Interaction Testing in Variation Analysis

Drago Plečko
Department of Computer Science
Columbia University
New York, NY 10027, USA
dp3144@columbia.edu

Abstract

Relationships of cause and effect are of prime importance for explaining scientific phenomena. Often, rather than just understanding the effects of causes, researchers may also wish to understand how a cause X affects an outcome Y mechanistically – i.e., what are the causal pathways that are activated between X and Y. For analyzing such questions, a range of methods has been developed over decades under the rubric of causal mediation analysis. Traditional mediation analysis focuses on decomposing the average treatment effect (ATE) into direct and indirect effects, and therefore focuses on the ATE as the central quantity. This corresponds to providing explanations for associations in the interventional regime, such as when the treatment X is randomized. Often, however, researchers may be interested in explaining associations in the natural, observational regime. In this paper we introduce variation analysis, an extension of mediation analysis that focuses on the total variation (TV) measure between X and Y, written as $\mathbb{E}[Y \mid x_1] - \mathbb{E}[Y \mid x_0]$. The TV measure encompasses both causal and confounded effects, and is therefore suitable for providing explanations in the natural regime. Our focus is on decomposing the TV measure, in a way that explicitly includes interaction terms between causal and spurious pathways. We then introduce the concept of interaction testing, which involves hypothesis tests to determine if interaction terms are significantly different from zero. If interactions are not significant, a more parsimonious decomposition of the TV measure can be used. The paper further provides a structural basis for these interaction tests and demonstrates their applicability through algorithmic and empirical analyses. Additionally, the implications of the framework to risk ratio and odds ratio scales for binary outcomes is discussed, offering a comprehensive approach to understanding the interplay of direct, indirect, and spurious effects in causal inference.

1. Introduction

Understanding relationships of cause and effect is one of the fundamental tasks found throughout the sciences. The process of establishing mechanistic links between causes and their consequences is at the core of our ability to explain why events occur as they do. In this context, mediation analysis, a widely used tool, helps unravel the pathways through which causal effects are transmitted. By identifying intermediary variables, mediation analysis offers deeper insights into the underlying mechanisms driving the observed cause-effect relationships. This approach is crucial in fields ranging from epidemiology to social sciences, where understanding the nuances of causal relationships can inform interventions and policy decisions.

Interestingly, most of the literature on mediation analysis focuses on understanding the variations contained in the average treatment effect (ATE), also known as the total effect (TE), given by

$$\mathbb{E}[y \mid do(x_1)] - \mathbb{E}[y \mid do(x_0)], \tag{1}$$

where $do(\cdot)$ symbolizes the do-operator (Pearl, 2000), and x_0, x_1 are two distinct values attained by the variable X. Instead of just quantifying the causal effect, researchers are more broadly interested in determining which causal mechanisms transmit the causal influences from X to Y. Various

(a) Causal diagram of mediation.

(b) Causal diagram with observed confounding.

Figure 1: Causal diagrams for Ex. 1.

approaches for solving this problem have been proposed under the rubric of causal mediation analysis (Baron and Kenny, 1986; Robins and Greenland, 1992; Pearl, 2001; VanderWeele, 2015). A common goal for many of the mediation methods is to *decompose variations* that are contained in the ATE into variations that mediated by other variables (known as the indirect or mediated effect) and variations that are not mediated by other variables (known as the direct effect).

Interestingly, mediation analysis focuses solely on causal variations, and the ATE captures the association of X and Y in an interventional regime, such as the randomized control trial (RCT), in which values of X are randomized to either x_0 or x_1. This approach has proven tremendously useful for explaining causal effects, such as in testing the effects of drugs, understanding the impact of educational interventions, and evaluating policy changes. Often, however, researchers may be interested in explaining the association of X and Y in the natural, observational regime, without a specific intervention in mind. Somewhat surprisingly, the common approach for mediation falls short of answering simple question such as "why do patients receiving chemotherapy have higher mortality rates than those not receiving it?", or "why do coffee drinkers have higher rates of cardiovascular disease?". In both of the examples, the causal relationship may account for only a part of the observed association, while non-causal (or spurious/confounded) effects also play an important role in explaining the phenomenon. In the former example, illness severity increases both the probability of receiving chemotherapy and dying, while in the latter, coffee drinkers are more likely to also smoke, which is known determinant for cardiovascular disease. When explaining associations in the natural, observational regime, we may be interested in the quantity

$$\mathbb{E}[y \mid x_1] - \mathbb{E}[y \mid x_0], \tag{2}$$

which we will refer to as the total variation (TV) measure, instead of the typically used ATE. The key difference between the ATE and the TV is that the latter encompasses confounded variations and not just the causal ones. As the approaches focusing on the ATE were called mediation analysis, in this paper, we call the approaches focusing on the TV measure *variation analysis*, to provide a distinction for a class of methods also concerned with effects that are neither direct nor mediated, but confounded.

