Disentangled Representation for Causal Mediation Analysis

Ziqi Xu[1], **Debo Cheng**[1,2*], **Jiuyong Li**[1*], **Jixue Liu**[1], **Lin Liu**[1], **Ke Wang**[3]

[1] University of South Australia
[2] Guangxi Normal University
[3] Simon Fraser University

ziqi.xu@mymail.unisa.edu.au,{Jiuyong.li,Jixue.Liu,Lin.Liu}@unisa.edu.au,chengdb2016@gmail.com,wangk@cs.sfu.ca

Abstract

Estimating direct and indirect causal effects from observational data is crucial to understanding the causal mechanisms and predicting the behaviour under different interventions. Causal mediation analysis is a method that is often used to reveal direct and indirect effects. Deep learning shows promise in mediation analysis, but the current methods only assume latent confounders that affect treatment, mediator and outcome simultaneously, and fail to identify different types of latent confounders (e.g., confounders that only affect the mediator or outcome). Furthermore, current methods are based on the sequential ignorability assumption, which is not feasible for dealing with multiple types of latent confounders. This work aims to circumvent the sequential ignorability assumption and applies the piecemeal deconfounding assumption as an alternative. We propose the Disentangled Mediation Analysis Variational AutoEncoder (DMAVAE), which disentangles the representations of latent confounders into three types to accurately estimate the natural direct effect, natural indirect effect and total effect. Experimental results show that the proposed method outperforms existing methods and has strong generalisation ability. We further apply the method to a real-world dataset to show its potential application.

Introduction

Causal mediation analysis (CMA) aims to reveal how other attributes mediate the causal effect of a treatment on the outcome. It can help understand how the mechanism works and enable outcome prediction under various interventions in practice. In recent years, CMA has been widely used in operational decision (Yin and Hong 2019) and policy evaluation (Cheng, Guo, and Liu 2022). For example, in the UC Berkeley admission process, it was found that female applicants were more likely to be rejected than male applicants (Bickel, Hammel, and O'Connell 1975). However, investigators through CMA understood that the choice of disciplines, as a mediator was affected by gender and affected the acceptance rate. There was in fact no discrimination because female applicants preferred to apply for admission into disciplines with a higher level of competition and lower acceptance rates.
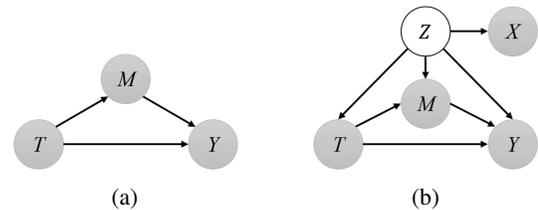
Figure 1: (a) A causal graph for the UC Berkeley admission process example. T denotes gender (treatment), M denotes discipline choice (mediator), and Y denotes acceptance rate (outcome). (b) Causal graph for CMAVAE (Cheng, Guo, and Liu 2022). T denotes treatment, M denotes mediator, Y denotes outcome, X denotes proxy attributes and Z denotes the representation of latent confounders which affect T, M and Y simultaneously.

The above example illustrates three essential ingredients of CMA, the natural direct effect (NDE), the natural indirect effect (NIE) and the total effect (TE), respectively. We use Figure 1a to illustrate the above example in terms of the three types of effects. Here, TE measures the expected increase in the acceptance rate as gender changes from male to female; NDE measures the expected increase in acceptance rate as gender changes from male to female while setting the discipline choice to whatever value it would have attained prior to the change (i.e., under gender being male); NIE measures the expected increase in acceptance rate when gender is male, and discipline choice changes to whatever value it would have attained under gender being female (Pearl 2014). To be fair for all applicants, the NDE should be zero since gender should not directly affect the acceptance rate, while the NIE is allowed to be non-zero if discipline choice is considered a non-sensitive attribute.

The major challenge in CMA is latent confounders. In the above example, an applicant's economic status affects his or her discipline choice and acceptance rate, and without measuring it, we cannot estimate NDE, NIE or TE. Fortunately, leveraging the strength of representation learning, researchers are able to learn the representations of latent confounders from other attributes. For example, the representation of the applicants' economic status can be learned from the collected information of their addresses and parental
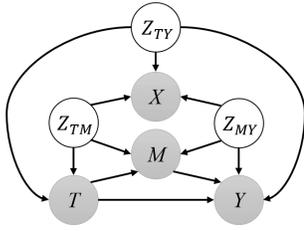
Figure 2: The causal graph for the proposed Disentangled Mediation Analysis Variational AutoEncoder (DMAVAE). T is the treatment, M is the mediator, Y is the outcome and X represents the set of the proxy attributes. Z_{TM}, Z_{TY} and Z_{MY} denote the disentangled representations of the three types of latent confounders.

information. Deep learning has made significant progress in representation learning in the past decade. For example, Kingma and Welling (2014) proposed Variational AutoEncoder (VAE) to learn representations with neural networks. Louizos et al. (2017) first combined causal inference and VAE and proposed CEVAE. CEVAE can learn the representation of latent confounders and inference individual treatment effect (ITE) and average treatment effect (ATE). Zhang, Liu, and Li (2021) proposed TEDVAE, which disentangles latent representations into three parts to achieve more accurate estimation of causal effects than CEVAE does. However, both CEVAE and TEDVAE do not consider mediator and cannot estimate NDE or NIE.

Cheng, Guo, and Liu (2022) proposed CMAVAE, which considers mediator and the estimation of NDE and NIE. It is the first work that combines deep learning and CMA, and CMAVAE assumes that the latent confounders affect treatment, mediator and outcome simultaneously as shown in Figure 1b. Their work extended sequential ignorability (Imai, Keele, and Tingley 2010) to representation learning and used proxy attributes X to learn the representation of latent confounders. However, CMAVAE has limited generalisation ability since it does not consider different types of confounders, which, however, are often seen in practice. For example, the demand of the job market can be considered as a confounder that affects M (discipline choice) and Y (acceptance rate) in Figure 1a, because the popularity of a job area affects students' choice of disciplines, the university may need to limit the acceptance rate due to the limitation of teaching resources. Furthermore, this confounder does not affect an applicant's gender. CMAVAE cannot handle this case since it assumes a confounder that affects treatment (gender in this example), mediator and outcome simultaneously.

