Learning Joint Nonlinear Effects from Single-variable Interventions in the Presence of Hidden Confounders

Sorawit Saengkyongam
University College London

Ricardo Silva
University College London
The Alan Turing Institute

Abstract

We propose an approach to estimate the effect of multiple simultaneous interventions in the presence of hidden confounders. To overcome the problem of hidden confounding, we consider the setting where we have access to not only the observational data but also sets of *single-variable* interventions in which each of the treatment variables is intervened on separately. We prove identifiability under the assumption that the data is generated from a nonlinear continuous structural causal model with additive Gaussian noise. In addition, we propose a simple parameter estimation method by pooling all the data from different regimes and jointly maximizing the combined likelihood. We also conduct comprehensive experiments to verify the identifiability result as well as to compare the performance of our approach against a baseline on both synthetic and real-world data.

1 MOTIVATION & CONTRIBUTION

To adopt data-driven approaches in many scientific or business domains, one often has to deal with causal inference problems where the causal effects of different treatments are of interest. Estimating causal effects purely from observational data, however, often requires strong untestable assumptions. One such major assumption is an absence of *hidden confounders* (unobserved common causes) which is often violated in practice. For instance, consider the problem of inferring the effect of medical treatments from past observational data. It is difficult to account for all confounders since we often do not know precisely how the doctors prescribed the treatments in the first place.

A gold standard method to overcome the issue of hidden confounding is to conduct a controlled experiment in which the variables of interest are directly intervened by the examiner, and thus all of the confounders are fully observed by design or eliminated altogether (in the case of randomized experiments). In the previous example, one could conduct a randomized controlled trial where the treatments are randomly assigned to the patients. By doing so, all of the confounders are eliminated and treatment effects can be estimated with standard statistical inference methods.

In many real-world scenarios, however, we may not be able to obtain experimental data due to feasibility issues such as budget, time, and ethical constraints. These feasibility constraints are especially more prominent when we consider the effect of joint interventions. One such example is multiple gene knockout experiments where we aim to estimate the effect of simultaneous gene knockouts on a phenotype of interest. The number of possible combinations of simultaneous knockouts increases exponentially with the number of genes. Conducting such a large number of knockout experiments are, therefore, not practically feasible. The problem is even more prominent when there are more than two treatment levels per variable. In this case, using causal models to predict joint interventional effects from observational data is an attractive option. However, as mentioned before, existing methods often assume the absence of hidden confounders for some pairs of variables, or return uninformative results if such a condition is not satisfied. To remove this strong assumption, we consider a setting where, in addition to the observational data, we have access to sets of *single-variable* interventional data in which each treatment variable is intervened on separately. We believe this is a practically reasonable setting since the number of interventional experiments needed only grows linearly with the number of treatment variables.

Example 1. To illustrate the setting of our problem, we consider the following hypothetical example with two treatment variables. A biologist seeks to predict a phenotypic response (Y) to simultaneous perturbations of genes X_1 and X_2, where all measurements are continuous. Because of some practical constraints, the biologist can only obtain data generated from single-gene perturbation experiments in addition to the observational data. In other words, the biologist is given observational data where Y, X_1, X_2 are distributed according to the natural regime, and two sets of interventional data where X_1 and X_2, are being perturbed separately, at different levels. The goal is to predict the effect of joint interventions at combined levels of genes X_1 and X_2 on the phenotype Y. Figure 1 depicts a graphical representation of this simplified example.

Contribution: In this work, we consider a problem of learning joint interventional effects in the presence of hidden confounders from a combination of observational data and *single-variable* interventions. We first show that, without any restrictions on the underlying structural equations, the effect of joint interventions is not identifiable. However, by introducing some reasonably weak assumptions, namely additive noise models, the effect is then identifiable. Furthermore, we introduce an inference algorithm to learn the parameters of the proposed model by pooling data across different regimes and jointly maximizing the combined likelihood. Our result provides the means to dramatically reduce the number of interventional experiments needed to estimate the effect of joint interventions compared to a blackbox model.

2 BACKGROUND & RELATED WORK

2.1 BACKGROUND

Structural Causal Models: Our work relies on the structural causal model (SCM) framework described by Pearl (2009). SCMs allow us to encode causal assumptions by describing the data generating process not only in terms of probability distributions, but also its causal mechanisms through a set of structural equations. Specifically, an SCM $\mathcal{M}$ is a 4-tuple $\langle \boldsymbol{X}, \boldsymbol{U}, \boldsymbol{f}, P_{\boldsymbol{U}} \rangle$, where

1. $\boldsymbol{X}$ represents endogenous variables.
2. $\boldsymbol{U}$ represents exogenous ("noise") variables.
3. $\boldsymbol{f}$ is a set of structural equations, each of which describing how the corresponding endogenous variable causally depends on other variables, i.e. $X_i = f_i(\mathbf{PA}_{X_i}, \boldsymbol{U}_i)$ (the set $\mathbf{PA}_{X_i} \in \boldsymbol{X} \backslash X_i$ are called
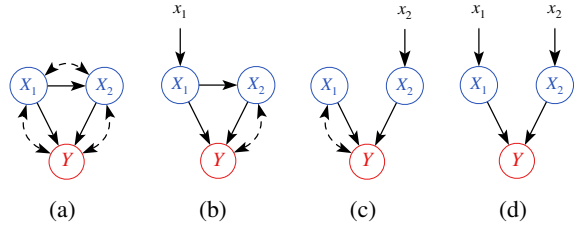

Figure 1: An example with two treatment variables. (a) the causal graph associated with the unintervened SCM $\mathcal{M}$. (b, c) the causal graphs associated with the singly-intervened SCMs $\mathcal{M}_{do(X_1)}$, $\mathcal{M}_{do(X_2)}$. (d) the causal graph associated with the jointly-intervened SCM $\mathcal{M}_{do(X_1,X_2)}$. We are given data from (a), (b), (c) and aim to make predictions in (d).

parents or *direct causes* of X_i, and $\boldsymbol{U}_i \in \boldsymbol{U}$ are the corresponding noise variables).