While mediation analysis has been successful in quantifying direct and indirect effects, far fewer approaches for handling *interactions* between direct and indirect effects have been proposed (VanderWeele, 2015). Furthermore, in the context of variation analysis, no methods exist to date that are intended for analyzing interactions, such as the interactions of spurious and causal effects. In the following example, we motivate some of the key developments appearing in this paper:

Example 1 (Total Effect and Total Variation Decompositions) *Consider the causal diagram in Fig. 1a, with treatment X, outcome Y, and mediator W. A key result from Pearl (2001) demonstrated that the average treatment effect (ATE) can be decomposed into*

$$\mathbb{E}[y \mid do(x_1)] - \mathbb{E}[y \mid do(x_0)] = \underbrace{\mathbb{E}[Y_{x_1,W_{x_0}} - Y_{x_0,W_{x_0}}]}_{\textit{natural direct effect}} - \underbrace{\mathbb{E}[Y_{x_1,W_{x_0}} - Y_{x_1,W_{x_1}}]}_{\textit{natural indirect effect}}. \tag{3}$$

The natural direct effect (NDE) compares the potential outcome $Y_{x_0,W_{x_0}}$ in which $X = x_0$ along both direct ($X \to Y$) and indirect ($X \to W \to Y$) pathways, vs. the potential outcome $Y_{x_1,W_{x_0}}$ in

which $X = x_1$ along the direct pathway, while $X = x_0$. In this way, the NDE quantifies the effect of changing X from x_0 to x_1 along the direct pathway. To obtain the ATE, from the NDE, one subtracts the natural indirect effect (NIE) with a reverse transition, which measures the effect of changing $X = x_1$ to $X = x_0$ along the indirect path by comparing potential outcomes $Y_{x_1,W_{x_1}}$ vs. $Y_{x_1,W_{x_0}}$. Often, the need to consider a transition $x_0 \to x_1$ in the NDE while subtracting a reverse transition $x_1 \to x_0$ for the NIE is criticized as a shortcoming of the widely used effect decomposition in Eq. 3.

A result of VanderWeele (2015) sheds lights on the issue of opposite transitions appearing in the decomposition of the ATE. In particular, the ATE can be decomposed in a different way:

$$ATE_{x_0,x_1}(y) = \underbrace{\mathbb{E}[Y_{x_1,W_{x_0}} - Y_{x_0,W_{x_0}}]}_{\textit{natural direct effect}} + \underbrace{\mathbb{E}[Y_{x_0,W_{x_1}} - Y_{x_0,W_{x_0}}]}_{\textit{natural indirect effect}} \tag{4}$$

$$+ \underbrace{\mathbb{E}[Y_{x_1,W_{x_0}} - Y_{x_0,W_{x_0}} - (Y_{x_1,W_{x_1}} - Y_{x_0,W_{x_1}})]}_{\textit{interaction effect}}. \tag{5}$$

Notably, in Eq. 4, the NDE and the NIE both appear with a transition of $x_0 \to x_1$, with a baseline of $Y_{x_0,W_{x_0}}$. There is an additional term, however, which compares how the direct effect changes according to the behavior of W, namely

$$\mathbb{E}[\underbrace{Y_{x_1,W_{x_0}} - Y_{x_0,W_{x_0}}}_{x_0 \to x_1 \ \textit{DE with}\ W_{x_0}} - \underbrace{(Y_{x_1,W_{x_1}} - Y_{x_0,W_{x_1}})}_{x_0 \to x_1 \ \textit{DE with}\ W_{x_1}}]. \tag{6}$$

In other words, the interactive effect compares how strongly the direct effect of a $x_0 \to x_1$ transition changes in the setting of W_{x_0} vs. W_{x_1}. The result of VanderWeele (2015) demonstrates that the key issue in interpreting Pearl's decomposition is due to interactions.