In this paper, we consider a general case with three types of latent confounders as shown in Figure 2, where Z_{TM} is the representation of the latent confounders that affect treatment and mediator; Z_{TY} is the representation of the latent confounders that affect treatment and outcome; and Z_{MY} is the representation of the latent confounders that affect mediator and outcome. We can access the observed proxy attributes X and assume that the proxy attributes

and latent confounders are inherently correlated (Angrist and Pischke 2009). Our proposed disentanglement scheme in Figure 2 follows the piecemeal deconfounding assumption (Pearl 2014), which is a more relaxed assumption than sequential ignorability. We propose a novel method, namely Disentangled Mediation Analysis Variational AutoEncoder (DMAVAE) to learn the disentangled representations of the three types of latent confounders as illustrated in Figure 2 by using a VAE-based approach, for estimating NDE, NIE and TE. We make the following contributions in this paper:

- We study a general case of CMA under the piecemeal deconfounding assumption, where there are three types of latent confounders.

- We propose a novel CMA method, DMAVAE, to learn the disentangled representations of the three types of latent confounders from proxy attributes to more accurately estimate NDE, NIE and TE.

- We evaluate the effectiveness of the DMAVAE method on synthetic datasets. Experiments show that DMAVAE outperforms existing CMA methods and has a strong generalisation ability. Furthermore, we conduct a case study to show the application scenarios of DMAVAE.

Background

We use upper case letters to represent attributes and lower case letters to represent the values of the attributes. We follow the work in (Pearl 2014) to introduce the notations, definitions and assumptions.

Notations and Definitions

Let T_i represent the treatment status of the i-th individual, with $T_i = 1$ indicating the i-th individual receiving the treatment and $T_i = 0$ otherwise. We use M_i to represent the mediator for the i-th individual, $M_i(t)$ for the value of the mediator when T_i is fixed to t, where $t \in \{0, 1\}$. Similarly, Y_i is used to represent the outcome of the i-th individual and $Y_i(t, m)$ for the outcome when $T_i = t$ and $M_i = m$. Finally, we use X_i to denote the set of proxy attributes. When the context is clear we omit the subscript i.

Definition 1 (Total Effect (Pearl 2014)).

$$\text{TE} = \mathbb{E}[Y(1, M(1)) - Y(0, M(0))]. \tag{1}$$

As stated in (Pearl 2014), TE is the measure of "the expected increase in the outcome Y when the treatment changes from $T = 0$ to $T = 1$, while the mediator is allowed to track the change in T."

Definition 2 (Natural Direct Effect (Pearl 2014)).

$$\text{NDE} = \mathbb{E}[Y(1, M(0)) - Y(0, M(0))]. \tag{2}$$

Based on (Pearl 2014), NDE is the measure of "the expected increase in the outcome Y that the treatment changes from $T = 0$ to $T = 1$, while setting the mediator to whatever value it would have attained (for each individual) prior to the change, i.e., under $T = 0$."

Definition 3 (Natural Indirect Effect (Pearl 2014)).

$$\text{NIE} = \mathbb{E}[Y(0, M(1)) - Y(0, M(0))]. \tag{3}$$

As explained in (Pearl 2014), NIE is the measure of "the expected increase in the outcome Y when the treatment is held constant, at $T = 0$, and M changes to whatever value it would have attained (for each individual) under $T = 1$."

Additionally, TE can be decomposed as follows:

$$\text{TE} = \text{NDE} - \text{NIE}_r, \tag{4}$$

where NIE_r denotes the reverse version of NIE, as follows:

$$\text{NIE}_r = \mathbb{E}[Y(1, M(0)) - Y(1, M(1))]. \tag{5}$$

This implies that TE is identifiable whenever NDE and NIE_r are identifiable (Pearl 2014). Note that NDE and NIE are defined in counterfactual expressions. To allow these counterfactual expressions to be estimated, Pearl (2014) proposed three steps for the counterfactual derivation of NDE and NIE: First, we find an adjustment set under suitable assumptions; Second, we reduce all counterfactuals to do-expressions by using the adjustment set found; Finally, we identify the do-expressions from the data.

Assumptions

Assumption 1 (Sequential Ignorability (Imai, Keele, and Tingley 2010))**.** *There exists a set W of measured attributes such that:*

1. *W and T deconfound the mediator-outcome relationship, keeping T fixed: $[Y(t', m) \perp\!\!\!\perp M(t) | T = t, W = w]$;*
2. *W deconfounds the treatment-{mediator, outcome} relationship: $[T \perp\!\!\!\perp (Y(t', m), M(t)) | W = w]$,*

where $0 < p(M = m | do(T = t, W = w)) < 1$, $0 < p(Y = y | do(T = t, M = m, W = w)) < 1$, and $t, t' \in \{0, 1\}$.

Pearl (2014) proposed the following assumption, which is more relaxed than Assumption 1.

Assumption 2 (Piecemeal Deconfounding (Pearl 2014))**.** *There exists three sets of measured attributes $W = \{W_1, W_2, W_3\}$ such that:*

1. *No member of W_1 is affected by the treatment;*
2. *W_1 deconfounds the $M \rightarrow Y$ relationship (holding T constant);*
3. *$\{W_2, W_1\}$ deconfounds the $T \rightarrow M$ relationship;*
4. *$\{W_3, W_1\}$ deconfounds the $\{T, M\} \rightarrow Y$ relationship.*

We note that Assumption 2 follows the idea of divide-and-conquer, and allows the use of different sets of attributes to deconfound the mediator and outcome processes separately, rather than having to use the same set of deconfounding attributes as required by Assumption 1 (Pearl 2014).

Method

CMA via Representations

As mentioned above, the piecemeal deconfounding assumption is more relaxed than the sequential ignorability assumption. Our goal is to extend the piecemeal deconfounding assumption to disentangled representation learning. For this, we propose the causal graph for the disentanglement as shown in Figure 2, and introduce the following assumption for the disentangled representations.

Assumption 3 (Assumption for DMAVAE)**.**

1. *There exist three types of confounders, i.e., Z_{TM}, Z_{TY} and Z_{MY} as shown in Figure 2.*
2. *There exist proxy attributes X that approximate Z_{TM}, Z_{TY} and Z_{MY}.*
3. *$p(T, M, Y, X, Z_{TM}, Z_{TY}, Z_{MY})$ can be recovered from all observed attributes (T, M, Y, X).*

For our proposed disentanglement, we have the following theorem.

