Supplementary Material: 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

INTRODUCTION

In this supplementary material, we provide the completed proof of Theorem 1 as well as the detail of the unidentifiability result (Section 4 in the main paper). Furthermore, we present more information on the settings of the synthetic experiment. Lastly, supplemental results on the additional baseline and synthetic dataset are provided.

1 PROOF OF THEOREM 1

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

$$
\begin{aligned}
Y &= f_Y(\boldsymbol{X}^K) + U_Y \\
X_k &= f_k(\boldsymbol{X}^{k-1}) + U_k, \textit{for } k = 1, \ldots, K
\end{aligned}
$$

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 $Q(\mathcal{M}^K) = E[Y \mid do(X_1, \ldots, X_K)]$ is identifiable from a combination of the observational distribution $P_{(\boldsymbol{X},Y)}^{\mathcal{M}^K}$ and the set of single-variable interventional distributions $\{P_{(\boldsymbol{X},Y)}^{\mathcal{M}^K_{do(X_i)}}\}_{i=1}^K$, for any integer $K \geq 2$.

Proof. 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, we show that given that $\mathbb{E}[Y \mid do(X_1, \ldots, X_K)]$ is identifiable, $\mathbb{E}[Y \mid do(X_1, \ldots, X_{K+1}]$ is also identifiable.

Base Step:

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 $P_{(\boldsymbol{X},Y)}^{\mathcal{M}^2}$,

$$
\begin{aligned}
&\mathbb{E}[Y \mid X_1 = x_1, X_2 = x_2] \\
&\qquad = f_Y(x_1, x_2) + \mathbb{E}[U_Y \mid X_1 = x_1, X_2 = x_2]
\end{aligned}
$$

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 $P_{(\boldsymbol{X},Y)}^{\mathcal{M}^2}$, $P_{(\boldsymbol{X},Y)}^{\mathcal{M}^2_{do(X_1)}}$ and $P_{(\boldsymbol{X},Y)}^{\mathcal{M}^2_{do(X_2)}}$.

From the additive Gaussian noise assumptions, we have

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

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 $\sigma_{ij} = \mathrm{Cov}(U_i, U_j)$.

From Equation (1), 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 $P_{(\boldsymbol{X},Y)}^{\mathcal{M}^2_{do(X_1)}}$,

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

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

$$
\begin{aligned}
U_2 &= X_2 - f_2(X_1) \\
U_1 &= X_1
\end{aligned}
$$

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

Identifying σ_{Y1}

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

$$\mathbb{E}[Y \mid do(X_1 = x_1)] = \mathbb{E}[f_y(x_1, f_2(x_1) + U_2)] \tag{2}$$

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

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

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.

$$\begin{aligned}&\mathbb{E}_{p(U_2)}[\mathbb{E}[Y \mid X_1 = x_1, do(X_2 = f_2(x_1) + U_2)]]\\ &\quad = \mathbb{E}[f_y(x_1, f_2(x_1) + U_2)] + \mathbb{E}[U_Y \mid X_1 = x_1].\end{aligned} \tag{3}$$

Subtracting (2) from (3), we get

$$(3) - (2) = \mathbb{E}[U_Y \mid X_1 = x_1] = \mathbb{E}[U_Y \mid U_1 = x_1].$$

Lemma 1. *Let U_1 and U_2 be zero-mean random variables. The covariance $\mathrm{Cov}(U_1, U_2)$ can be identified from the conditional expectation $\mathbb{E}[U_2 \mid U_1]$ and the marginal distribution $p(U_1)$.*

Proof.

$$\begin{aligned}\mathrm{Cov}(U_1, U_2) &= \mathbb{E}[U_1 U_2]\\ &= \int_{u_2, u_1} u_2 u_1 p(u_2, u_1)\\ &= \int_{u_1} u_1 \int_{u_2} u_2 p(u_2 \mid u_1) p(u_1)\\ &= \int_{u_1} u_1\, \mathbb{E}[U_2 \mid u_1] p(u_1)\\ &= \mathbb{E}_{U_1 \sim p(U_1)}\left[U_1\, \mathbb{E}[U_2 \mid U_1]\right].\end{aligned}$$

□

By Lemma 1, the covariance $\sigma_{Y1} = \mathrm{Cov}(U_y, U_1)$ is identified from $\mathbb{E}[U_y \mid U_1]$ and $p(U_1)$.