The total variation measure, when considering the causal diagram in Fig. 1b, can be decomposed as (Zhang and Bareinboim, 2018; Plečko and Bareinboim, 2024):

$$\mathbb{E}[y \mid x_1] - \mathbb{E}[y \mid x_0] = \underbrace{\mathbb{E}[Y_{x_1} - Y_{x_0} \mid x_0]}_{\textit{counterfactual total effect}} - \underbrace{\mathbb{E}[Y_{x_1} \mid x_0] - \mathbb{E}[Y_{x_1} \mid x_1]}_{\textit{counterfactual spurious effect}} \tag{7}$$

The counterfactual total effect (also known as the effect of treatment on the treated, ETT) computes the effect of a $x_0 \to x_1$ transition for the subpopulation of individuals with $X = x_0$. To obtain the TV, from this effect we subtract the counterfactual spurious effect which compares how conditioning on $X = x_0$ differs from $X = x_1$ while setting $X = x_1$ along the causal pathways. Similarly as in Eq. 3, the two effects appear with reverse transitions.

The problem can be remedied again, by noting that the TV can also be decomposed as:

$$TV_{x_0,x_1}(y) = \underbrace{\mathbb{E}[Y_{x_1} - Y_{x_0} \mid x_0]}_{\textit{counterfactual total effect}} + \underbrace{\mathbb{E}[Y_{x_0} \mid x_1] - \mathbb{E}[Y_{x_0} \mid x_0]}_{\textit{counterfactual spurious effect}} \tag{8}$$

$$+ \underbrace{\mathbb{E}[Y_{x_1} - Y_{x_0} \mid x_1] - \mathbb{E}[Y_{x_1} - Y_{x_0} \mid x_0]}_{\textit{causal/spurious interaction}}. \tag{9}$$

The additional term appearing in this new decomposition compares

$$\underbrace{\mathbb{E}[Y_{x_1} - Y_{x_0} \mid x_1]}_{x_0 \to x_1 \ \textit{TE with}\ X = x_1 \ \textit{conditioning}} \quad \textit{vs.} \quad \underbrace{\mathbb{E}[Y_{x_1} - Y_{x_0} \mid x_0]}_{x_0 \to x_1 \ \textit{TE with}\ X = x_0 \ \textit{conditioning}} \tag{10}$$

and quantifies how much the total effect of a $x_0 \to x_1$ transition changes when conditioning on $X = x_1$ vs. $X = x_0$. In this way, we can measure the strength of the interaction between spurious and causal paths. □

Our goal in this manuscript will be to provide a coherent umbrella for understanding interactions of causal pathways in variation analysis, namely direct, indirect, and spurious. Our contributions are the following:

(i) We prove a first decomposition of the total variation (TV) measure that contains an explicit term for the interaction of causal and spurious pathways (Thm. 1),

(ii) We develop the concept of *interaction testing* (Def. 10), in which an explicit interaction term appearing in a decomposition is subject to a hypothesis test of being equal to 0,

(iii) We develop an algorithm (Alg. 1) for testing for interactions in a non-parametric way, showing a common preference for parsimony in statistics: if the interaction term is not significantly different from 0, a more parsimonious TV decomposition may be used.

(iv) We relate the effect interactions to the structural causal mechanisms of the underlying system (Def. 11), and demonstrate when there is a correspondence between a mechanism property and the corresponding interaction test (Prop. 2),

(v) We provide the most general decomposition of the TV measure that provides accounts for all interactions between direct, indirect, and spurious effects (Thm. 2),

(vi) We translate our results for the risk ratio and odds ratio scales (Sec. 5),

(vii) We perform an in-depth empirical analysis with the attempt to discover how commonly effects interact in practice (Sec. 6).

2. Preliminaries

We use the language of structural causal models (SCMs) as our basic semantical framework (Pearl, 2000). A structural causal model (SCM) is defined as:

Definition 1 (Structural Causal Model (SCM) (Pearl, 2000)) *A structural causal model $\mathcal{M}$ is a 4-tuple $\langle V, U, \mathcal{F}, P(u)\rangle$, where*

1. *U is a set of exogenous variables, also called background variables, that are determined by factors outside the model;*
2. *$V = \{V_1, ..., V_n\}$ is a set of endogenous (observed) variables, that are determined by variables in the model (i.e. by the variables in $U \cup V$);*
3. *$\mathcal{F} = \{f_{V_1}, ..., f_{V_n}\}$ is the set of structural functions determining V, $v_i \leftarrow f_{V_i}(\text{pa}(v_i), u_i)$, where $\text{pa}(V_i) \subseteq V \setminus V_i$ and $U_i \subseteq U$ are the functional arguments of f_{V_i};*
4. *$P(u)$ is a distribution over the exogenous variables U.*