Theorem 1. *The disentangled representation Z_{MY} following Figure 2 satisfies Conditions 1 and 2 of Assumption 2.*

Proof. Z_{MY} is a non-descendant of T as shown in Figure 2, hence Condition 1 of Assumption 2 is satisfied (considering Z_{MY} as W_1). All the back-door paths between M and Y when holding T constant are blocked by Z_{MY}, i.e., $M \leftarrow Z_{MY} \rightarrow Y$ is blocked by Z_{MY}, $M \leftarrow Z_{MY} \rightarrow X \leftarrow Z_{TY} \rightarrow Y$, $M \leftarrow Z_{TM} \rightarrow X \leftarrow Z_{TY} \rightarrow Y$ and $M \leftarrow Z_{TM} \rightarrow X \leftarrow Z_{MY} \rightarrow Y$ are all blocked by $\emptyset$ since X is a collider. Hence, Condition 2 of Assumption 2 is also satisfied when considering Z_{MY} as W_1. $\square$

Based on Theorem 1 and Equation 13 in (Pearl 2014), the counterfactual expression of NDE (Equation 2) and NIE_r (Equation 5) can be reduced to do-expressions as follows:

$$
\begin{aligned}
\text{NDE} = & [\mathbb{E}(Y | do(T = 1, M = m), Z_{MY} = z_{MY}) \\
& - \mathbb{E}(Y | do(T = 0, M = m), Z_{MY} = z_{MY})] \\
& \times p(M = m | do(T = 0), Z_{MY} = z_{MY}) \\
& \times p(Z_{MY} = z_{MY}),
\end{aligned} \tag{6}
$$

$$
\begin{aligned}
\text{NIE}_r = & \ \mathbb{E}(Y | do(T = 1, M = m), Z_{MY} = z_{MY}) \\
& \times [p(M = m | do(T = 0), Z_{MY} = z_{MY}) \\
& - p(M = m | do(T = 1), Z_{MY} = z_{MY})] \\
& \times p(Z_{MY} = z_{MY}),
\end{aligned} \tag{7}
$$

Following Pearl's back-door adjustment formula (Pearl 2009a,b), we propose the following theorem which is used to reduce the do-expressions to probability expressions.

Theorem 2. *If we can estimate $p(T, M, Y, X, Z_{TM}, Z_{TY}, Z_{MY})$ under the causal graph in Figure 2, NDE and NIE_r can be identified from data as follows:*

$$
\begin{aligned}
\text{NDE} = & [\mathbb{E}(Y | T = 1, M = m, Z_{MY} = z_{MY}, Z_{TY} = z_{TY}) \\
& - \mathbb{E}(Y | T = 0, M = m, Z_{MY} = z_{MY}, Z_{TY} = z_{TY})] \\
& \times p(M = m | T = 0, Z_{MY} = z_{MY}, Z_{TM} = z_{TM}) \\
& \times p(Z_{MY} = z_{MY}, Z_{TM} = z_{TM}, Z_{TY} = z_{TY}),
\end{aligned} \tag{8}
$$

$$
\begin{aligned}
\text{NIE}_r = & \ \mathbb{E}(Y | T = 1, M = m, Z_{MY} = z_{MY}, Z_{TY} = z_{TY}) \\
& \times [p(M = m | T = 0, Z_{MY} = z_{MY}, Z_{TM} = z_{TM}) \\
& - p(M = m | T = 1, Z_{MY} = z_{MY}, Z_{TM} = z_{TM})] \\
& \times p(Z_{MY} = z_{MY}, Z_{TM} = z_{TM}, Z_{TY} = z_{TY}).
\end{aligned} \tag{9}
$$

Proof. We prove that Equations 6 and 7 can be identified, i.e., the do-expressions can be converted to the do-free expressions in Equations 8 and 9, respectively. For this, we prove that $p(M = m | do(T = 0), Z_{MY} = z_{MY})$ and $p(Y = y | do(T = 0, M = m), Z_{MY} = z_{MY})$ are identifiable. The case for $T = 1$ is identical and the expectation
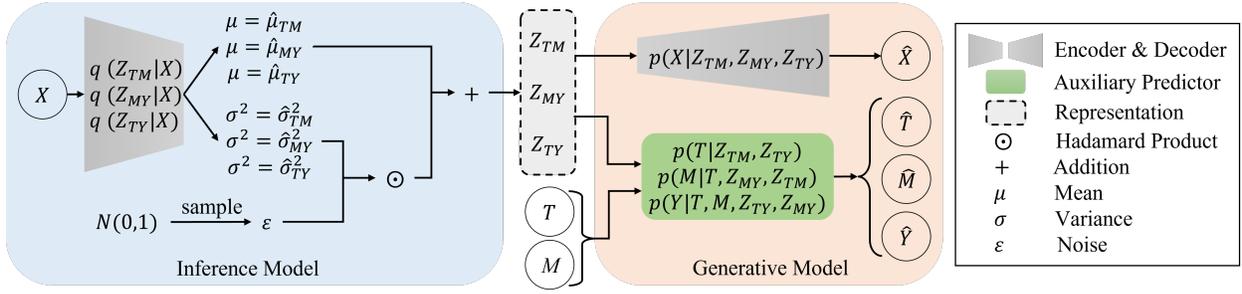
Figure 3: The architecture of Disentangled Mediation Analysis Variational AutoEncoder (DMAVAE).

expressions with the do-operators are identifiable when the corresponding probability expressions are identifiable. Once NDE and NIE$_r$ are identified, TE is calculated as shown in Equation 4.

$p(M = m|do(T = 0), Z_{MY} = z_{MY})$ is identifiable by adjusting Z_{TM}, since all the back-door paths between T and M when conditioning on Z_{MY} are blocked by Z_{TM}, i.e., $T \leftarrow Z_{TM} \rightarrow M$ is blocked by Z_{TM}, $T \leftarrow Z_{TM} \rightarrow X \leftarrow Z_{TY} \rightarrow Y \leftarrow M$, $T \leftarrow Z_{TY} \rightarrow X \leftarrow Z_{TM} \rightarrow M$ and $T \leftarrow Z_{TY} \rightarrow Y \leftarrow M$ are blocked by $\emptyset$ since X or Y is a collider. Hence, $p(M = m|do(T = 0), Z_{MY} = z_{MY}) = p(M = m|T = 0, Z_{MY} = z_{MY}, Z_{TM} = z_{TM})$.