Identifying σ_{Y2}

From the regime $P^{\mathcal{M}^2_{do(X_1)}}_{(\boldsymbol{X},Y)}$, and the one-to-one mapping between X_2 and U_2 for a fixed x_1, we get

$$\begin{aligned}&\mathbb{E}[Y \mid do(X_1 = x_1), U_2 = u_2]\\ &= \mathbb{E}[f_y(X_1, f_2(X_1) + U_2) + U_y \mid do(X_1 = x_1), U_2 = u_2]\\ &= f_y(x_1, f_2(x_1) + u_2) + \mathbb{E}[U_y \mid U_2 = u_2].\end{aligned}$$

Since we can trivially obtain $p(X_1)$, we can then take an expectation over X_1 according to that distribution,

$$\begin{aligned}&\mathbb{E}_{p(X_1)}[\mathbb{E}[Y \mid do(X_1 = x_1), U_2 = u_2]]\\ &= \mathbb{E}[f_y(X_1, f_2(X_1) + u_2)] + \mathbb{E}[U_y \mid U_2 = u_2].\end{aligned} \tag{4}$$

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

$$\begin{aligned}&\mathbb{E}[Y \mid X_1 = x_1, do(X_2 = x_2)]\\ &= \mathbb{E}[f_y(X_1, X_2) + U_y \mid X_1 = x_1, do(X_2 = x_2)]\\ &= f_y(x_1, x_2) + \mathbb{E}[U_y \mid X_1 = x_1]\end{aligned}$$

Since we have identified f_2 and $p(X_1)$, we can then theoretically choose $x_2 = f_2(x_1) + u_2$ and take the expectation over $p(X_1)$,

$$\begin{aligned}&\mathbb{E}_{p(X_1)}[\mathbb{E}[Y \mid X_1, do(X_2 = f_2(X_1) + u_2)]]\\ &= \mathbb{E}[f_y(X_1, f_2(X_1) + u_2)].\end{aligned} \tag{5}$$

From (4)-(5), we have

$$(4) - (5) = \mathbb{E}[U_y \mid U_2 = u_2].$$

$\sigma_{Y2} = \mathrm{Cov}(U_y, U_2)$ is now identifiable using Lemma 1.

Inductive Step:

Claim 1. *Given that the causal query $Q(\mathcal{M}^K) = \mathbb{E}[Y \mid do(X_1, \ldots, X_K)]$ is identifiable for any model obeying the Main Assumptions in Section 5 where $|\mathbf{X}| = K$, we have that $Q(\mathcal{M}^{K+1}) = \mathbb{E}[Y|do(X_1, \ldots, X_{K+1})]$ is also identifiable for any model where $|\mathbf{X}| = K + 1$.*

Proof of Claim 1.

Similar to the base case, we can write $Q(\mathcal{M}^{K+1})$ as

$$\begin{aligned}Q(\mathcal{M}^{K+1}) &= \mathbb{E}[Y \mid do(X_1 = x_1, \ldots, X_{K+1} = x_{K+1})]\\ &= f_y(\boldsymbol{X}^{K+1})\end{aligned}$$

From observational distribution $P^{\mathcal{M}^{K+1}}_{(\boldsymbol{X},Y)}$, we get

$$\begin{aligned}&\mathbb{E}[Y \mid X_1 = x_1, \ldots, X_{K+1} = x_{K+1}]\\ &= f_y(\boldsymbol{X}^{K+1}) + \mathbb{E}[U_y \mid X_1 = x_1, \ldots, X_{K+1} = x_{K+1}]\end{aligned}$$

We then only need to show that $\mathbb{E}[U_y \mid X_1 = x_1, \dots, X_{K+1} = x_{K+1}]$ is identifiable from the observational and *single-variable* interventional distributions.

