Conditional Counterfactual Causal Effect for Individual Attribution

Ruiqi Zhao [1] Lei Zhang [2] Shengyu Zhu [3] Zitong Lu [4] Zhenhua Dong [3] Chaoliang Zhang [3] Jun Xu [2] Zhi Geng [5] Yangbo He [1]

[1]School of Mathematical Sciences, Peking University [2]Renmin University of China [3]Huawei Noah's Ark Lab
[4]City University of Hong Kong [5]School of Mathematics and Statistics, Beijing Technology and Business University

Abstract

Identifying the causes of an event, also termed as causal attribution, is a commonly encountered task in many application problems. Available methods, mostly in Bayesian or causal inference literature, suffer from two main drawbacks: (1) cannot attribute for individuals, (2) attributing one single cause at a time and cannot deal with the interaction effect among multiple causes. In this paper, based on our proposed new measurement, called conditional counterfactual causal effect (CCCE), we introduce an individual causal attribution method, which is able to utilize the individual observation as the evidence and consider common influence and interaction effect of multiple causes simultaneously. We discuss the identifiability of CCCE and also give the identification formulas under proper assumptions. Finally, we conduct experiments on simulated and real data to illustrate the effectiveness of CCCE and the results show that our proposed method outperforms significantly over state-of-the-art methods.

Preliminary

Unless otherwise stated, we use capital letters such as X to denote random variables, and use boldface letters like $\mathbf{X}$ to denote random variable sets or vectors. An instantiation of a variable or a vector is denoted by a lowercase letter, e.g., x and $\mathbf{x}$.

Notations

- $\mathbf{X}$: binary variables $\{X_1, X_2, \cdots, X_p\}$ which are possible causes of outcome.
- Y: binary variable which is the outcome variable.
- $\mathbf{Z}$: binary variables $\{Z_1, Z_2, \cdots, Z_q\}$ that do not affect $\mathbf{X}$ and Y.
- π: a given topological ordering $(\mathbf{X}, Y, \mathbf{Z})$ of variables so that W_i is not affected by W_j for any $W_i, W_j \in \pi$ with $i < j$.
- $\mathbf{A}_k$: $\{X_1, X_2, \cdots, X_{k-1}\}$ in π.
- $\mathbf{D}_k$: $\mathbf{X} \backslash \mathbf{A}_k$ in π, that is, $\{X_k, X_{k+1}, \cdot, X_p\}$.
- $\mathbf{X}_{-k}$: the set of variables X_i's without X_k, that is, $\mathbf{X} \backslash \{X_k\}$.
- $Y_{\mathbf{x}}$: potential outcome of Y under $\mathbf{X} = \mathbf{x}$.
- $\mathbf{x} \preceq \mathbf{x}'$: $x_i \leq x_i'$ for all i.
- $|\mathbf{X}|$: the cardinality of set $\mathbf{X}$.

Given a topological order π, the data generating mechanism for $\mathbf{X}$ and Y can be described as

$$Y = f_Y(\mathbf{X}, \epsilon_Y),$$
$$X_k = f_k(\mathbf{A}_k, \epsilon_k), k = 1, 2, \ldots, p,$$

where ϵ_k's and ϵ_Y are independent noise variables. The data generating mechanisms for variables in $\mathbf{Z}$ can be similarly defined. Besides, we assume the ***consistency*** holds, that is, for any variable sets $\mathbf{U}$ and $\mathbf{V}$, we have $\mathbf{U}_{\mathbf{v}} = \mathbf{U}$ if $\mathbf{V} = \mathbf{v}$ is observed. We also assume the ***composition*** holds, that is, for any variable sets $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$, we have that $\mathbf{U}_{\mathbf{v}} = \mathbf{u}$ implies $\mathbf{W}_{\mathbf{uv}} = \mathbf{W}_{\mathbf{v}}$.