$p(Y = y|do(T = 0, M = m), Z_{MY} = z_{MY})$ is identifiable by adjusting Z_{TY}, since all the back-door paths between $\{T, M\}$ and Y when conditioning on Z_{MY} are blocked by Z_{TY}, i.e., $T \leftarrow Z_{TY} \rightarrow Y$ is blocked by Z_{TY}, $T \leftarrow Z_{TM} \rightarrow X \leftarrow Z_{TY} \rightarrow Y$ and $M \leftarrow Z_{TM} \rightarrow X \leftarrow Z_{TY} \rightarrow Y$ are blocked by $\emptyset$ since X is a collider. Hence, $p(Y = y|do(T = 0, M = m), Z_{MY} = z_{MY}) = p(Y = y|T = 0, M = m, Z_{MY} = z_{MY}, Z_{TY} = z_{TY})$. □

Learning Disentangled Representations

In the previous discussion, we have assumed that X contains (or captures) the information of the latent confounders, which can be disentangled into three types, i.e., Z_{TM}, Z_{TY} and Z_{MY}. With these representations, we can estimate NDE and NIE$_r$ by Equations 8 and 9, respectively.

To learn these representations, we propose DMAVAE based on Variational Autoencoder (VAE) (Kingma and Welling 2014), which uses variational inference to learn representations. DMAVAE parameterises the causal graph in Figure 2 as a model with neural network functions connecting the attributes of interest. The objective function of DMAVAE is the reconstruction error of the observed (X, T, M, Y) and the inferred $(\hat{X}, \hat{T}, \hat{M}, \hat{Y})$. The architecture of DMAVAE is shown in Figure 3. We note that Y is not shown in Figure 3 since Y is not used as an input to the generative model.

In the inference model, we design three separate encoders $q(Z_{TM}|X)$, $q(Z_{MY}|X)$ and $q(Z_{TY}|X)$ that serve as variational posteriors over the three representations. The variational approximations of the posteriors are defined as:

$$q(Z_{TM}|X) = \prod_{j=1}^{D_{Z_{TM}}} \mathcal{N}(\mu = \hat{\mu}_{TM_j}, \sigma^2 = \hat{\sigma}^2_{TM_j});$$

$$q(Z_{MY}|X) = \prod_{j=1}^{D_{Z_{MY}}} \mathcal{N}(\mu = \hat{\mu}_{MY_j}, \sigma^2 = \hat{\sigma}^2_{MY_j});$$

$$q(Z_{TY}|X) = \prod_{j=1}^{D_{Z_{TY}}} \mathcal{N}(\mu = \hat{\mu}_{TY_j}, \sigma^2 = \hat{\sigma}^2_{TY_j}),$$

where $\hat{\mu}_{TM}, \hat{\mu}_{MY}, \hat{\mu}_{TY}$ and $\hat{\sigma}^2_{TM}, \hat{\sigma}^2_{MY}, \hat{\sigma}^2_{TY}$ are the means and variances of the Gaussian distributions parameterised by neural networks. $D_{Z_{TM}}$, $D_{Z_{MY}}$ and $D_{Z_{TY}}$ are the dimensions of Z_{TM}, Z_{TM} and Z_{TY}, respectively.

The generative models for T and X are defined as:
$$p(T|Z_{TM}, Z_{TY}) = \text{Bern}(\sigma(g(Z_{TM}, Z_{TY})));$$

$$p(X|Z_{TM}, Z_{TY}, Z_{MY}) = \prod_{j=1}^{D_X} p(X_j|Z_{TM}, Z_{TY}, Z_{MY}),$$

where $g(\cdot)$ is a neural network parameterised by its own parameters, $\sigma(\cdot)$ is the logistic function and D_X is the dimension of X.

Specifically, the generative models for M and Y vary depending on the types of the attribute values. For continuous M and Y, the models are defined as:

$$p(M|T, Z_{TM}, Z_{MY}) = \mathcal{N}(\mu = \hat{\mu}_M, \sigma^2 = \hat{\sigma}^2_M);$$
$$\hat{\mu}_M = (Tg(Z_{TM}, Z_{MY}) + (1 - T)g(Z_{TM}, Z_{MY}));$$
$$\hat{\sigma}^2_M = (Tg(Z_{TM}, Z_{MY}) + (1 - T)g(Z_{TM}, Z_{MY})),$$

$$p(Y|T, M, Z_{TY}, Z_{MY}) = \mathcal{N}(\mu = \hat{\mu}_Y, \sigma^2 = \hat{\sigma}^2_Y);$$
$$\hat{\mu}_Y = (Tg(M, Z_{TY}, Z_{MY}) + (1 - T)g(M, Z_{TY}, Z_{MY}));$$
$$\hat{\sigma}^2_Y = (Tg(M, Z_{TM}, Z_{MY}) + (1 - T)g(M, Z_{TM}, Z_{MY})),$$

For binary M and Y the models are defined as:

$$p(M|T, Z_{TM}, Z_{MY}) = \text{Bern}(\sigma(g(T, Z_{TM}, Z_{MY})));$$
$$p(Y|T, M, Z_{TY}, Z_{MY}) = \text{Bern}(\sigma(g(T, M, Z_{TY}, Z_{MY}))).$$

Following the same setting in (Kingma and Welling 2014), we choose Gaussian distribution as the prior distributions of the representations, which are defined as:

$$p(Z_{TM}) = \prod_{j=1}^{D_{Z_{TM}}} \mathcal{N}(z_{TM_j}|0, 1);$$

$$p(Z_{MY}) = \prod_{j=1}^{D_{Z_{MY}}} \mathcal{N}(z_{MY_j}|0, 1);$$

$$p(Z_{TY}) = \prod_{j=1}^{D_{Z_{TY}}} \mathcal{N}(z_{TY_j}|0, 1).$$

We can now form the evidence lower bound (ELBO) for the above inference and generative networks:

$$
\begin{aligned}
\mathcal{M} = \ &\mathbb{E}_q[\log p(X|Z_{TM}, Z_{TY}, Z_{MY})] \\
&- D_{\text{KL}}[q(Z_{TM}|X)||p(Z_{TM})] \\
&- D_{\text{KL}}[q(Z_{MY}|X)||p(Z_{MY})] \\
&- D_{\text{KL}}[q(Z_{TY}|X)||p(Z_{TY})],
\end{aligned}
\tag{10}
$$

where $\mathbb{E}_q$ stands for $\mathbb{E}_{q(Z_{TM}|X)q(Z_{MY}|X)q(Z_{TY}|X)}$. The first term denotes the reconstruction loss between X and $\hat{X}$; each of the other terms is used to calculate the KL divergence between the prior knowledge and the learned representations.