From the additive Gaussian noise assumption, we have

$$\mathbb{E}[U_y \mid X_1 = x_1, \dots, X_{K+1} = x_{K+1}] = \Sigma_{u_y}\Sigma_{u_x}^{-1}\boldsymbol{u}_x$$

where $\boldsymbol{u}_x = [x_1 \ \cdots \ x_{K+1} - f_{K+1}(x_K)]^\top$, $\Sigma_{u_y} = [\sigma_{y1} \ \cdots \ \sigma_{y(K+1)}]$, $\Sigma_{u_x} = \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1(K+1)} \\ \vdots & \ddots & \vdots \\ \sigma_{(K+1)1} & \cdots & \sigma_{K+1}^2 \end{bmatrix}$. To identify Σ_{u_x} and $\sigma_{y1}, \dots, \sigma_{yK}$, we make use of the marginalization of a SCM defined in Bongers et al. (2016). Since $\mathcal{M}^{K+1}$ is acyclic (as we only consider an acyclic SCM), we can then marginalize out any subset of the endogenous variables associated with $\mathcal{M}^{K+1}$.

Identifying Σ_{u_x}

We marginalize Y and obtain the marginalization $\mathcal{M}^{K+1}_{marg(Y)}$. Since Y has no child node, marginalizing out Y has not effect on the rest of structural equations $\{f_1, \dots, f_{K+1}\}$. We can then treat X_{K+1} as a new Y and $\mathcal{M}^{K+1}_{marg(Y)}$ will now be equivalent to $\mathcal{M}^K$. Since we assume that Q($\mathcal{M}^K$) is identifiable, the covariance Σ_{u_x} is then identifiable.

Identifying $\sigma_{y1}, \dots, \sigma_{yK}$

We marginalize X_{K+1} and obtain the marginalization $\mathcal{M}^{K+1}_{marg(X_{K+1})}$ which preserves causal semantics of $\mathcal{M}^{K+1}$ and has the following structural equations

$$\begin{aligned} Y &= f_y(\boldsymbol{X}^K, f_{K+1}(\boldsymbol{X}^K) + U_{K+1}) + U_y \\ X_k &= f_k(\boldsymbol{X}^{k-1}) + U_k, \text{ for } k = 1, \dots, K. \end{aligned}$$

We then have that,

$$\begin{aligned} &\mathbb{E}[Y \mid X_1 = x_1, \dots, X_K = x_K] \\ &= \mathbb{E}[f_y(\boldsymbol{x}^K, f_{K+1}(\boldsymbol{x}^K) + U_{K+1})] \\ &\qquad + \mathbb{E}[U_y \mid X_1 = x_1, \dots, X_K = x_K]. \end{aligned}$$

Define $g_y(\boldsymbol{x}^K) := \mathbb{E}[f_y(\boldsymbol{x}^K, f_{K+1}(\boldsymbol{x}^K) + U_{K+1})]$, then

$$\begin{aligned} &\mathbb{E}[Y \mid X_1 = x_1, \dots, X_K = x_K] \\ &= g_y(\boldsymbol{x}^K) + \mathbb{E}[U_y \mid X_1 = x_1, \dots, X_K = x_K]. \end{aligned}$$

Since the model

$$\begin{aligned} Y &= g_y(\mathbf{X}^K) + U_y \\ X_k &= f_k(\boldsymbol{X}^{k-1}) + U_k, \text{ for } k = 1, \dots, K, \end{aligned}$$

satisfies the main assumptions where $|\mathbf{X}^K| = K$, by the induction step we have that $g_y(\mathbf{x}^K)$ is identifiable. It follows that $\mathbb{E}[U_y \mid X_1 = x_1, \dots X_K = x_K]$ is identifiable, and in turn $\sigma_{y1}, \dots \sigma_{yK}$ are identifiable. Next, we will show that $\sigma_{y(K+1)}$ is identifiable which will conclude the proof.