□

The assignment mechanisms $\mathcal{F}$ determine how each of the observed variables V_i attains its value, based on other observed variables and the latent variables U. Together with the probability distribution $P(u)$ over the exogenous variables U, it specifies the entire behavior of the underlying phenomenon. In particular, the SCM also specifies the *observational distribution* of the underlying phenomenon, defined through:

Definition 2 (Observational Distribution (Bareinboim et al., 2022)) *An SCM $\mathcal{M}$ that is a 4-tuple $\langle V, U, \mathcal{F}, P(u)\rangle$ induces a joint probability distribution $P(V)$ such that for each $Y \subseteq V$,*

$$P^{\mathcal{M}}(y) = \sum_{u} \mathbb{1}\Big(Y(u) = y\Big) P(u), \tag{11}$$

where $Y(u)$ is the solution for Y after evaluating $\mathcal{F}$ with $U = u$. □

A further important notion building on the concept of the SCM is that of a submodel, which is defined next:

Definition 3 (Submodel (Pearl, 2000)) *Let $\mathcal{M}$ be a structural causal model, X a set of variables in V, and x a particular value of X. A submodel $\mathcal{M}_x$ (of $\mathcal{M}$) is a 4-tuple:*

$$\mathcal{M}_x = \langle V, U, \mathcal{F}_x, P(u) \rangle \tag{12}$$

where

$$\mathcal{F}_x = \{f_i : V_i \notin X\} \cup \{X \leftarrow x\}, \tag{13}$$

and all other components are preserved from $\mathcal{M}$. □

Building on submodels, we introduce next the notion of a potential outcome:

Definition 4 (Potential Outcome / Response (Rubin, 1974; Pearl, 2000)) *Let X and Y be two sets of variables in V and $u \in \mathcal{U}$ be a unit. The potential outcome/response $Y_x(u)$ is defined as the solution for Y of the set of equations $\mathcal{F}_x$ evaluated with $U = u$. That is, $Y_x(u)$ denotes the solution of Y in the submodel $\mathcal{M}_x$ of $\mathcal{M}$.* □

In words, $Y_x(u)$ is the value variable Y would take if (possibly contrary to observed facts) X is set to x, for a specific unit u. We further define how counterfactual distributions over various possible potential outcomes are computed:

Definition 5 (Counterfactual Distributions (Bareinboim et al., 2022)) *Consider an SCM $\mathcal{M} = \langle V, U, \mathcal{F}, P(u) \rangle$, and let $Y_1, \ldots, Y_k \subset V$, and $X_1, \ldots, X_k \subset V$ be subsets of the observables, and let $x_1, \ldots, x_k$ be specific values of X_is. Denote by $(Y_i)_{x_i}$ the potential response of variables Y_i when setting $X_i = x_i$. The SCM $\mathcal{M}$ induces a family of joint distributions over counterfactual events $(Y_1)_{x_1}, \ldots, (Y_k)_{x_k}$:*

$$P^{\mathcal{M}}((y_1)_{x_1}, \ldots, (y_k)_{x_k}) = \sum_u \mathbb{1}\Big(\bigwedge_{i=1}^{k} (Y_i)_{x_i}(u) = y_i \Big) P(u). \tag{14}$$

□

The l.h.s. in Eq. 14 contains variables with different subscripts, which syntactically represent different potential responses (Def. 4), or counterfactual worlds. Finally, there is one more prerequisite notion for our discussion. The mechanisms $\mathcal{F}$ and the distribution over the exogenous variables $P(u)$ are almost never observed. However, to perform causal inference, we need a way of encoding assumptions about the underlying SCM. A common way of doing so is through an object called a causal diagram, which is defined next:

Definition 6 (Causal Diagram (Pearl, 2000; Bareinboim et al., 2022)) *Let an SCM $\mathcal{M}$ be a 4-tuple $\langle V, U, \mathcal{F}, P(u) \rangle$. A graph $\mathcal{G}$ is said to be a causal diagram (of $\mathcal{M}$) if:*

(1) there is a vertex for every endogenous variable $V_i \in V$,

(2) there is an edge $V_i \rightarrow V_j$ if V_i appears as an argument of $f_j \in \mathcal{F}$,

(3) there is a bidirected edge $V_i \dashleftarrow\dashrightarrow V_j$ if the corresponding $U_i, U_j \subset U$ are correlated or the corresponding functions f_i, f_j share some $U_{ij} \in U$ as an argument.