Following the works in (Louizos et al. 2017; Zhang, Liu, and Li 2021), we design three auxiliary predictors to encourage the disentanglement of the representations. To optimise the parameters in the auxiliary predictors, we add them to the loss function. Since maximising $\mathcal{M}$ is equal to minimising $-\mathcal{M}$, the final loss function of DMAVAE is defined as:

$$\begin{aligned}
\mathcal{L}_{\text{DMAVAE}} = & -\mathcal{M} + \log q(T|Z_{TM}, Z_{TY}) \\
& + \log q(M|T, Z_{\text{TM}}, Z_{\text{MY}}) \\
& + \log q(Y|T, M, Z_{TY}, Z_{MY}).
\end{aligned} \quad (11)$$

Experiments

Evaluating estimated causal effects has always been a huge challenge because we have no access to ground truth in practice (Guo et al. 2020). Evaluating the results of CMA is considered a more difficult task, as it requires accurate estimation of causal effects with respect to different paths. Therefore the evaluation of causal effect estimation and CMA mainly relies on simulated datasets (Huber, Lechner, and Mellace 2016). Following the works in (Louizos et al. 2017; Cheng, Guo, and Liu 2022), we construct simulated datasets to evaluate the performance of DMAVAE and compare it with the baseline methods. We also evaluate the generalisation ability of DMAVAE. Finally, we apply DMAVAE to a real-world dataset and verify its usability. The code is available at https://github.com/IRON13/DMAVAE.

Experiment Setup

Methods for Comparison We compare DMAVAE with traditional and deep learning-based CMA methods. For traditional CMA methods, we select the commonly used parametric methods LSEM (Baron and Kenny 1986), LSEM-I (Imai, Keele, and Tingley 2010), NEW-I, NEW-W (Lange, Vansteelandt, and Bekaert 2012), and semi-parametric method IPW (Robins, Rotnitzky, and Zhao 1994; Huber 2014). For deep learning-based CMA methods, we compare our method with CMAVAE (Cheng, Guo, and Liu 2022), which is the only VAE-based CMA method. Since there are no other deep learning-based CMA methods, we select two well-known causal effect estimators (CEVAE (Louizos et al. 2017) and TEDVAE (Zhang, Liu, and Li 2021)) as baselines for TE estimation only.

Evaluation Criterion For evaluating the performance of DMAVAE and the baselines, we use the estimation bias $|(\hat{\beta} - \beta)/\beta| \times 100\%$ as the metric, where $\hat{\beta}$ is the estimated results and β is the ground truth.

Implementation Details We use PyTorch (Paszke et al. 2019) and Pyro (Bingham et al. 2018) to implement DMAVAE. For LSME and LSME-I, we use their implementations in R package "mediation" (Tingley et al. 2014). For NEW-I and NEW-W, we use their implementations in the R package "Medflex" (Steen et al. 2017). For IPW, we use the implementation in the R package "causal-weight" (Bodory

and Huber 2018). CEVAE[1] and TEDVAE[2] are open-source on GitHub, and CMAVAE is implemented in Pyro and PyTorch by us.

Performance Evaluation

We firstly generate synthetic datasets based on the causal graph in Figure 2 (hence we know the ground truth) to evaluate the performance of our methods, in comparison with the baselines. We follow the same procedure in (Louizos et al. 2017; Cheng, Guo, and Liu 2022) to generate synthetic datasets with a variety range of sample sizes (denoted as N): 2k, 3k, 4k, 5k, 6k, 7k, 8k, 9k and 10k.

To avoid the bias brought by data generation, we repeatedly generate 30 datasets for each sample size. The results for the mean and the standard deviation of estimation bias for NDE, NIE_r and TE are shown in Table 1.

Results DMAVAE achieves the lowest estimation bias in both NDE and NIE_r compared with the baseline methods as shown in Tables 1a and 1b. We note that DMAVAE performs clearly better than CMAVAE under the current data generation mechanism. A possible explanation is that since there are three types of latent confounders, but CMAVAE only learns one type of representation. As we mentioned above, different adjustment sets are needed for different estimation tasks, so only considering one type of representation can lead to a relatively large bias in estimating NDE and NIE_r in this case.

In Table 1c, LSEM, LSEM-I, CMAVAE and DMAVAE show better performance for estimating TE compared with the other baseline methods. With the increase in sample size, CMAVAE shows lower estimation bias than DMAVAE. The reason could be that estimating TE requires the representation of all confounders, as DMAVAE disentangles three types of representations, it may produce more cumulative errors compared to CMAVAE, which only assumes one type of latent confounders.

Evaluation on Generalisability

We also conduct experiments on multiple cases to verify the effectiveness and the generalisation ability of DMAVAE to data generated from structures with different types of latent confounders. We note that among the baseline methods used previously, only CMAVAE can achieve a similar level of estimation bias as our methods for NDE and NIE_r. Hence, we select CMAVAE as the compared model in this part of the evaluation.

We have already compared DMAVAE and CMAVAE under the latent confounder assumption shown in Figure 2. We have seen that our proposed DMAVAE method outperforms CMAVAE in estimating NDE and NIE_r and achieves similar performance in estimating TE. In this section, we firstly compare DMAVAE and CMAVAE with the data generated based on the causal graph shown in Figure 1b. The results are shown in Table 2. We note that DMAVAE has a higher estimation bias compared to the results in Table 1, while

---

[1]https://github.com/AMLab-Amsterdam/CEVAE
[2]https://github.com/WeijiaZhang24/TEDVAE

N	2k	3k	4k	5k	6k	7k	8k	9k	10k
LSEM	7.1 ± 1.6	7.1 ± 1.3	7.3 ± 1.2	6.9 ± 1.3	7.4 ± 0.9	7.5 ± 0.9	7.2 ± 1.0	7.2 ± 0.7	7.4 ± 0.9
LSEM-I	8.3 ± 1.8	7.9 ± 1.3	8.1 ± 1.3	7.8 ± 1.1	8.2 ± 1.0	8.6 ± 0.9	8.0 ± 1.0	8.1 ± 0.8	8.3 ± 0.8
NEW-I	10.9 ± 8.4	7.8 ± 6.7	7.9 ± 6.3	7.5 ± 5.5	7.6 ± 4.5	8.8 ± 5.9	6.8 ± 4.5	6.4 ± 4.2	8.2 ± 5.2
NEW-W	32.1 ± 12.0	21.5 ± 8.9	22.2 ± 7.7	20.3 ± 8.4	20.7 ± 9.6	16.8 ± 7.0	12.9 ± 5.0	18.6 ± 6.5	15.6 ± 6.2
IPW	7.5 ± 2.5	7.0 ± 1.7	7.7 ± 1.4	7.3 ± 1.4	7.5 ± 1.4	7.9 ± 1.4	7.3 ± 1.3	7.6 ± 1.0	8.0 ± 0.9
CMAVAE	8.9 ± 5.0	5.5 ± 2.3	4.4 ± 2.2	3.8 ± 2.0	3.9 ± 1.9	4.6 ± 1.7	4.5 ± 2.1	4.7 ± 1.4	5.2 ± 2.0
DMAVAE	**4.2 ± 2.7**	**1.7 ± 1.3**	**1.7 ± 1.3**	**1.8 ± 1.4**	**1.8 ± 1.1**	**1.4 ± 0.8**	**1.7 ± 1.3**	**1.3 ± 1.0**	**1.5 ± 1.1**