CCCE can be regarded as a generalization of PostTCE. Specifically, in the case of $\mathbf{Z} = \varnothing$ and $\mathbf{X_S} = \{X_1\}$, if the evidence $\mathbf{W} = \mathbf{w}$ is in the form of $\{\mathbf{U} = \mathbf{u}, Y = 1 : \mathbf{U} \subset \mathbf{X}\}$, then CCCE is reduced to $\mathrm{E}(Y_{X_1=1} - Y_{X_1=0} \mid \mathbf{U} = \mathbf{u}, Y = 1)$, which is exactly PostTCE$(X_1 \Rightarrow Y \mid \mathbf{U} = \mathbf{u}, Y = 1)$. The details will be given in next section. Further, in the case of $\mathbf{Z} = \varnothing$ and $\mathbf{X_S} = \{X_1\}$, if the evidence is given by $\{X_1 = 1, Y = 1\}$, then CCCE is reduced to $\mathrm{E}(Y_{X_1=1} - Y_{X_1=0} \mid X_1 = 1, Y = 1) = \mathrm{P}(Y_{X_1=0} = 0 \mid X_1 = 1, Y = 1)$, which is exactly PN$(X_1 \Rightarrow Y)$. What's more, if the evidence is given by $\{X_1 = 0, Y = 0\}$, then CCCE is reduced to $\mathrm{E}(Y_{X_1=1} - Y_{X_1=0} \mid X_1 = 0, Y = 0) = \mathrm{P}(Y_{X_1=1} = 1 \mid X_1 = 0, Y = 0)$, which is exactly PS$(X_1 \Rightarrow Y)$. Thus, CCCE can be considered as a generalization of PN, PS and PostTCE. It should be noted that this generalization is not trivial, because CCCE calculates the impact of random multiple causes on the result simultaneously, and also adds $\mathbf{Z}$ into the evidence set, which makes the identifiability of CCCE more complex than PostTCE.

Definition

Conditional Counterfactual Causal Effect

Given the variable set $\mathbf{V} = \{\mathbf{X}, Y, \mathbf{Z}\}$, the evidence $\mathbf{W} = \mathbf{w}$ and a set of causes $\mathbf{X_S} \subseteq \mathbf{X}$, the ***conditional counterfactual causal effect*** (CCCE) of $\mathbf{X_S}$ on Y is defined as

$$\mathrm{CCCE}(\mathbf{X_S} \Rightarrow Y \mid \mathbf{W} = \mathbf{w}) = \mathbb{E}\left(Y_{\mathbf{x}_\mathbf{S}^1} - Y_{\mathbf{x}_\mathbf{S}^0} \mid \mathbf{W} = \mathbf{w}\right), \tag{1}$$

in which $\mathbf{W} \subseteq \mathbf{V}$ and $\mathbf{x}_\mathbf{S}^1 \succeq \mathbf{x}_\mathbf{S}^0$.

In many applications, we would set $\mathbf{x}_\mathbf{S}^1$ and $\mathbf{x}_\mathbf{S}^0$ as $\mathbf{1}_{|\mathbf{S}|}$ and $\mathbf{0}_{|\mathbf{S}|}$, which are vectors with all entries being 1 and 0, respectively. By this definition, CCCE can measure the joint influence of all causes in $\mathbf{X_S}$ on the outcome Y. We remark that $\mathbf{x}_\mathbf{S}^1$ and $\mathbf{x}_\mathbf{S}^0$ can be set to other values given other cases of interest. In addition, the evidence $\mathbf{W} = \mathbf{w}$ always contains the observation of Y. For example, in order to evaluate the effect of advertising, we are more concerned about the conversion rate of recommendations among people with buying behavior ($Y = 1$). Hence, CCCE measures the causation for causes in $\mathbf{X_S}$. The larger the CCCE of $\mathbf{X_S}$ on Y is, the larger the attribution of the effect to the causes $\mathbf{X_S}$ is. Note that the evidence $\mathbf{W} = \mathbf{w}$ can contain the observation of $\mathbf{X_S}$. Therefore, CCCE is different from the conditional causal effect. The former is a counterfactual variable, while the latter only conditions on covariates except $\mathbf{X_S}$, that is, the conditional causal effect only involves intervention variables and can be identified with observational and interventional data under suitable assumptions.

Assumptions

To give the identification formula of CCCE, the following assumptions are required.