4. $P_{\boldsymbol{U}}$ is a joint probability distribution over the noise variables $\boldsymbol{U}$.

An SCM $\mathcal{M}$ entails a joint distribution over the endogenous variables denoted by $P_{\boldsymbol{X}}^{\mathcal{M}}$. If $\mathcal{M}$ is unaffected by any intervention, we refer to the entailed distribution as the *observational distribution*. An SCM also induces a unique causal graph G. Each node in G represents one of the endogenous variables, while edges denote the causal mechanisms defined by the structural equations $\boldsymbol{f}$. Specifically, the directed edge is drawn from X_j to X_i if $X_j \in \mathbf{PA}_{X_i}$. Furthermore, if exogenous variables are not jointly independent, a bi-directed edge will be drawn between the corresponding pairs of endogenous variables. In other words, a bi-directed edge is drawn between X_i and X_j if $\boldsymbol{U}_i \not\perp\!\!\!\perp \boldsymbol{U}_j$. These dependencies between exogenous variables indicate the presence of *hidden confounders*.

Perfect Intervention: A crucial property of SCMs is the invariance of structural equations, that is, each structural equation is unaffected by any changes made on other structural equations. This property allows us to define a mathematical operator for performing perfect intervention called the "$do(\cdot)$" operator. Applying "$do(X_i = x)$" operator on an SCM $\mathcal{M}$ results in a post-intervened SCM $\mathcal{M}_{do(X_i=x)}$ in which the original equation for X_i is replaced by the constant x. From $\mathcal{M}_{do(X_i)}$, we can obtain the entailed distribution $P_{\boldsymbol{X}}^{\mathcal{M}_{do(X_i)}}$. This is referred to as an *interventional distribution*.

2.2 RELATED WORK

Identification of nonparametric SCMs: The identification of causal effects under nonparametric structural

causal models has been extensively studied over the past decades. In the nonparametric case, the problem of identification asks whether the causal queries can be uniquely computed from available data and a given causal graph regardless of the underlying structural equations and the distribution of the noise variables. The early work focused on identifying causal effects purely from observational data (Pearl, 1995; Tian & Pearl, 2002; Huang & Valtorta, 2008). Later on, Bareinboim & Pearl (2012) extended the notion of identifiability to consider not only observational data but also interventional data called z-identifiability. Recently, more general results were introduced by Lee et al. (2019) in which a necessary and sufficient graphical condition for identifying causal effects from arbitrary combinations of observational and interventional distributions was provided.

Despite the great progress in nonparametric identification over the past decade, causal effects in many real-world settings are not identifiable under this framework due to its reliance on independence constraints only. In particular, we show in Section 4 that the effect of joint interventions is not nonparametrically identifiable under our problem setting.

Additive Noise Models: By imposing some restrictions on the structural equations, one could obtain more powerful identifiability results. One of the most commonly used restrictions are the additive noise models (ANMs) (Hoyer et al., 2009) which limit the form of the structural equations to be additive with respect to the noise variables. Based on ANMs, Janzing et al. (2009) proposed a method for inferring a latent confounder between two observed random variables which is otherwise not possible without additional assumptions. Concerning the task of learning causal structure, ANMs also permit us to recover the underlying causal graph from observational data beyond the Markov equivalence class (Zhang & Hyvärinen, 2009; Mooij et al., 2011; Peters & Bühlmann, 2013; Peters et al., 2014). Inspired by the previous successes, our work relies on this model to provide identifiability of joint interventional effects.

Effect Estimation in SCMs: Related to our problem, Nandy et al. (2017) developed a method to estimate the effect of joint interventions from observational data when the causal structure is not known, joint-IDA, which is an extension of the IDA algorithm of Maathuis et al. (2009). In IDA, the assumed causal structure is obtained via a causal discovery algorithm that returns all graphs in the Markov equivalence class. The algorithm exploits properties of a linear SCM with Gaussian noise to avoid the naïve approach of enumerating all graphs in the Markov equivalence class. Hidden confounders are allowed in the extension by Malinsky & Spirtes (2016),

where the Markov equivalence class is less informative and a reasonable degree of sparsity (in the observable marginal distribution) is to be expected if nontrivial solution sets are to be returned. Another closely related work is the work of Hyttinen et al. (2012). They considered the problem of learning the structure and parameters of linear cyclic SCMs, where they make use of interventional data to remove the hidden confounding assumption. However, their approach relies on the linearity assumption, which is a particularly strong restriction.

The goal of our work is broadly similar to that by Nandy et al. (2017) and Hyttinen et al. (2012). However, we drop the linearity assumption used in both of the works. Besides, no sparsity is required for identification. The price to be paid is our requirement for single-variable interventions to be observed, which we believe is a reasonable setting in important real-world applications. This complements the literature on methods for sparse Markov equivalence classes with purely observational data.

3 PROBLEM SETTING

We now define the setting we consider in our work.

Let $\boldsymbol{X} = \{X_1, \ldots, X_K\}$ be a set of treatment variables and Y be an outcome variable. Denote $\boldsymbol{X}^k := \{X_i\}_{i=1}^k$. $\boldsymbol{X}$ and Y are assumed to be generated from an SCM $\mathcal{M}$,

$$Y = f_Y(\boldsymbol{X}^K, U_Y)$$
$$X_k = f_k(\boldsymbol{X}^{k-1}, U_k), \text{ for } k = 1, \ldots, K, \quad (1)$$

where $U_1, \ldots, U_Y$ are the noise variables (note that we do not assume the noise variables to be mutually independent).

We are given i.i.d. samples from observational regime $D_{obs} \sim P_{(\boldsymbol{X}, Y)}^{\mathcal{M}}$. In addition, we are also given a set of i.i.d. samples from single-variable interventional regimes where we intervene on $X_1, \ldots, X_K$ separately, i.e. $\boldsymbol{D}_{int} = \{D_{int}^k \sim P_{(\boldsymbol{X}, Y)}^{\mathcal{M}_{do(X_k)}}\}_{k=1}^K$.