Identifying $\sigma_{y(K+1)}$

From the regime $P^{\mathcal{M}^{K+1}_{do(X_{K+1})}}_{(\boldsymbol{X},Y)}$, we have

$$\begin{aligned} &\mathbb{E}[Y \mid X_1 = x_1, \dots, do(X_{K+1} = x_{K+1})] \\ &= \mathbb{E}[f_y(\boldsymbol{X}^{K+1}) + U_y \mid X_1 = x_1, \dots, do(X_{K+1} = x_{K+1})] \\ &= f_y(\boldsymbol{X}^{K+1}) + \mathbb{E}[U_y \mid X_1 = x_1, \dots, X_K = x_K] \end{aligned}$$

The identifiability of $Q(\mathcal{M}^K)$ implies that f_{K+1} is identifiable. We can then choose $x_{K+1} = f_{K+1}(\boldsymbol{x}^K) + u_{K+1}$ for some u_{K+1} in the domain of U_{K+1} and take expectation over $p(\boldsymbol{X}^K)$,

$$\begin{aligned} &\mathbb{E}_{p(\boldsymbol{X}^K)}[\mathbb{E}[Y \mid X_1, \dots, X_K, do(X_{K+1} = x_{K+1})]] \\ &= \mathbb{E}[f_y(X_1, f_2(X_1) + U_2, \dots, f_{K+1}(\boldsymbol{X}^K) + u_{K+1})] \end{aligned} \tag{6}$$

From the regime $P^{\mathcal{M}^{K+1}_{do(X_K)}}_{(\boldsymbol{X},Y)}$, we have

$$\begin{aligned} &\mathbb{E}[Y \mid X_1 = x_1, \dots, do(X_K = x_K), U_{K+1} = u_{K+1}] \\ &= f_y(\boldsymbol{x}^K, f_{K+1}(\boldsymbol{x}^K) + u_{K+1}) \\ &\qquad + \mathbb{E}[U_y \mid X_1 = x_1, \dots, U_{K+1} = u_{K+1}] \end{aligned}$$

Taking expectation over $p(\boldsymbol{X}^K)$ yields,

$$\begin{aligned} &\mathbb{E}_{p(\boldsymbol{X}^K)}[\mathbb{E}[Y \mid X_1, \dots, X_K, U_{K+1} = u_{K+1})] \\ &= \mathbb{E}[f_y(X_1, f_2(X_1) + U_2, \dots, f_{K+1}(\boldsymbol{X}^K) + u_{K+1})] \\ &\qquad + \mathbb{E}[U_y \mid U_{K+1} = u_{K+1}] \end{aligned} \tag{7}$$

From (7)-(6), we can then get,

$$(7) - (6) = \mathbb{E}[U_y \mid U_{K+1} = u_{K+1}]$$

Finally, we can obtain $\sigma_{y(K+1)}$ by Lemma 1. □

Conclusion:

Since both the base case $K = 2$ and the inductive step are identifiable, by induction, the causal query $Q(\mathcal{M}^K)$ is identifiable for any integer $K \geq 2$.

□

2 UNIDENTIFIABILITY UNDER UNCONSTRAINED SCMs

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

Let $\ddot{\mathcal{M}}$ be an SCM with the following form,

$$\begin{aligned} Y &= X_1 \wedge X_2 \wedge U_y \\ X_2 &= X_1 \wedge U_2 \\ X_1 &= U_1 \end{aligned}$$

where $U_y = U_2 = U_1 \sim$ Bernoulli(p).

Let $P^{\ddot{\mathcal{M}}}$ be the joint distribution induced by $\ddot{\mathcal{M}}$, we have

$$P^{\ddot{\mathcal{M}}}(y, x_1, x_2) = \begin{cases} p, & \text{if } (y, x_1, x_2) = (1,1,1) \\ 1-p, & \text{if } (y, x_1, x_2) = (0,0,0) \\ 0, & \text{otherwise} \end{cases}$$

Intervening on X_1 results in the SCM $\ddot{\mathcal{M}}_{do(X_1=x_1)}$ with the following form,

$$\begin{aligned} Y &= X_1 \wedge X_2 \wedge U_y \\ X_2 &= X_1 \wedge U_2 \\ X_1 &= x_1 \end{aligned}$$

with the interventional joint distribution $P^{\ddot{\mathcal{M}}_{do(X_1=x_1)}}$,

when $x_1 = 1$,

$$P^{\ddot{\mathcal{M}}_{do(X_1=1)}}(y, x_2) = \begin{cases} p, & \text{if } (y, x_2) = (1,1) \\ 1-p, & \text{if } (y, x_2) = (0,0) \\ 0, & \text{otherwise} \end{cases}$$

when $x_1 = 0$,

$$P^{\ddot{\mathcal{M}}_{do(X_1=0)}}(y, x_2) = \begin{cases} 1, & \text{if } (y, x_2) = (0,0) \\ 0, & \text{otherwise} \end{cases}$$