Assumption 1 (No confounding)

(a) There is no confounding among variables in $\mathbf{X}$, that is, $(\mathbf{X}_k)_{\mathbf{a}_k} \perp \mathbf{A}_k$ for all $\mathbf{a}_k$;

(b) There is no confounding between Y and $\mathbf{X}$, that is, $Y_{\mathbf{x}} \perp \mathbf{X}$ for all $\mathbf{x}$;

(c) Given $\mathbf{X}$ and Y, there is no confounding between $(\mathbf{X}, Y)$ and $\mathbf{Z}$, that is, $(Y_{\mathbf{x}}, \mathbf{X}_{\mathbf{x_S}}) \perp \mathbf{Z} \mid \mathbf{X}, Y$ for all $\mathbf{x}, \mathbf{x_S}$ and $\mathbf{X_S} \subseteq \mathbf{X}$.

The assumption of no confounding is also known as the assumption of ignorability or exogeneity, implying that there is no unobserved confounders. The Assumptions 1(a) and 1(b) are satisfied if $\epsilon_1, \ldots, \epsilon_p$ are mutually independent and $(\epsilon_1, \ldots, \epsilon_p)$ and ϵ_Y are independent, respectively, while the Assumption 1(c) is satisfied when $(\epsilon_1, \ldots, \epsilon_p, \epsilon_Y)$ and the noise variables of $\mathbf{Z}$ are independent.

Assumption 2 (Monotonicity)

(a) The variables in $\mathbf{X}$ satisfy the monotonicity, that is, $(X_k)_{\mathbf{a}_k} \leq (X_k)_{\mathbf{a}_k'}$ for all $k = 1, \ldots, p$ whenever $\mathbf{a}_k \preceq \mathbf{a}_k'$;

(b) The outcome variable Y satisfies the monotonicity, that is, $Y_{\mathbf{x}} \leq Y_{\mathbf{x}'}$ whenever $\mathbf{x} \preceq \mathbf{x}'$.

The assumption of monotonicity is often assumed in practice, implying that the causes cannot prevent the effect.

Main Results

Lemma 1. Given a causal ordering π, let $\mathbf{X_S} \subseteq \mathbf{X}$ and $\mathbf{x}_\mathbf{S}^0 \preceq \mathbf{x_S}$. The conditional probability $\mathrm{P}(Y_{\mathbf{x}_\mathbf{S}^0} = 1 \mid \mathbf{X} = \mathbf{x})$ is identifiable, and its identification formula is

$$\begin{aligned}\mathrm{P}(Y_{\mathbf{x}_\mathbf{S}^0} = 1 \mid \mathbf{X} = \mathbf{x}) = \sum_{\mathbf{c}_{k:p} \preceq \mathbf{d}_k} \Big\{ & \mathrm{P}(Y = 1 \mid \mathbf{A}_k = \mathbf{a}_k, \mathbf{D}_k = \mathbf{c}_{k:p}) \\ \times \prod_{i \in \{k, \ldots, p\} \backslash \mathbf{S}} & \left[1 - x_i c_i + x_i (-1)^{1 - c_i} \times \frac{\mathrm{P}(X_i = 1 \mid \mathbf{A}_k = \mathbf{a}_k, \mathbf{X}_{k:i-1} = \mathbf{c}_{k:i-1})}{\mathrm{P}(X_i = x_i \mid \mathbf{A}_i = \mathbf{a}_i)}\right] \Big\},\end{aligned} \tag{2}$$

where $k = \min \mathbf{S}, \mathbf{X}_{k:i-1} = \{X_k, \ldots, X_{i-1}\}$ and $\mathbf{c}_{k:p} = (c_k, \ldots, c_p)$ satisfying $c_i = x_i^0$ if $i \in \mathbf{S}$.

Lemma 2. Given a causal ordering π, let $\mathbf{X_S} \subseteq \mathbf{X}$ and $\mathbf{x}_\mathbf{S}^1 \succeq \mathbf{x_S}$. The conditional probability $\mathrm{P}(Y_{\mathbf{x}_\mathbf{S}^1} = 1 \mid \mathbf{X} = \mathbf{x})$ is identifiable, and its identification formula is