The causal query of interest is the effect of joint interventions in terms of the conditional expectation $Q(\mathcal{M}) = \mathbb{E}[Y \mid do(X_1, \ldots, X_K)]$.

Figure 1 illustrates an example with two treatment variables.

4 UNIDENTIFIABILITY UNDER UNCONSTRAINED SCMs

We will now show that, without additional assumptions on the form of structural equations, the joint interventional effects are not identifiable. To illustrate the

unidentifiability, we consider the case where there are two treatment variables. We will show that there exists a pair of SCMs $\ddot{\mathcal{M}}, \tilde{\mathcal{M}}$ such that they entail identical observational distribution ($P^{\ddot{\mathcal{M}}} = P^{\tilde{\mathcal{M}}}$) as well as identical single-variable interventional distributions ($P^{\ddot{\mathcal{M}}_{do(X_1)}} = P^{\tilde{\mathcal{M}}_{do(X_1)}}$ and $P^{\ddot{\mathcal{M}}_{do(X_2)}} = P^{\tilde{\mathcal{M}}_{do(X_2)}}$), but induce different joint interventional distributions ($P^{\ddot{\mathcal{M}}_{do(X_1,X_2)}} \neq P^{\tilde{\mathcal{M}}_{do(X_1,X_2)}}$).

Define SCMs $\ddot{\mathcal{M}}, \tilde{\mathcal{M}}$ as follows,

$\ddot{\mathcal{M}}$	$\tilde{\mathcal{M}}$
$Y = X_1 \wedge X_2 \wedge U_Y$	$Y = X_2 \wedge U_Y$
$X_2 = X_1 \wedge U_2$	$X_2 = X_1 \wedge U_2$
$X_1 = U_1$	$X_1 = U_1$

where $U_Y = U_2 = U_1 \sim \text{Bernoulli}(p)$.

Tables 1, 2, 3 show the joint probabilities in the observational and single-variable interventional regimes. The two given SCMs yield identical observational distributions and identical single-variable interventional distributions. However, they yield different joint interventional distributions,

$$P^{\ddot{\mathcal{M}}_{do(X_1=0,X_2=1)}}(y) = \begin{cases} 1, & \text{if } y = 0 \\ 0, & \text{otherwise} \end{cases}$$

$$\neq P^{\tilde{\mathcal{M}}_{do(X_1=0,X_2=1)}}(y) = \begin{cases} p, & \text{if } y = 1 \\ 1-p, & \text{if } y = 0 \\ 0, & \text{otherwise} \end{cases}$$

Hence, the join interventional effects are non-identifiable under unconstrained SCMs.

5 PROPOSED APPROACH

The result in the previous section motivates us to consider a restricted class of causal models, particularly additive noise models. In additive noise models, we impose an additional assumption on the structural equations, that is, the noise variables (exogenous variables) affect the observables (endogenous variables) in an additive way. Considering our problem definition, we essentially replace (1) with

$$Y = f_Y(\mathbf{X}^K) + U_Y$$
$$X_k = f_k(\mathbf{X}^{k-1}) + U_k, \text{ for } k = 1, \dots, K \quad (2)$$

while the rest remain the same. Furthermore, the noise distribution is assumed to be a zero-mean Gaussian: $(U_1, \dots, U_Y) \sim \mathcal{N}(\mathbf{0}, \Sigma)$, where Σ is an arbitrary covariance matrix of size $(K+1) \times (K+1)$.

5.1 MAIN ASSUMPTIONS

Our approach relies on the following main assumptions.

Table 1: Observational Distribution: $P^{\ddot{\mathcal{M}}}, P^{\tilde{\mathcal{M}}}$

$P(X_1, X_2, Y)$	$Y = 0$	$Y = 1$
$X_1, X_2 = 0, 0$	$1 - p$	0
$X_1, X_2 = 1, 0$	0	0
$X_1, X_2 = 0, 1$	0	0
$X_1, X_2 = 1, 1$	0	p

Table 2: Interventional Distribution: $P^{\ddot{\mathcal{M}}_{do(X_1)}}, P^{\tilde{\mathcal{M}}_{do(X_1)}}$

	$P(X_2, Y)$	$Y = 0$	$Y = 1$
$do(X_1 = 0)$	$X_2 = 0$	1	0
	$X_2 = 1$	0	0
$do(X_1 = 1)$	$P(X_2, Y)$	$Y = 0$	$Y = 1$
	$X_2 = 0$	$1 - p$	0
	$X_2 = 1$	0	p

Table 3: Interventional Distribution: $P^{\ddot{\mathcal{M}}_{do(X_2)}}, P^{\tilde{\mathcal{M}}_{do(X_2)}}$

	$P(X_1, Y)$	$Y = 0$	$Y = 1$
$do(X_2 = 0)$	$X_1 = 0$	$1 - p$	0
	$X_1 = 1$	p	0
$do(X_2 = 1)$	$P(X_1, Y)$	$Y = 0$	$Y = 1$
	$X_1 = 0$	$1 - p$	0
	$X_1 = 1$	0	p

1. (Additive noise) the underlying model is assumed to be an additive noise model.

2. (Gaussian noise) the noise distribution is assumed to be a zero-mean Gaussian with arbitrary covariance matrix.

3. (Known causal structure) the underlying causal structure is assumed to be given.

4. (Acyclicity) the underlying causal structure can be represented by an acyclic directed mixed graph (ADMG, Richardson, 2003).

Assumptions 1 and 2 play a major role in providing the identifiability of joint interventional effects. The main implication of the additive Gaussian noise assumption is that we can only consider continuous variables. Nonetheless, we believe that the Gaussianity assumption can be relaxed to some extent which may shed some light on how to handle discrete variables. We leave this for future work. Furthermore, we note that we assume Assumption 3 for clarity of exposition where we focus on the identifiability and estimation of the joint effects. However, it is conceptually straightforward to obtain an estimated causal structure in our setting as we have access to the single-variable interventions for all treatment variables (Eberhardt et al., 2006).

