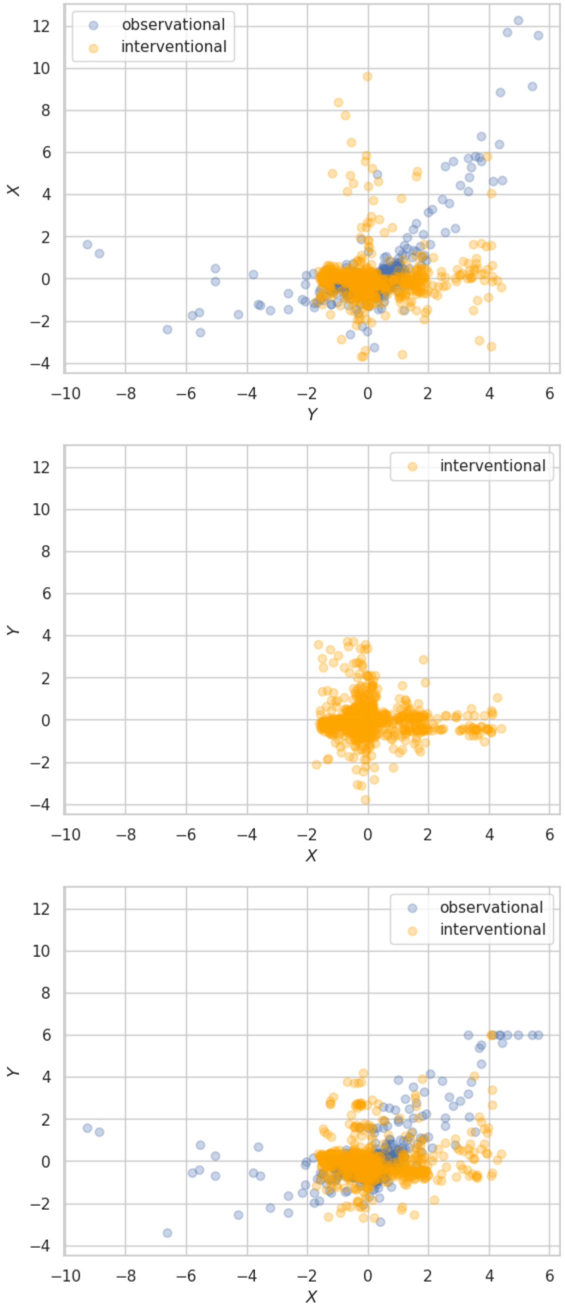

Figure 29. Dataset 13: Interventional and observational samples. (Top) training data. (Center) samples from flow model trained with only interventional data. (Bottom) samples from flow model trained with observational and interventional data.

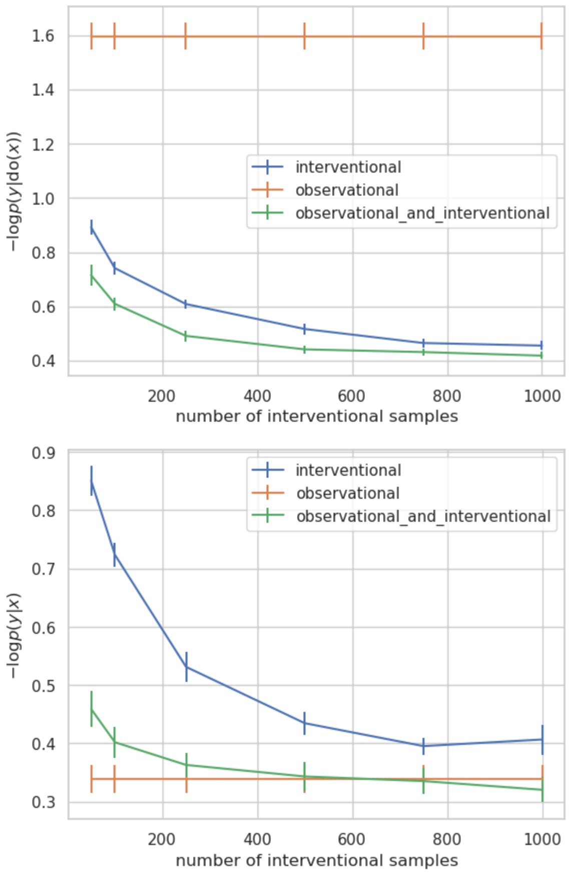

Figure 30. Dataset 13: Performances measured in terms of negative log-likelihood on the interventional (top) and the observational (bottom) test sets.