$$\begin{aligned}\mathrm{P}(Y_{\mathbf{x}_\mathbf{S}^1} = 1 \mid \mathbf{X} = \mathbf{x}) = \sum_{\mathbf{c}_{k:p} \succeq \mathbf{d}_k} \Big\{ & \mathrm{P}(Y = 1 \mid \mathbf{A}_k = \mathbf{a}_k, \mathbf{D}_k = \mathbf{c}_{k:p}) \\ \times \prod_{i \in \{k, \ldots, p\} \backslash \mathbf{S}} & \left[x_i + c_i - x_i c_i + (1 - x_i)(-1)^{c_i} \times \frac{\mathrm{P}(X_i = 0 \mid \mathbf{A}_k = \mathbf{a}_k, \mathbf{X}_{k:i-1} = \mathbf{c}_{k:i-1})}{\mathrm{P}(X_i = x_i \mid \mathbf{A}_i = \mathbf{a}_i)}\right] \Big\},\end{aligned} \tag{3}$$

where $k = \min \mathbf{S}, \mathbf{X}_{k:i-1} = \{X_k, \ldots, X_{i-1}\}$ and $\mathbf{c}_{k:p} = (c_k, \ldots, c_p)$ satisfying $c_i = x_i^1$ if $i \in \mathbf{S}$.

Taking Lemma 1 as an example, the conditional probability $\mathrm{P}(Y = 1 \mid \mathbf{A}_k = \mathbf{a}_k, \mathbf{D}_k = \mathbf{c}_{k:p})$ in Equation (2) contains $\mathbf{D}_k$ in the condition part, which may be affected by $\mathbf{X_S}$. Especially, if the observed sample $\mathbf{x_S} = \mathbf{x}_\mathbf{S}^0$ appears in the evidence $\mathbf{W} = \mathbf{w}$, then we have $\mathrm{P}(Y_{\mathbf{x}_\mathbf{S}^0} = 1 \mid \mathbf{X} = \mathbf{x}) = \mathrm{P}(Y = 1 \mid \mathbf{X} = \mathbf{x})$ by consistency. If $\mathbf{Z} = \varnothing$ and $|\mathbf{S}| = 1$, that is, $\mathbf{X_S} = \{X_1\}$, then Lemma 1 degenerates to Lemma 1 in Lu et al. (2022). A similar observation holds for Lemma 2.

Theorem 1. Under Assumption 2(b), the following equation holds:

$$\mathrm{CCCE}(\mathbf{X}_S \Rightarrow Y \mid \mathbf{X} = \mathbf{x}, Y = y) = 1 - \frac{\mathrm{P}(Y_{\mathbf{x}_\mathbf{S}^{1-y}} = y \mid \mathbf{X} = \mathbf{x})}{\mathrm{P}(Y = y \mid \mathbf{X} = \mathbf{x})}; \tag{4}$$

where $\mathbf{X_S} \subseteq \mathbf{X}, \mathbf{x}_\mathbf{S}^1 \succeq \mathbf{x_S} \succeq \mathbf{x}_\mathbf{S}^0$ and $\mathbf{x_S} \subseteq \mathbf{x}$.

Theorem 2. Given a causal ordering π, let $\mathbf{X_S} \subseteq \mathbf{X}$, and $\mathbf{W}$ is an arbitrary subset of $\{\mathbf{X}, Y, \mathbf{Z}\}$. CCCE of $\mathbf{X_S}$ on Y based on the evidence $\mathbf{W} = \mathbf{w}$ has the following equation:

$$\begin{aligned}\mathrm{CCCE}(\mathbf{X_S} \Rightarrow Y \mid \mathbf{W} = \mathbf{w}) = & \sum_{(\mathbf{x}, y, \mathbf{z}) : (\mathbf{x}, y, \mathbf{z}) \supseteq \mathbf{w}} \mathrm{CCCE}(\mathbf{X_S} \Rightarrow Y \mid \mathbf{X} = \mathbf{x}, Y = y) \\ & \times \mathrm{P}(\mathbf{X} = \mathbf{x}, Y = y, \mathbf{Z} = \mathbf{z} \mid \mathbf{W} = \mathbf{w}),\end{aligned} \tag{5}$$

which is identifiable according to Theorem 1.

According to the results above, CCCE only uses the topological ordering of the variables for the attribution, but a causal graph may has several different valid topological ordering. In fact, for a given graph, the value of CCCE is invariant for different valid topological orderings, as long as the evidence set $\mathbf{W} = \mathbf{w}$ contains the variable Y and its all ancestors. Note that, for any topological ordering of a given graph, the order of ancestors of Y always precede the order of Y. Therefore, for any given valid topological ordering, we only need to make the evidence set $\mathbf{W} = \mathbf{w}$ contain Y and the variables before Y in this ordering.