(a)

N	2k	3k	4k	5k	6k	7k	8k	9k	10k
LSEM	11.0 ± 3.0	10.7 ± 2.2	10.6 ± 2.5	10.4 ± 2.7	11.0 ± 1.6	10.4 ± 1.9	10.8 ± 1.7	10.5 ± 1.8	11.1 ± 1.5
LSEM-I	13.2 ± 3.1	12.1 ± 2.3	12.5 ± 2.4	12.2 ± 2.5	12.7 ± 1.8	12.1 ± 1.9	12.5 ± 1.6	12.2 ± 1.6	13.0 ± 1.5
NEW-I	11.7 ± 9.1	9.5 ± 7.6	10.0 ± 7.0	9.0 ± 6.6	10.5 ± 4.7	12.1 ± 5.6	9.1 ± 5.4	9.6 ± 4.1	11.2 ± 5.1
NEW-W	24.6 ± 9.3	18.7 ± 5.8	18.3 ± 7.2	20.8 ± 7.0	19.2 ± 6.4	16.6 ± 5.5	16.5 ± 5.7	13.5 ± 5.5	18.7 ± 6.4
IPW	58.0 ± 3.7	57.1 ± 3.6	58.2 ± 2.5	57.7 ± 2.5	57.7 ± 2.2	57.7 ± 1.9	57.6 ± 1.7	58.0 ± 2.0	58.1 ± 1.2
CMAVAE	15.1 ± 8.7	8.2 ± 5.0	6.1 ± 4.1	5.7 ± 3.5	6.0 ± 2.9	7.7 ± 3.7	9.6 ± 3.3	9.4 ± 3.3	9.2 ± 3.4
DMAVAE	**6.4 ± 4.9**	**3.9 ± 2.6**	**5.1 ± 3.5**	**4.8 ± 3.1**	**4.2 ± 2.4**	**4.5 ± 2.5**	**4.3 ± 2.1**	**4.5 ± 2.3**	**4.6 ± 2.3**

(b)

N	2k	3k	4k	5k	6k	7k	8k	9k	10k
LSEM	**1.1 ± 0.9**	1.0 ± 0.8	1.3 ± 0.8	1.1 ± 0.7	**1.1 ± 0.6**	1.4 ± 0.6	**1.1 ± 0.5**	1.2 ± 0.5	**1.1 ± 0.5**
LSEM-I	1.2 ± 0.9	1.1 ± 0.7	**1.2 ± 0.7**	1.1 ± 0.6	1.2 ± 0.7	1.5 ± 0.7	1.3 ± 0.6	1.2 ± 0.5	**1.1 ± 0.5**
NEW-I	10.9 ± 8.6	8.2 ± 7.0	8.6 ± 6.4	7.9 ± 5.9	8.5 ± 4.6	9.9 ± 5.8	7.5 ± 4.8	7.3 ± 4.3	9.2 ± 5.2
NEW-W	27.9 ± 11.2	18.8 ± 7.5	19.0 ± 7.3	18.5 ± 7.9	18.3 ± 8.3	15.4 ± 5.6	11.6 ± 5.0	14.6 ± 5.7	14.9 ± 5.9
IPW	14.8 ± 1.4	14.8 ± 1.0	14.8 ± 0.6	14.9 ± 0.7	14.7 ± 0.7	14.5 ± 0.7	14.8 ± 0.5	14.8 ± 0.5	14.5 ± 0.5
CEVAE	24.1 ± 5.6	27.0 ± 4.3	28.1 ± 1.1	28.3 ± 1.0	28.1 ± 0.8	27.8 ± 0.8	27.9 ± 1.1	27.5 ± 0.8	27.7 ± 1.1
TEDVAE	28.3 ± 1.3	28.3 ± 1.3	27.9 ± 1.3	28.1 ± 0.9	27.9 ± 1.0	27.6 ± 0.8	28.3 ± 1.0	28.0 ± 0.9	27.9 ± 1.0
CMAVAE	**1.1 ± 0.9**	**1.0 ± 0.7**	**1.2 ± 0.7**	0.9 ± 0.7	1.2 ± 0.8	**1.0 ± 0.6**	1.1 ± 0.8	**0.9 ± 0.6**	1.1 ± 0.8
DMAVAE	**1.1 ± 0.9**	**1.0 ± 0.7**	1.3 ± 0.7	**0.8 ± 0.7**	1.2 ± 0.7	1.3 ± 0.9	1.3 ± 0.9	1.2 ± 0.8	1.2 ± 0.9

(c)

Table 1: (a) Estimation bias (%) for the estimated NDE. (b) Estimation bias (%) for the estimated NIE_r. (c) Estimation bias (%) for the estimated TE. The best results are shown in boldface, and the runner-up results are underlined.

CMAVAE achieves a lower estimation bias than that in Table 1. Such results are reasonable since the data generation mechanism fully complies with the assumption of CMAVAE (i.e., there only exists one type of confounder that affects T, M and Y simultaneously). However, we find that even under this assumption, DMAVAE achieves lower estimation bias in some datasets compared to CMAVAE, which implies that DMAVAE has good generalisation ability.

Furthermore, we consider multiple cases when comparing DMAVAE and CMAVAE in terms of their generalisability. Similar to the job market example mentioned in Introduction, the causal graph for real applications may contain one type or two types or all the three types of latent confounders. When considering the different combinations of the three types of latent confounders, we get six cases (indicated in Figure 4 as Case 1 to Case 6). The datasets for each case are generated separately following the respective causal graphs. For Case 1 to Case 3, we assume only one type of latent confounders exists. For Case 4 to Case 6, we assume there exist two types of latent confounders. The results for the six cases are shown in Figure 4. From the figure, we see that DMAVAE outperforms CMAVAE in all 6 cases. This indicates that DMAVAE has better generalisation ability, and can deal with different cases in practice more effectively than CMAVAE.