5.2 MODEL IDENTIFICATION

Our main theoretical result is the identification of the effect of joint interventions on all control variables $\boldsymbol{X}$ from a combination of observational data D_{obs} and single-variable interventional data $\boldsymbol{D}_{int}$.

Theorem 1 (Identifiability of joint nonlinear effects under additive noise models)
Let $\mathcal{M}^K = \langle \{\boldsymbol{X}, Y\}, \boldsymbol{U}, \boldsymbol{f}, P_U \rangle$ be an additive noise SCM with K treatment variables,

$$Y = f_Y(\boldsymbol{X}^K) + U_Y$$
$$X_k = f_k(\boldsymbol{X}^{k-1}) + U_k, \, for \, k = 1, \dots, K$$

with $P_{\boldsymbol{U}} \sim \mathcal{N}(\boldsymbol{0}, \Sigma)$, where Σ is an arbitrary covariance matrix and $\boldsymbol{X}^k := \{X_i\}_{i=1}^k$. The causal query $Q(\mathcal{M}^K) = \mathbb{E}[Y \mid do(X_1, \dots, X_K)]$ is identifiable from a combination of the observational distribution $P_{(\boldsymbol{X},Y)}^{\mathcal{M}^K}$ and the set of single-variable interventional distributions $\{P_{(\boldsymbol{X},Y)}^{\mathcal{M}_{do(X_i)}^K}\}_{i=1}^K$, for any integer $K \geq 2$.

We prove the theorem by induction. In the base case, we show that $\mathbb{E}[Y \mid do(X_1, X_2)]$ is identifiable. In the inductive step, given that $\mathbb{E}[Y \mid do(X_1, \dots, X_K)]$ is identifiable, we show that $\mathbb{E}[Y \mid do(X_1, \dots, X_{K+1})]$ is also identifiable. A sketch of the proof for the base case is provided below. Please refer to the supplementary material for the full proof.

Proof sketch for the base case. The query of interest is

$$Q(\mathcal{M}^2) = \mathbb{E}[Y \mid do(X_1 = x_1, X_2 = x_2)] = f_Y(x_1, x_2).$$

Due to unobserved confounders, the above query is not identifiable solely from the observational distribution $P_{(\boldsymbol{X},Y)}^{\mathcal{M}^2}$,

$$\mathbb{E}[Y \mid X_1 = x_1, X_2 = x_2]$$
$$= f_Y(x_1, x_2) + \mathbb{E}[U_Y \mid X_1 = x_1, X_2 = x_2].$$

However, if we are able to identify $\mathbb{E}[U_Y \mid X_1 = x_1, X_2 = x_2]$, we would then be able to identify our query of interest $f_Y(x_1, x_2)$. We then need to show that the expected noise $\mathbb{E}[U_Y \mid X_1 = x_1, X_2 = x_2]$ can be uniquely computed from a combination of $P_{(\boldsymbol{X},Y)}^{\mathcal{M}^2}$, $P_{(\boldsymbol{X},Y)}^{\mathcal{M}_{do(X_1)}^2}$ and $P_{(\boldsymbol{X},Y)}^{\mathcal{M}_{do(X_2)}^2}$.

From the additive Gaussian noise assumptions, we have

$$\mathbb{E}[U_Y \mid X_1 = x_1, X_2 = x_2] = \Sigma_{u_y} \Sigma_{u_x}^{-1} \boldsymbol{u}_x, \quad (3)$$

with $\boldsymbol{u}_x = [x_1 \quad x_2 - f_2(x_1)]^\top$, $\Sigma_{u_y} = [\sigma_{Y1} \quad \sigma_{Y2}]$ and $\Sigma_{u_x} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$, where we define $\sigma_{ij} := \text{Cov}(U_i, U_j)$.

From Equation (3), the quantities that we need to show the identifiability are f_2, Σ_{u_x} and Σ_{u_y}.

Identifying f_2 and Σ_{u_x}

f_2 can be trivially obtained from $P_{(\boldsymbol{X},Y)}^{\mathcal{M}_{do(X_1)}^2}$,

$$\mathbb{E}[X_2 \mid do(X_1 = x_1)] = f_2(x_1).$$

Since f_2 is identified, we can then identify the joint distribution $p(U_1, U_2)$ from $P_{(\boldsymbol{X},Y)}^{\mathcal{M}^2}$,

$$U_2 = X_2 - f_2(X_1)$$
$$U_1 = X_1$$

And thus, the covariance matrix Σ_{u_x} is identifiable.

Identifying σ_{Y1} and σ_{Y2}

The idea is to identify these covariances by contrasting the expected treatment responses in the different interventional regimes. We give the sketch for the identification of σ_{Y1} below (σ_{Y2} can be obtained in a similar fashion).

From the regime $P_{(\boldsymbol{X},Y)}^{\mathcal{M}_{do(X_1)}^2}$, we have

$$\mathbb{E}[Y \mid do(X_1 = x_1)] = \mathbb{E}[f_y(x_1, f_2(x_1) + U_2)] \quad (4)$$

From the regime $P_{(\boldsymbol{X},Y)}^{\mathcal{M}_{do(X_2)}^2}$, we have

$$\mathbb{E}[Y \mid X_1 = x_1, do(X_2 = x_2)]$$
$$= f_y(x_1, x_2) + \mathbb{E}[U_Y \mid X_1 = x_1]$$

We can hypothetically choose $x_2 = f_2(x_1) + u_2$ and treat the above solely as a mathematical expression that can take a random variable as an input, in this case U_2:

$$\mathbb{E}[Y \mid X_1 = x_1, do(X_2 = f_2(x_1) + U_2)].$$

We then take its expectation with respect to the identifiable marginal $p(U_2)$. The left-hand side of the equation below is also identifiable since f_2 is and we observe all single-variable interventions.

$$\mathbb{E}_{p(U_2)}[\mathbb{E}[Y \mid X_1 = x_1, do(X_2 = f_2(x_1) + U_2)]]$$
$$= \mathbb{E}[f_y(x_1, f_2(x_1) + U_2)] + \mathbb{E}[U_Y \mid X_1 = x_1]. \quad (5)$$

Subtracting (4) from (5), we get

$$(5) - (4) = \mathbb{E}[U_Y \mid X_1 = x_1] = \mathbb{E}[U_Y \mid U_1 = x_1].$$

The covariance σ_{Y1} can then be obtained from $\mathbb{E}[U_Y | U_1]$ and $P(U_1)$. $\qquad \square$

5.3 MODEL ESTIMATION

In this section, we propose a joint modeling approach to learn the parameters of the additive noise causal models from a combination of observational and interventional data. In particular, we define a likelihood for each data point according to its corresponding regime and then maximize the combined likelihood in which the model parameters are shared across different regimes and jointly optimized.