D.3.14 Dataset 14: # of latent confounders = 3, # of observed confounders = 3, random seed = 5

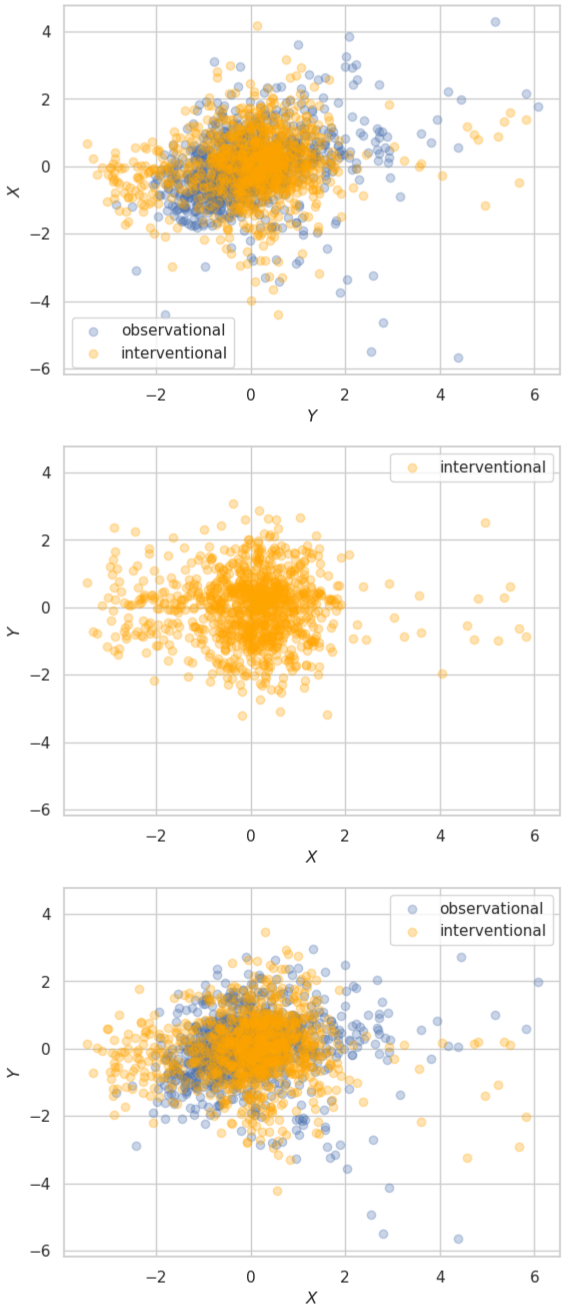

Figure 31. Dataset 14: Interventional and observational samples. (Top) training data. (Center) samples from flow model trained with only interventional data. (Bottom) samples from flow model trained with observational and interventional data.

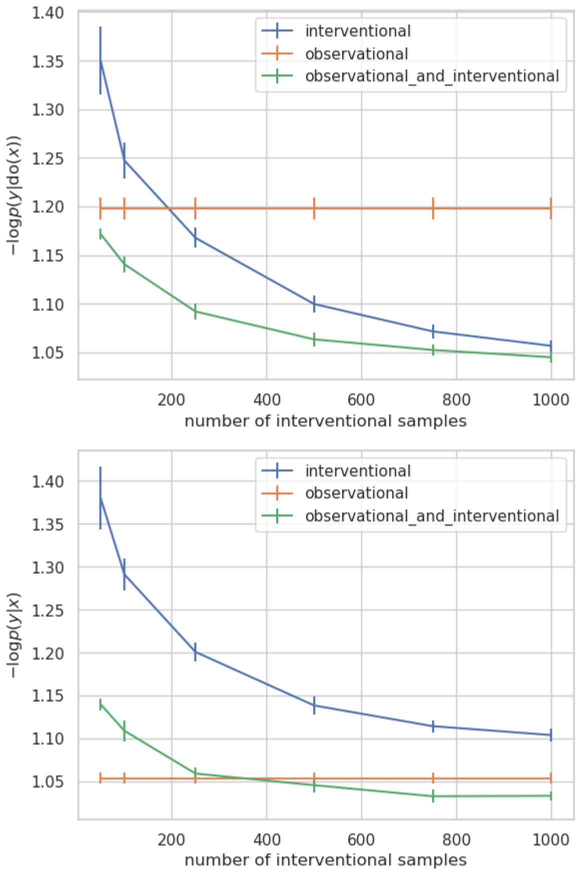

Figure 32. Dataset 14: Performances measured in terms of negative log-likelihood on the interventional (top) and the observational (bottom) test sets.