Model	DMAVAE		CMAVAE	
N	NDE	NIE_r	NDE	NIE_r
2k	8.5 ± 4.8	**7.4 ± 5.5**	**5.2 ± 4.4**	19.9 ± 8.8
3k	4.6 ± 3.6	**5.4 ± 3.5**	**2.6 ± 2.3**	18.7 ± 9.3
4k	4.3 ± 2.4	**4.5 ± 3.5**	**2.9 ± 2.1**	9.9 ± 5.8
5k	4.3 ± 2.1	**4.5 ± 2.6**	**3.8 ± 2.2**	4.8 ± 1.0
6k	**4.2 ± 2.0**	4.8 ± 3.5	4.3 ± 2.5	4.8 ± 4.1
7k	**3.6 ± 1.9**	4.9 ± 3.2	3.8 ± 2.0	5.4 ± 3.8
8k	3.9 ± 1.7	**4.6 ± 2.6**	**3.5 ± 1.7**	4.7 ± 4.4
9k	4.1 ± 1.8	4.7 ± 2.8	**3.9 ± 1.8**	**3.9 ± 3.9**
10k	**4.0 ± 1.9**	4.7 ± 3.0	**4.0 ± 1.8**	3.8 ± 3.2

Table 2: Estimation bias (%) for the estimated NDE and NIE_r with data generated based on the causal graph shown in Figure 1b. The best results are shown in boldface.

Case Study

Adult Dataset To illustrate the potential of DMAVAE for real-world applications, we show that DMAVAE can be used for detecting discrimination. We apply DMAVAE on the real-world dataset Adult. The dataset is retrieved from the UCI repository (Dua and Graff 2017) and it contains 14 attributes including personal, educational and economic information for 48842 individuals. We use the sensitive attribute gender as T, occupation as M, income as Y and all the other attributes as the proxy attributes X.
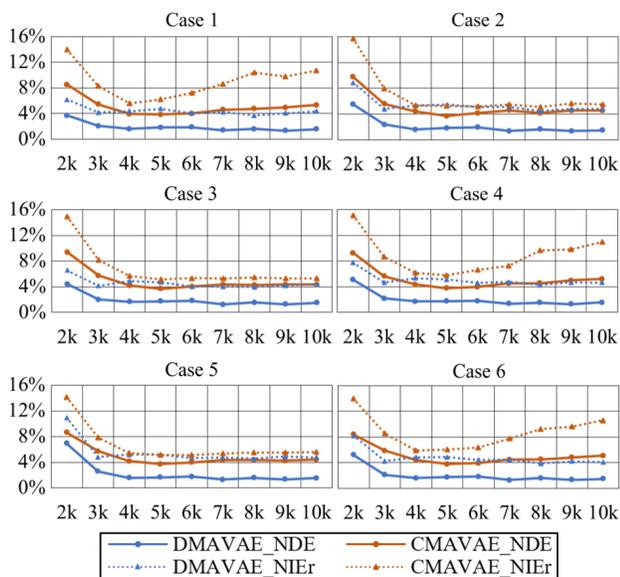
Figure 4: Results for the six cases. The horizontal axis indicates sample size (N) and the vertical axis indicates estimation bias (%).

For causality-based discrimination detection, NDE indicates the causal effect that the sensitive attribute directly has on the outcome and NIE refers to the causal effect through the mediator on the outcome. Hence the approach detects discrimination in a dataset by estimating NDE and NIE. There is direct discrimination when the NDE is greater than a given threshold τ, and there exists indirect discrimination when the mediator is also considered as a sensitive attribute and the NIE is greater than τ (Zhang, Wu, and Wu 2018).

We obtain that NDE = 0.804 and NIE = 0.002 by applying DMAVAE on the Adult dataset ($\tau = 0.05$). The results indicate that there is significant direct discrimination against gender, while the discrimination against gender through occupation is ignorable.

Related Work

Mediation analysis has its origins in structural equation models (SEMs), and the specific methods of CMA can be traced back to path analysis (Wright 1934), which has been widely used in social sciences for decades (Baron and Kenny 1986; Bollen 1989). Under the linear models, causal effects are expressed as the coefficients of structural equations. These coefficients are defined, identified, and estimated based on the assumption of sequential ignorability and commitment to a specific distribution. Specifically, LSEM is a well-established method based on linear SEMs for mediation analysis (Baron and Kenny 1986; James, Mulaik, and Brett 1983; Judd and Kenny 1981; MacKinnon 2012; MacKinnon and Dwyer 1993). Although LSEM is extremely simple to use, the additional assumptions are often not satisfied in reality, leading to biased estimation (Huber, Lechner, and Mellace 2016; Rudolph et al. 2019).

The above situation has been improved in recent decades.

Counterfactual thinking in statistics (Rubin 1974; Holland 1988) and epidemiology (Robins and Greenland 1992), together with the formal definitions based on non-parametric structural equations provide theoretical support for mediation analysis to be expanded from linear to nonlinear models. With counterfactual thinking, defining direct and indirect effects requires no commitment to the distributional forms, and thus applies under functions with arbitrary nonlinear interactions, and both continuous and categorical variables (Pearl 2014; Rubin 2005). For example, Imai, Keele, and Tingley (2010) extended LSEM and proposed a general approach that brings the definition, identification, estimation, and sensitivity analysis of causal mediation effects closely together within a single method; Huber (2014) demonstrated that the applicability of inverse probability weighting depends on whether some confounders are themselves influenced by the treatment or not. For more details on mediation analysis, please refer to the papers (MacKinnon, Fairchild, and Fritz 2007; VanderWeele 2016).

In recent years, deep learning has demonstrated powerful ability in fitting non-linear models, and researchers are devoted to estimating causal effects through deep learning techniques. CEVAE (Louizos et al. 2017) is the first method that uses VAE to learn the representation of latent confounders and adjust for confounding bias using the learned representation. Cheng, Guo, and Liu (2022) used VAE-based techniques for mediation analysis and proposed CMAVAE that can learn the representation of latent confounders to estimate direct and indirect effects under the sequential ignorability assumption (Imai, Keele, and Tingley 2010).