Intervening on X_2 results in the SCM $\ddot{\mathcal{M}}_{do(X_2=x_2)}$ with the following form,

$$\begin{aligned} Y &= X_1 \wedge X_2 \wedge U_y \\ X_2 &= x_2 \\ X_1 &= U_1 \end{aligned}$$

with the interventional joint distribution $P^{\ddot{\mathcal{M}}_{do(X_2=x_2)}}$,

when $x_2 = 1$,

$$P^{\ddot{\mathcal{M}}_{do(X_2=1)}}(y, x_1) = \begin{cases} p, & \text{if } (y, x_1) = (1,1) \\ 1-p, & \text{if } (y, x_1) = (0,0) \\ 0, & \text{otherwise} \end{cases}$$

when $x_2 = 0$,

$$P^{\ddot{\mathcal{M}}_{do(X_2=0)}}(y, x_1) = \begin{cases} p, & \text{if } (y, x_1) = (0,1) \\ 1-p, & \text{if } (y, x_1) = (0,0) \\ 0, & \text{otherwise} \end{cases}$$

Let $\bar{\mathcal{M}}$ be another SCM with the following form,

$$\begin{aligned} Y &= X_2 \wedge U_y \\ X_2 &= X_1 \wedge U_2 \\ X_1 &= U_1 \end{aligned}$$

where $U_y = U_2 = U_1 \sim$ Bernoulli(p)

Let $P^{\bar{\mathcal{M}}}$ be the joint distribution induced by $\bar{\mathcal{M}}$, we have

$$P^{\bar{\mathcal{M}}}(y, x_1, x_2) = \begin{cases} p, & \text{if } (y, x_1, x_2) = (1,1,1) \\ 1-p, & \text{if } (y, x_1, x_2) = (0,0,0) \\ 0, & \text{otherwise} \end{cases}$$

Intervening on X_1 results in the SCM $\bar{\mathcal{M}}_{do(X_1=x_1)}$ with the following form,

$$\begin{aligned} Y &= X_2 \wedge U_y \\ X_2 &= X_1 \wedge U_2 \\ X_1 &= x_1 \end{aligned}$$

with the interventional joint distribution $P^{\bar{\mathcal{M}}_{do(X_1=x_1)}}$,

when $x_1 = 1$,

$$P^{\bar{\mathcal{M}}_{do(X_1=1)}}(y, x_2) = \begin{cases} p, & \text{if } (y, x_2) = (1,1) \\ 1-p, & \text{if } (y, x_2) = (0,0) \\ 0, & \text{otherwise} \end{cases}$$

when $x_1 = 0$,

$$P^{\bar{\mathcal{M}}_{do(X_1=0)}}(y, x_2) = \begin{cases} 1, & \text{if } (y, x_2) = (0,0) \\ 0, & \text{otherwise} \end{cases}$$

Intervening on X_2 results in the SCM $\bar{\mathcal{M}}_{do(X_2=x_2)}$ with the following form,

$$\begin{aligned} Y &= X_2 \wedge U_y \\ X_2 &= x_2 \\ X_1 &= U_1 \end{aligned}$$

with the interventional joint distribution $P^{\bar{\mathcal{M}}_{do(X_2=x_2)}}$, when $x_2 = 1$,

$$P^{\bar{\mathcal{M}}_{do(X_2=1)}}(y, x_1) = \begin{cases} p, & \text{if } (y, x_1) = (1, 1) \\ 1-p, & \text{if } (y, x_1) = (0, 0) \\ 0, & \text{otherwise} \end{cases}$$

when $x_2 = 0$,

$$P^{\bar{\mathcal{M}}_{do(X_2=0)}}(y, x_1) = \begin{cases} p, & \text{if } (y, x_1) = (0, 1) \\ 1-p, & \text{if } (y, x_1) = (0, 0) \\ 0, & \text{otherwise} \end{cases}$$