D.3.15 Dataset 15: # of latent confounders = 4, # of observed confounders = 4, random seed = 2

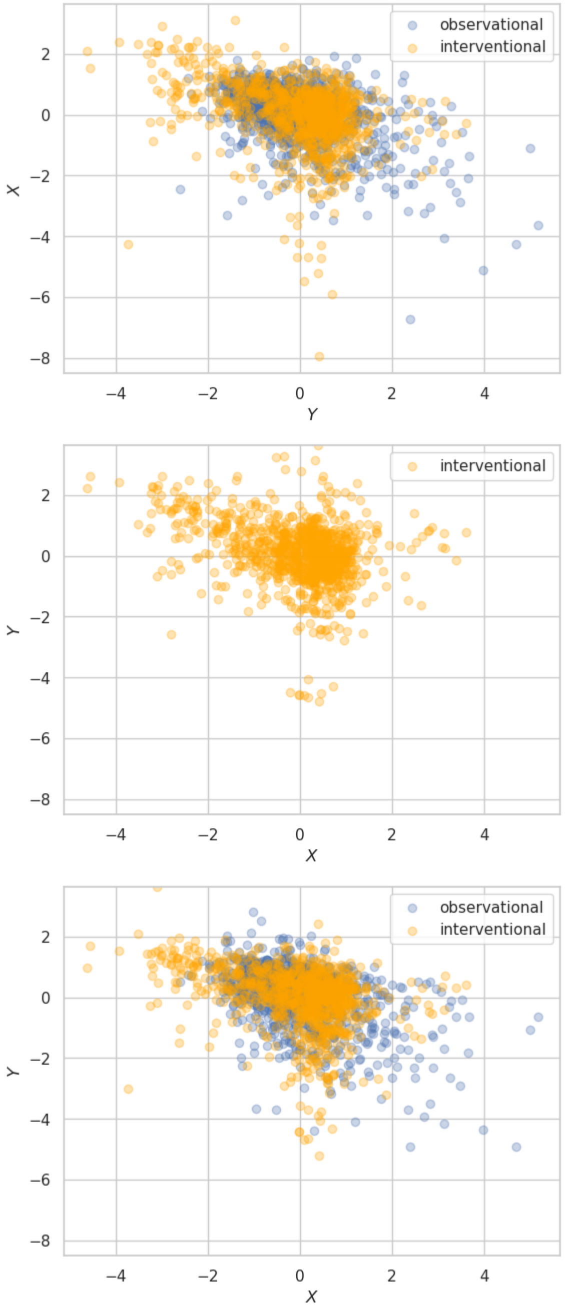

Figure 33. Dataset 15: Interventional and observational samples. (Top) training data. (Center) samples from flow model trained with only interventional data. (Bottom) samples from flow model trained with observational and interventional data.

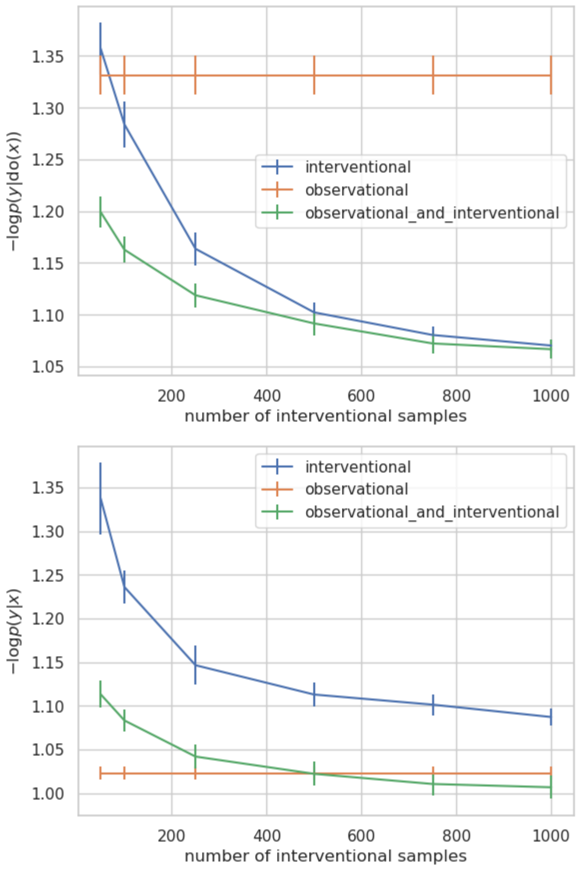

Figure 34. Dataset 15: Performances measured in terms of negative log-likelihood on the interventional (top) and the observational (bottom) test sets.