However, CMAVAE assumes that there is only one type of latent confounders, which is not applicable to the different cases in practice. Our work improves CMAVAE. We extend the piecemeal deconfounding assumption to disentangled representation learning. The goal is to achieve more accurate estimation results and enhance the generalisation ability of the method.

Conclusion

This work investigates an important causal mediation analysis problem, i.e., how to achieve accurate estimation of NDE, NIE and TE. We use a relaxed assumption, piecemeal deconfounding rather than sequential ignorability for causal mediation analysis. We extend the piecemeal deconfounding assumption to representation learning in a general case with three types of latent confounders. Then, we introduce DMAVAE, a VAE-based method that can learn disentangled representations from proxy attributes for estimating NDE, NIE and TE. With extensive experiments, we show that the proposed method significantly outperforms other mediation analysis methods on synthetic datasets. We demonstrate that DMAVAE has a strong generalisation ability and can be applied to different cases. Furthermore, we verify the ability of DMAVAE on a real-world application. For future work, we will apply this model to scenarios of interest in mediation analysis to help understand mechanisms or guide policy formulation.

Acknowledgments

References

Angrist, J. D.; and Pischke, J.-S. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Baron, R. M.; and Kenny, D. A. 1986. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6): 1173–1182.

Bickel, P. J.; Hammel, E. A.; and O'Connell, J. W. 1975. Sex bias in graduate admissions: Data from Berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation. *Science*, 187(4175): 398–404.

Bingham, E.; Chen, J. P.; Jankowiak, M.; Obermeyer, F.; Pradhan, N.; Karaletsos, T.; Singh, R.; Szerlip, P.; Horsfall, P.; and Goodman, N. D. 2018. Pyro: Deep universal probabilistic programming. *Journal of Machine Learning Research*, 1–6.

Bodory, H.; and Huber, M. 2018. The causalweight package for causal inference in R. FSES Working Papers 493, Faculty of Economics and Social Sciences, University of Freiburg/Fribourg Switzerland.

Bollen, K. A. 1989. *Structural equations with latent variables*, volume 210. John Wiley & Sons.

Cheng, L.; Guo, R.; and Liu, H. 2022. Causal Mediation Analysis with Hidden Confounders. In *The Fifteenth ACM International Conference on Web Search and Data Mining, WSDM 2022*, 113–122.

Dua, D.; and Graff, C. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml. University of California, Irvine, School of Information and Computer Sciences.

Guo, R.; Cheng, L.; Li, J.; Hahn, P. R.; and Liu, H. 2020. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)*, 53(4): 1–37.

Holland, P. W. 1988. Causal inference, path analysis and recursive structural equations models. *ETS Research Report Series*, 1988(1): 1–50.

Huber, M. 2014. Identifying causal mechanisms (primarily) based on inverse probability weighting. *Journal of Applied Econometrics*, 29(6): 920–943.

Huber, M.; Lechner, M.; and Mellace, G. 2016. The finite sample performance of estimators for mediation analysis under sequential conditional independence. *Journal of Business & Economic Statistics*, 34(1): 139–160.

Imai, K.; Keele, L.; and Tingley, D. 2010. A general approach to causal mediation analysis. *Psychological Methods*, 15(4): 309–334.

James, L. R.; Mulaik, S. A.; and Brett, J. M. 1983. *Causal analysis: Assumptions, models, and data*. Beverly Hills (Calif.): Sage, 1983.

Judd, C. M.; and Kenny, D. A. 1981. Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review*, 5(5): 602–619.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014*, 1–14.

Lange, T.; Vansteelandt, S.; and Bekaert, M. 2012. A simple unified approach for estimating natural direct and indirect effects. *American Journal of Epidemiology*, 176(3): 190–195.

Louizos, C.; Shalit, U.; Mooij, J. M.; Sontag, D. A.; Zemel, R. S.; and Welling, M. 2017. Causal Effect Inference with Deep Latent-Variable Models. In *Advances in Neural Information Processing Systems 30, NeurIPS 2017*, 6446–6456.

MacKinnon, D. P. 2012. *Introduction to statistical mediation analysis*. Routledge.

MacKinnon, D. P.; and Dwyer, J. H. 1993. Estimating mediated effects in prevention studies. *Evaluation Review*, 17(2): 144–158.

MacKinnon, D. P.; Fairchild, A. J.; and Fritz, M. S. 2007. Mediation analysis. *Annual Review of Psychology*, 58: 1–22.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32, NeurIPS 2019*, 8024–8035.

Pearl, J. 2009a. Causal inference in statistics: An overview. *Statistics Surveys*, 3: 96–146.

Pearl, J. 2009b. *Causality*. Cambridge University Press.

Pearl, J. 2014. Interpretation and identification of causal mediation. *Psychological Methods*, 19(4): 459–481.

Robins, J. M.; and Greenland, S. 1992. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 143–155.

Robins, J. M.; Rotnitzky, A.; and Zhao, L. P. 1994. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427): 846–866.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5): 688.

Rubin, D. B. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469): 322–331.

Rudolph, K. E.; Goin, D. E.; Paksarian, D.; Crowder, R.; Merikangas, K. R.; and Stuart, E. A. 2019. Causal mediation analysis with observational data: considerations and illustration examining mechanisms linking neighborhood poverty

to adolescent substance use. *American Journal of Epidemiology*, 188(3): 598–608.

Steen, J.; Loeys, T.; Moerkerke, B.; and Vansteelandt, S. 2017. Medflex: an R package for flexible mediation analysis using natural effect models. *Journal of Statistical Software*, 76: 1–46.

Tingley, D.; Yamamoto, T.; Hirose, K.; Keele, L.; and Imai, K. 2014. Mediation: R Package for Causal Mediation Analysis. *Journal of Statistical Software*, 59(5): 1–38.

VanderWeele, T. J. 2016. Mediation analysis: a practitioner's guide. *Annual Review of Public Health*, 37: 17–32.

Wright, S. 1934. The method of path coefficients. *The Annals of Mathematical Statistics*, 5(3): 161–215.

Yin, X.; and Hong, L. 2019. The Identification and Estimation of Direct and Indirect Effects in A/B Tests through Causal Mediation Analysis. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019*, 2989–2999.

Zhang, L.; Wu, Y.; and Wu, X. 2018. Causal modeling-based discrimination discovery and removal: criteria, bounds, and algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 31(11): 2035–2050.

Zhang, W.; Liu, L.; and Li, J. 2021. Treatment Effect Estimation with Disentangled Latent Factors. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, 10923–10930.