We can see that the given two SCMs yield the same observational distribution as well as interventional distributions when intervening on each of the treatment variables (i.e. $P^{\ddot{\mathcal{M}}} = P^{\bar{\mathcal{M}}}$, $P^{\ddot{\mathcal{M}}_{do(X_1)}} = P^{\bar{\mathcal{M}}_{do(X_1)}}$ and $P^{\ddot{\mathcal{M}}_{do(X_2)}} = P^{\bar{\mathcal{M}}_{do(X_2)}}$). However, they yield different joint interventional distributions,

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

Hence, the effect of joint interventions is not identifiable under unconstrained SCMs.

3 ADDITIONAL BASELINE

In addition to the direct regression baseline (REG) described in the main paper, we also consider another baseline where the regime indicators are used in the regression model. We refer to this baseline as REG_IND. Specifically, REG_IND models the conditional expectation $\mathbb{E}[Y \mid \mathbf{PA}_Y]$ as follow,

$$\mathbb{E}[Y \mid \mathbf{PA}_Y = \mathbf{x}] = f_Y(\mathbf{x}; \boldsymbol{\theta}_0) + \sum_{k=1}^{K} z_k f_Y(\mathbf{x}; \boldsymbol{\theta}_k)$$

where z_k is the regime indicator variable; $z_k = 1$ if the treatment variable X_k is intervened on, otherwise $z_k = 0$. For example, if the observation $(\mathbf{x}, y)$ is sampled from $P^{\mathcal{M}_{do(X_1)}}_{(\mathbf{X},Y)}$, then $z_1 = 1$ and $z_k = 0$ for all $k \neq 1$.

Figure 1 illustrates the results of the experiment described in Section 6.2.2 with the additional baseline (REG_IND). It is clear from the plots that REG_IND is consistently worse than the other baseline (REG); hence, we decided to not include this additional baseline in the main paper.

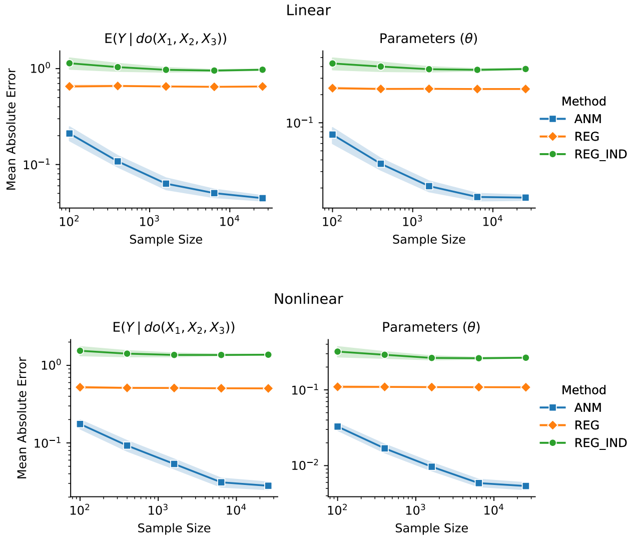

Figure 1: MAE of the predicted joint effect (left) and the parameter estimates (right) as the sample size increases. The solid lines represent the mean absolute error averaged over 50 Monte Carlo experiments and the filled regions depict its 95% confidence interval (note that both vertical and horizontal axes are in logarithmic scale).

4 SYNTHETIC EXPERIMENT

In this section, we provide more details on the data generating process used in the synthetic experiment. Furthermore, we present additional experiment results on the synthetic data where we compare the performance of our approach to the baseline.