Experiments

Simulations

We generate ten different causal graphs and conducte ten experiments on each with 1000 samples. For each individual sample, we use Rand (randomly select in variables preceding Y), Post, PN, PS, PNS, ACE, PostTCE and CCCE for attribution and take the observational sample $(x_1, x_2, x_3, x_4, y, x_6)$ as the evidence. After finding the causes, we use the Change Rate (CR, higher is better) to measure the effectiveness and accuracy of attribution, that is, we set the cause variables $\mathbf{X_S}$ to $\mathbf{0}_{|\mathbf{S}|}$, regenerate the counterfactual data, and calculate the proportion of samples whose Y changes from 1 to 0 in the whole sample. The value of CR measures the proportion of samples whose Y will change from 1 to 0 if we set $\mathbf{X_S} = \mathbf{0}_{|\mathbf{S}|}$ when $\mathbf{X_S} = \mathbf{1}_{|\mathbf{S}|}$ is observed. In other words, CR calculates the proportion of people whose Y will happen if $\mathbf{X_S} = \mathbf{0}_{|\mathbf{S}|}$ happens and Y will not happen if $\mathbf{X_S} = \mathbf{0}_{|\mathbf{S}|}$ does not happen. The higher the proportion of this kind of people, the greater the impact of $\mathbf{X_S}$ on Y. In Table 1, CR1-o and CR2-o are the average change rate of different methods for attributing one and two causes without interaction effect, while CR2-w is the change rate of mathods for attributing two causes with interaction effect.

Table 1. Change rate of methods for attribution with simulated data.

	Rand	Post	PN	PS	PNS	ACE	PostTCE	CCCE
CR1-o	0.239	0.450	0.460	0.444	0.460	0.452	0.664	**0.671**
std.	(0.014)	(0.028)	(0.025)	(0.042)	(0.028)	(0.028)	(0.029)	(0.030)
CR2-o	0.384	0.671	0.643	0.642	0.639	0.671	0.703	**0.740**
std.	(0.014)	(0.036)	(0.026)	(0.027)	(0.027)	(0.035)	(0.028)	(0.020)
CR2-w	0.396	0.609	0.563	0.570	0.571	0.570	0.629	**0.709**
std.	(0.016)	(0.038)	(0.034)	(0.035)	(0.036)	(0.039)	(0.023)	(0.025)

In the case without interaction effect, recall that when only one cause is attributed, CCCE degenerates into PostTCE. It can be seen that PostTCE and CCCE perform much better than other methods and have similar change rates. In addition, when two causes need to be attributed, due to monotonicity, all methods perform better than attributing one cause, and PostTCE and CCCE are still much better than others. Besides, CR2 of CCCE is about 3.7% higher than that of PostTCE, as counterfacual probabilities of all causes are generated independently and randomly, and there is no interaction effect in this case. In the case with interaction effect, it can be seen that the change rate of CCCE (0.709) is the highest, at least 8% higher than other methods.

Real Data

In this experiment, we apply our method to a real world dataset about the expression levels of proteins and phospholipids in Sachs (2005). The ground truth causal graph has 11 vertices and 21 edges. We aim to attribute the expression of the variable Mek. Here we only use the observational data with 853 samples for our attribution. We binarize the data by setting the data greater than the median value to 1, and 0 otherwise. The results are shown in Table 2.

Table 2. Change rate of methods for attributing the expression of Mek with real data.

	Rand	Post	PN	PS	PNS	ACE	PostTCE	CCCE
CR1	0.183	0.0.534	0.526	0.538	0.532	0.532	0.632	**0.648**
std.	(0.023)	(0.018)	(0.018)	(0.012)	(0.022)	(0.014)	(0.011)	(0.019)
CR2	0.325	0.0.537	0.555	0.546	0.552	0.553	0.650	**0.703**
std.	(0.020)	(0.011)	(0.020)	(0.022)	(0.026)	(0.021)	(0.019)	(0.021)

Similarly, CR1 in Table 2 is the average change rate for attributing one cause. Again, PostTCE and CCCE outperform other methods. Note that the change rates of PostTCE and CCCE are close. As for CR2, the average change rate for attributing two causes, CCCE performs best and its change rate is 5.3% higher than PostTCE. This implies that there is an interaction effect between the causes of Mek.