To define the combined likelihood, we first rewrite (2) using vector notation where we define $X_Y := Y$ for notational convenience:

$$X = f(X) + U$$

with $X = \begin{bmatrix} X_1 \\ \vdots \\ X_Y \end{bmatrix}, f(X) = \begin{bmatrix} f_1(\mathbf{PA}_{X_1}) \\ \vdots \\ f_Y(\mathbf{PA}_Y) \end{bmatrix}, U = \begin{bmatrix} U_1 \\ \vdots \\ U_Y \end{bmatrix}.$

By the change of variables, the distribution of the endogenous variables can be expressed in terms of the distribution of the noise variables as,

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = p_{\boldsymbol{U}}(\boldsymbol{x} - \boldsymbol{f}(\boldsymbol{x})) |\det(\boldsymbol{I} - \Delta \boldsymbol{f}(\boldsymbol{x}))|$$

where $\boldsymbol{I}$ denotes the identity mapping and $\Delta \boldsymbol{f}(\boldsymbol{x})$ denotes a Jacobian of $\boldsymbol{f}$ evaluated at $\boldsymbol{x}$. Since we assume that the underlying SCM is acyclic, the term $|\det(\boldsymbol{I} - \Delta \boldsymbol{f}(\boldsymbol{x}))|$ is equal to 1. We can then define the likelihood of each data point according to its corresponding regime as follows.

Likelihood

We are given an i.i.d sample of size N_0 from the observational regime denoted by $D_{obs} = \{\overset{0}{\boldsymbol{x}}{}^{(n)}\}_{n=1}^{N_0}$ and a set of K i.i.d samples of size $\{N_1, \dots, N_K\}$. Each of which is generated from the interventional regime where the treatment X_k is intervened on, denoted by $\boldsymbol{D}_{int} = \{D_{int}^k = \{\overset{k}{\boldsymbol{x}}{}^{(n)}\}_{n=1}^{N_k}\}_{k=1}^{K}$. In addition, we parameterize functions $\boldsymbol{f}$ with $\boldsymbol{\theta}$, denoted by $\boldsymbol{f}(\cdot; \boldsymbol{\theta})$.

The likelihood for each data point in D_{obs} is defined as,

$$L^0(\boldsymbol{\theta}, \Sigma^0; \overset{0}{\boldsymbol{x}}{}^{(n)}) = p_{\boldsymbol{U}}^0(\overset{0}{\boldsymbol{x}}{}^{(n)} - \boldsymbol{f}(\overset{0}{\boldsymbol{x}}{}^{(n)}; \boldsymbol{\theta}); \Sigma^0)$$

where $p_{\boldsymbol{U}}^0 \sim \mathcal{N}(\boldsymbol{0}, \Sigma^0)$, and Σ^0 is a covariance matrix of size $(K+1) \times (K+1)$.

As for the interventional regime, the intervened variable X_k is replaced by a constant, and thus the likelihood reduces to the joint probability of all other variables $\boldsymbol{X} \setminus X_k$ given the parameters.

Define $\overset{k}{\boldsymbol{x}}{}_{-k}^{(n)} := \overset{k}{\boldsymbol{x}}{}^{(n)} \setminus \overset{k}{x}{}_k^{(n)}, \boldsymbol{U}_{-k} := \boldsymbol{U} \setminus U_k, \boldsymbol{f}(\cdot)_{-k} := \boldsymbol{f}(\cdot) \setminus f_k(\cdot)$. The likelihood for each data point in D_{int}^k

can be defined as,

$$L^k(\boldsymbol{\theta}, \Sigma^k; \overset{k}{\boldsymbol{x}}{}^{(n)}) = p_{\boldsymbol{U}_{-k}}^k(\overset{k}{\boldsymbol{x}}{}_{-k}^{(n)} - \boldsymbol{f}_{-k}(\overset{k}{\boldsymbol{x}}{}^{(n)}; \boldsymbol{\theta}); \Sigma^k)$$

where $p_{\boldsymbol{U}_{-k}}^k = \mathcal{N}(\boldsymbol{0}, \Sigma^k)$, and Σ^k is a covariance matrix of size $K \times K$.

We can then define the combined log-likelihood as,

$$\boldsymbol{\ell}(\boldsymbol{\theta}, \boldsymbol{\Sigma}; D_{obs}, \boldsymbol{D}_{int}) = \sum_{k=0}^{K} \sum_{n=1}^{N_k} \log L^k(\boldsymbol{\theta}, \Sigma^k; \overset{k}{\boldsymbol{x}}{}^{(n)}) \tag{6}$$

The maximum likelihood estimators for $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ can be obtained by maximizing the combined log-likelihood [1]. To do so, we can alternately solve for $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ as follows.

Solving for $\boldsymbol{\theta}$: Fixing the covariance matrices $\{\boldsymbol{\Sigma}\}$, the combined log-likelihood (6) can be maximized by standard iterative methods. The choice of an optimization algorithm can be chosen based on the choice of the functions $\boldsymbol{f}$. For instance, if we represent $\boldsymbol{f}$ by deep neural networks, we may consider commonly used gradient-based optimization algorithms such as Adam (Kingma & Ba, 2014) or Adagrad (Duchi et al., 2011).