4.1 Data Generating Process

We first define a data generating process, from which we will simulate observational and interventional samples. In the synthetic experiment, we consider the case where the number of treatments $K = 3$. The pre-defined 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. Let $\boldsymbol{x} \in \mathbb{R}^d$ be a d-dimensional input vector, the linear functions are simply defined as,

$$f_k(\boldsymbol{x}; \boldsymbol{\theta}) := \sum_{i}^{d} \theta_i^k x_i$$

For the nonlinear ones, we add second-order interactions in addition to the main effects. Let $\phi : \mathbb{R}^d \to \mathbb{R}^{d+\binom{d}{2}}$ be

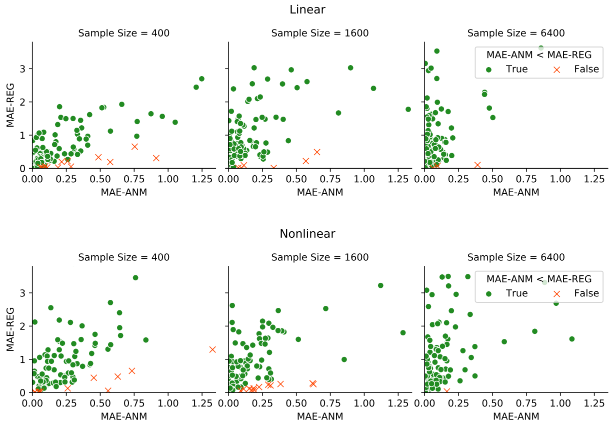

Figure 2: The prediction accuracy of `ANM` vs `REG` as the sample size increases. Each data point represents one Monte Carlo experiment in which the functions and covariance matrix of the underlying SCM are randomly generated.

a feature map defined by,

$$\phi(\boldsymbol{x}) = [x_1, \dots, x_d, x_1x_2, \dots, x_1x_d, x_2x_3, \\ \dots, x_2x_d, \dots, x_{d-1}x_d]$$

Our nonlinear functions are then defined as,

$$f_k(\boldsymbol{x}; \boldsymbol{\theta}) := \sum_{i}^{d+\binom{d}{2}} \theta_i^k \phi(\boldsymbol{x})_i$$

For assessing consistency and unbiasedness, we fix an SCM with the following parameters

Covariance Matrix:

$$\Sigma_u = \begin{bmatrix} 1. & 0.3 & 0.8 & -0.6 \\ 0.3 & 1. & 0.3 & -0.5 \\ 0.8 & 0.3 & 1. & -0.5 \\ -0.6 & -0.5 & -0.5 & 1. \end{bmatrix}$$

Linear functions:

$$\boldsymbol{\theta}_2 = [1.0] \,, \boldsymbol{\theta}_3 = [0.5 \quad -1.0] \,, \boldsymbol{\theta}_y = [1.5 \quad 1.0 \quad -0.5]$$

Nonlinear functions:

$$\boldsymbol{\theta}_2 = [1.0] \,, \boldsymbol{\theta}_3 = [0.5 \quad -1.0 \quad 1.0] \,, \\ \boldsymbol{\theta}_y = [1.5 \quad 1.0 \quad -0.5 \quad 0.5 \quad -1 \quad -1.5]$$

4.2 Assessing the accuracy of the causal predictions

We compare the performance of our model with the baseline where we measure the accuracy of the predicted joint interventional effects on multiple different SCMs. Instead of fixing the parameters and the covariance matrix of the underlying SCM, we now re-sample them in every iteration. To generate a random SCM in each iteration, we generate a random covariance matrix $\Sigma_u = WW^\top + diag(\boldsymbol{v})$ where W is a random matrix of size $D \times K$ with $D < K$ and $\boldsymbol{v}$ is a K-dimensional vector. Each element of W and $\boldsymbol{v}$ is drawn from the standard normal distribution. In addition, the parameters of structural equations are drawn from an independent normal distribution $\boldsymbol{\theta}_i \sim \mathcal{N}(0, 1.5I)$. We then sample 100 uniform random test points and compute the mean absolute error of each model with respect to those test points. Figure 2 compares the performance of the `ANM` and the `REG` models across 100 Monte Carlo experiments. In short, `ANM` consistently outperforms `REG` in terms of the mean absolute error on the test data. Furthermore, the performance gap between the two models is larger when we increase the sample size (the number of simulations in which `REG` outperforms `ANM` are smaller as n_{sample} increases). With the sample size of 6400, `ANM` outperforms `REG` approximately 99% of the time. This is simply because our model has lower variance when we have more training data.

References

Bongers, S., Peters, J., Schölkopf, B., and Mooij, J. M. Theoretical aspects of cyclic structural causal models. *arXiv preprint arXiv:1611.06221*, 2016.