Solving for $\boldsymbol{\Sigma}$: Fixing the functions $\boldsymbol{f}$, all the $K+1$ terms of the outer summation in (6) are disjoint and can thus be optimized separately. For each k, the problem boils down to an estimation of a multivariate normal covariance matrix for which the closed-form solution can be obtained (the maximum likelihood estimate of the covariance Σ^k is essentially the sample covariance matrix).

6 EXPERIMENTS

In this section, we conduct multiple experiments to empirically illustrate the identifiability result, and the dangers of not properly modeling the likelihood function as a combination of different regimes. Furthermore, we compare the performance of our approach against a baseline approach on both synthetic and real-world datasets.

6.1 BASELINE

Given observational data D_{obs} and sets of single-variable interventional data $\boldsymbol{D}_{int}$, a straightforward approach for estimating the effect of joint interventions is to pool all the data together $D_{pool} = \{D_{obs}, \boldsymbol{D}_{int}\}$ and directly fit a regression model on D_{pool} to estimate $\mathbb{E}[Y \mid \mathbf{PA}_Y]$. We

---

[1]We note that there could be a potential issue when the number of samples in the observational regime grows much faster than the interventional regimes, the observational regime could dominate the likelihood landscape and create undesired behavior during optimization. Our method is currently lacking treatment of such issues and we leave this for future work.

denote this approach as `REG` which is considered as our baseline throughout the experiments [2]. Even though the baseline and our approach (denoted by `ANM`) utilize the same datasets, the main difference is that the baseline does not take into account the full causal structure (it partially uses the causal structure by considering only the direct causes of the outcome in the regression model), whereas our approach does exploit the full causal structure which is crucial in obtaining consistent estimates of the joint effects as we will see in the experimental results.

6.2 SYNTHETIC EXPERIMENT

In this subsection, we illustrate our identifiability result on simulated data. Observational and interventional data are simulated according to a pre-defined SCM with correlated exogenous variables. We then compare the parameter estimates from our approach and the baseline on those simulated data. In particular, we empirically illustrate that our estimator is consistent. By using non-regularized polynomials in our function class, experiments will also empirically illustrate unbiasedness properties of our estimator even under a nonconvex optimization formulation.

6.2.1 Data Generating Process

In the synthetic experiment, we consider the case where the number of treatments $K = 3$. The data generating process is an additive noise SCM,

$$
\begin{aligned}
Y &= f_Y(X_1, X_2, X_3; \boldsymbol{\theta}_y) + U_Y \\
X_3 &= f_3(X_1, X_2; \boldsymbol{\theta}_3) + U_3 \\
X_2 &= f_2(X_1; \boldsymbol{\theta}_2) + U_2 \\
X_1 &= U_1
\end{aligned}
$$

where $(U_1, U_2, U_3, U_Y) \sim \mathcal{N}(\mathbf{0}, \Sigma_u)$.

We examine both linear and non-linear structural equations $\boldsymbol{f}$ in our experiments. Specifically, the nonlinear functions considered are polynomials with second-order interactions in addition to the main effects. The same functional forms are also used in the baseline when fitting the direct regression models.

6.2.2 Assessing Consistency and Unbiasedness

We fix a simulated SCM with pre-defined parameters and evaluate consistency of our estimator by comparing the fitted values and the ground truth values of both parameter estimates and the predicted joint effects as the sample

---
[2]We did consider another baseline where the regime indicators are used in the regression model. The results are consistently worse than `REG` approach; therefore we decided not to include the results in the main paper. Please see the supplementary material for further details.
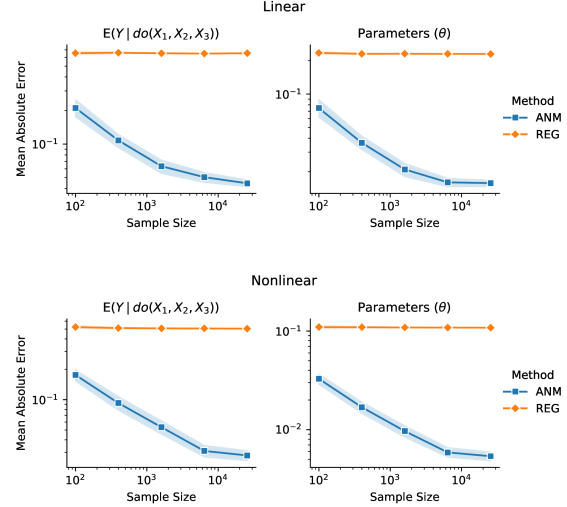
Figure 2: MAE of the predicted joint effect (left) and the parameter estimates (right) as the sample size increases. The solid lines represent the mean absolute error averaged over 50 Monte Carlo experiments and the filled regions depict its 95% confidence interval (note that both vertical and horizontal axes are in logarithmic scale).

size increases ($n_{sample} = 100, 400, 1600, 6400, 25600$). For parameter estimates, we compute the absolute differences and average among all the parameters. For the predicted joint effects, we evaluate the mean absolute error (MAE) on 100000 uniform random joint intervention test points. We perform 50 simulations for each sample size. In each simulation, the observational data of size n_{sample} is generated from the specified SCM. Furthermore, the intervention points $do(X_k = x_k)$ are randomly chosen according to the marginal distributions $P(X_k)$. The sets of K *single-variable* interventional data are then generated from the intervened SCMs, each of which has the size of n_{sample}. Figure 2 depicts the consistency result. For `ANM`, the errors both in terms of the predicted joint effect and the parameter estimates decrease and approach to zero as the sample size increase which indicates the consistency of our estimator. On the other hand, `REG` clearly suffers from inconsistency.

In addition to the consistency, we also empirically illustrate that our estimator can be unbiased[3] by comparing the sampling distributions of the parameter estimates with the ground truth values for a fixed a sample size ($n_{sample} = 1600$). The sampling distributions are simply obtained by Monte Carlo simulation. Figure 3 shows the sampling distributions of the parameter estimates. In short, we can see that our model (`ANM`) produces unbiased estimates of the model parameters, whereas the pa-

---
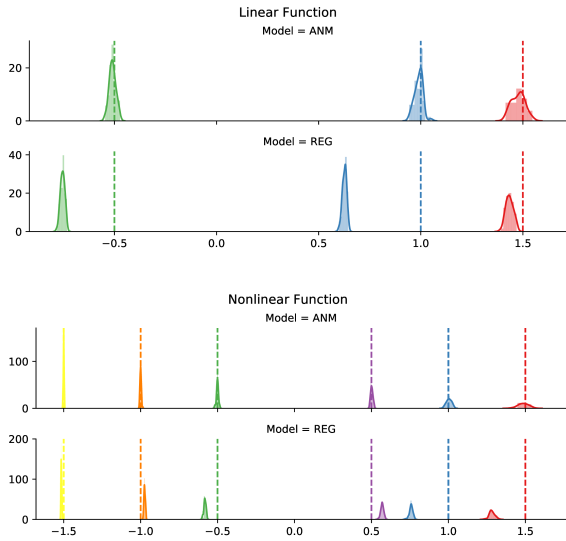[3]This will not happen in general if regularization is used.

Figure 3: Sampling distributions of the parameter estimates. Each parameter is coloured differently. The dot lines represent the ground truth parameter values.

rameter estimates obtained from the baseline (`REG`) are considerably biased regardless of sample size.

6.3 SEMI-SYNTHETIC EXPERIMENT

We further evaluate our proposed approach on data motivated by a real-world generative process. In order to have access to ground-truth, we create a semi-synthetic dataset by constructing a realistic data generating process (i.e. an SCM) from a dataset designed using expert knowledge, and use the learned SCM to generate plausible training and test data according to a best fit to an additive Gaussian error model.

6.3.1 Dataset and Setting

We base our semi-synthetic experiment on a gene expression dataset from the DREAM4 challenge (Marbach et al., 2010, 2009; Prill et al., 2010). The DREAM4 dataset consists of several gene regulation networks each of which are accompanied by several types of perturbation data, along with the gold-standard causal structure. In our experiment, we adopt one of the networks with 10 variables along with its corresponding multifactorial perturbations. The multifactorial perturbations are steady-state levels obtained after applying multifactorial perturbations which can be seen as gene expression profiles from different patients, i.e. observational data.

The multifactorial perturbations data is then used to learn an underlying SCM based on the gold standard causal structure. The underlying SCM is an additive noise

model with the nonlinear functional form described in the previous experiment. All the variables are standardized by subtracting the mean and dividing by the standard deviation. To simulate a confounding effect, we put a constraint on the correlation between the noise variables where the correlation matrix is randomly generated. Furthermore, as the gold standard causal structure may have cycles, we employ a heuristic algorithm proposed by Eades et al. (1993) to search for a minimal feedback arc set and remove them from the graph to obtain a DAG. Having obtained the underlying SCM, we then generate training data from the learned SCM where we define the outcome variable as the last node according to the topological ordering and all other nodes as treatment variables (the number of treatment variables $K = 9$). Observational data and single-variable interventions are generated in the same way as described in Section 6.2.

6.3.2 Comparing the performance of the models

We compare the performance of our model against the baseline as we vary the size of observational and single-variable interventional data ($n_{sample} = 100, 400, 1600, 6400$). The model performance is measured by how well they predict the effect of joint interventions on various test points. We generate a large number of uniform random joint intervention test points ($n_{test} = 100000$) and compute two evaluation metrics, MAE and Spearman's Rank correlation between the predicted joint interventional effects and the ground truth values. We consider Spearman's Rank correlation in addition to MAE because, in many applications, we are interested in the ranking of the predicted interventional effects rather the actual predicted values. Figure 4(a) presents the results. Our approach (`ANM`) clearly outperforms the baseline (`REG`) in both metrics. Moreover, for a large enough sample size, the performance of `ANM` is almost on a par with the oracle's performance.

6.3.3 Assessing the effect of hidden confounders

We now investigate the impact of hidden confounders on our model and the baseline. We fix the number of samples $n_{sample} = 1600$ and vary the confounding levels of the underlying SCM by varying the magnitude of the correlation coefficients of the noise variables. Specifically, we limit the size of the non-diagonal entries in the correlation matrix by a constant c (i.e. $|\text{corr}(U_i, U_j)| \leq c$ for all $i \neq j$). We consider four confounding levels $c = 0.1, 0.35, 0.65, 0.8$. Figure 4(b) depicts the impact of the unobserved confounding on both models. Observe that when the confounding level is low, the performance of `ANM` and `REG` are relatively the same as expected. However, the performance of `REG` deteriorates drastically as we increase the confounding levels, whereas the

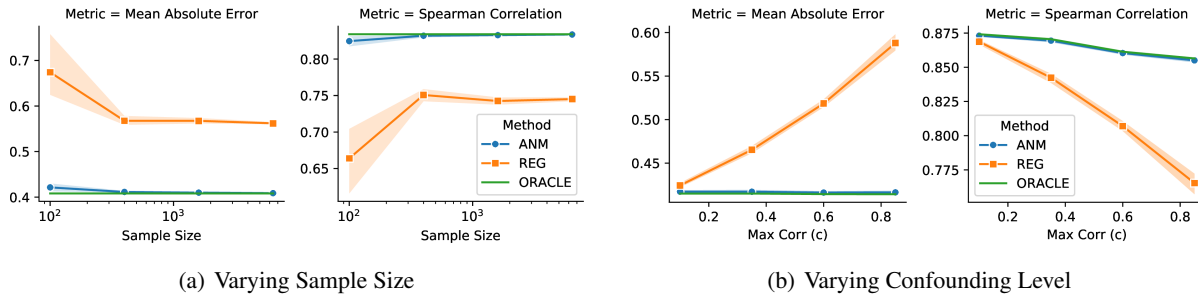(a) Varying Sample Size   (b) Varying Confounding Level

Figure 4: MAE and Spearman's Correlation (a) as the sample size increases (b) as the confounding level increases. The solid lines represent the metric values, and the filled regions depict its 95% confidence interval obtained from 50 bootstrap samples. The straight green lines represent the best possible values obtained when the ground-truth parameters are used in the model.

one of `ANM` remains unaffected by the change in the confounding levels. This demonstrates that our approach is able to disentangle the confounding effects.

6.3.4 Examining the uncertainty of the causal predictions

In the last experiment, we examine the prediction uncertainty of our proposed model. We first select 10 uniform random joint intervention test points and rank them according to the probability of being generated from the observational distribution. In particular, we estimate the probability density function of the observed variables using kernel density estimation with a Gaussian kernel. The test points are then evaluated on the estimated density and get ranked accordingly. For each test point, we estimate prediction intervals of the joint effects based on 50 bootstrap samples. Figure 5 shows the prediction intervals on the selected test points. As expected, we have high uncertainty in our predictions when the test points fall within a low-density region of the distribution of the training data. This illustrates a potential issue when there
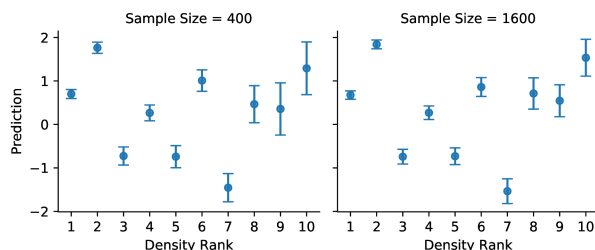
is little overlap between intervention points in the training and the test data and that we need to be careful when making predictions in the low overlap regions.

7 DISCUSSION

In this paper, we proposed an approach to combine single-variable interventions for learning the effect of joint interventions. Our approach relies on relatively weak assumptions compared to existing methods and, most importantly, does not assume the absence of hidden confounders.

In future work, we plan to tackle the setting where we may observe context variables (covariates) in addition to the treatment variables. In this setting, we are often interested in learning interventional effects conditioned on the covariates which can be high-dimensional. We hypothesize that methods for estimating heterogeneous treatment effects (Nie & Wager, 2017; Chernozhukov et al., 2018) can be integrated with our approach.

Furthermore, we believe that this line of work could play significant roles in improving exploration in reinforcement learning/bandits. Our work provides a principled way to predict unseen interventions by exploiting the causal models, which in turn reduces the number of interventions needed in exploring the action space. How to effectively incorporate our approach in reinforcement learning/bandits is a promising future research direction.

8 ACKNOWLEDGEMENTS

Figure 5: Prediction intervals of the `ANM` model. Each data point represents a test point where the x-axis represents how likely that test point is observed under the observational distribution (1:very likely, 10:very unlikely).

References

Bareinboim, E. and Pearl, J. Causal inference by surrogate experiments: z-identifiability. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pp. 113–120, 2012.

Chernozhukov, V., Demirer, M., Duflo, E., and Fernandez-Val, I. Generic machine learning inference on heterogenous treatment effects in randomized experiments. Technical report, National Bureau of Economic Research, 2018.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12 (Jul):2121–2159, 2011.

Eades, P., Lin, X., and Smyth, W. F. A fast and effective heuristic for the feedback arc set problem. *Information Processing Letters*, 47(6):319–323, 1993.

Eberhardt, F., Glymour, C., and Scheines, R. N-1 experiments suffice to determine the causal relations among n variables. In *Innovations in machine learning*, pp. 97–112. Springer, 2006.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pp. 689–696, 2009.

Huang, Y. and Valtorta, M. On the completeness of an identifiability algorithm for semi-markovian models. *Annals of Mathematics and Artificial Intelligence*, 54 (4):363–408, 2008.

Hyttinen, A., Eberhardt, F., and Hoyer, P. O. Learning linear cyclic causal models with latent variables. *Journal of Machine Learning Research*, 13(Nov):3387–3439, 2012.

Janzing, D., Peters, J., Mooij, J., and Schölkopf, B. Identifying confounders using additive noise models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 249–257, 2009.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Lee, S., Correa, J. D., and Bareinboim, E. General identifiability with arbitrary surrogate experiments. In *Proceedings of Thirty-fifth Conference on Uncertainty in Artificial Intelligence*, 2019.

Maathuis, M. H., Kalisch, M., Bühlmann, P., et al. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A): 3133–3164, 2009.

Malinsky, D. and Spirtes, P. Estimating causal effects with ancestral graph Markov models. *Proceedings of the Eighth International Conference on Probabilistic Graphical Models, PMLR*, 52:299–309, 2016.

Marbach, D., Schaffter, T., Mattiussi, C., and Floreano, D. Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *Journal of computational biology*, 16(2):229–239, 2009.

Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Floreano, D., and Stolovitzky, G. Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the national academy of sciences*, 107 (14):6286–6291, 2010.

Mooij, J. M., Janzing, D., Heskes, T., and Schölkopf, B. On causal discovery with cyclic additive noise models. In *Advances in neural information processing systems*, pp. 639–647, 2011.

Nandy, P., Maathuis, M. H., Richardson, T. S., et al. Estimating the effect of joint interventions from observational data in sparse high-dimensional settings. *The Annals of Statistics*, 45(2):647–674, 2017.

Nie, X. and Wager, S. Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*, 2017.

Pearl, J. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Pearl, J. Causality: Models, reasoning and inference. 2009.

Peters, J. and Bühlmann, P. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2013.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15 (1):2009–2053, 2014.

Prill, R. J., Marbach, D., Saez-Rodriguez, J., Sorger, P. K., Alexopoulos, L. G., Xue, X., Clarke, N. D., Altan-Bonnet, G., and Stolovitzky, G. Towards a rigorous assessment of systems biology models: the dream3 challenges. *PloS one*, 5(2):e9202, 2010.

Richardson, T. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30: 145–157, 2003.

Tian, J. and Pearl, J. A general identification condition for causal effects. In *Eighteenth national conference on Artificial intelligence*, pp. 567–573, 2002.

Zhang, K. and Hyvärinen, A. On the identifiability of the post-nonlinear causal model. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pp. 647–655, 2009.