□

We call $\text{pa}(V_i)$ the set of parents of V_i, while the sets of children $\text{ch}(V_i)$, ancestors $\text{an}(V_i)$, and descendants $\text{de}(V_i)$ are defined analogously.

Building on the notion of a potential response, one can further define the notions of counterfactual and factual contrasts, given by:

Definition 7 (Contrasts (Plečko and Bareinboim, 2024)) *Given an SCM $\mathcal{M}$, a contrast $\mathcal{C}$ is any quantity of the form*

$$\mathcal{C}(C_0, C_1, E_0, E_1) = \mathbb{E}[y_{C_1} \mid E_1] - \mathbb{E}[y_{C_0} \mid E_0], \tag{15}$$

where E_0, E_1 are observed (factual) clauses and C_0, C_1 are counterfactual clauses to which the outcome Y responds. Furthermore, whenever

(a) $E_0 = E_1$, the contrast $\mathcal{C}$ is said to be counterfactual;

(b) $C_0 = C_1$, the contrast $\mathcal{C}$ is said to be factual.

□

For instance, the contrast $(C_0 = \{x_0\}, C_1 = \{x_1\}, E_0 = \emptyset, E_1 = \emptyset)$ corresponds to the *average treatment effect (ATE)* $\mathbb{E}[y_{x_1} - y_{x_0}]$. Similarly, the contrast $(C_0 = \{x_0\}, C_1 = \{x_1\}, E_0 = \{x_0\}, E_1 = \{x_0\})$ corresponds to the *effect of treatment on the treated (ETT)* $\mathbb{E}[y_{x_1} - y_{x_0} \mid x_0]$. Many other important causal quantities can be represented as contrasts, as exemplified later on.

Proposition 1 (Structural Basis Expansion (Plečko and Bareinboim, 2024)) *Counterfactual and factual contrasts admit the following structural basis expansions, respectively:*

(a) Counterfactual contrast ($\mathcal{C}_{ctf}$), where $E_0 = E_1 = E$, can be expanded as

$$P(y_{C_1} \mid E) - P(y_{C_0} \mid E) = \sum_u \big(\underbrace{\mathbb{1}(Y_{C_1}(u) = y) - \mathbb{1}(Y_{C_0}(u) = y)}_{\textit{unit-level difference}} \big) \times \underbrace{P(u \mid E)}_{\textit{posterior}}, \tag{16}$$

(b) Factual contrast ($\mathcal{C}_{factual}$), where $C_0 = C_1 = C$, can be expanded as

$$P(y_C \mid E_1) - P(y_C \mid E_0) = \sum_u \underbrace{\mathbb{1}(Y_C(u) = y)}_{\textit{unit outcome}} \big(\underbrace{P(u \mid E_1) - P(u \mid E_0)}_{\textit{posterior difference}} \big). \tag{17}$$

□

As noted by Plečko and Bareinboim (2024), the event E in Eq. 16 can be used for controlling the granularity of the population for which the effect is quantified. For instance, the choice of $C_0 = x_0, C_1 = x_1$, and $E = \emptyset$ which yields the ATE (as discussed above) is simply an average of the unit-level difference $y_{x_1}(u) - y_{x_0}(u)$ over all units u of the population

$$\text{ATE}_{x_0, x_1}(y) = \sum_u (y_{x_1}(u) - y_{x_0}(u)) P(u). \tag{18}$$

The ETT, which is obtained with the same C_0, C_1 but with $E = \{X = x\}$, is a more granular notion of a total effect:

$$\text{ETT}_{x_0, x_1}(y \mid x) = \sum_u (y_{x_1}(u) - y_{x_0}(u)) P(u \mid x). \tag{19}$$

One is still averaging the unit-level difference $y_{x_1}(u) - y_{x_0}(u)$ but over the subset of units that have $X(u) = x$. The weight given to each unit is given by $P(u \mid X = x)$, which corresponds to the probability mass of the unit $U = u$ in the event $X = x$. Further, even more granular quantifications of the direct effect, are also possible, and Plečko and Bareinboim (2024) also discuss conditioning on $E = \{X = x, Z = z\}$, all the way to conditioning on all the observables, $E = \{X = x, Z = z, W = w, Y = y\}$. At each level of granularity, a more localized notion of total effect is possible. It is also important to note that controlling the granularity of the population works in the same for

quantifying direct and indirect effects, for which different choices of C_0, C_1 are used (in particular, for the direct effect one can use $C_0 = \{x_0\}, C_1 = \{x_1, W_{x_0}\}$).

The notion of a spurious effect behaves somewhat differently, since the decomposition in Eq. 17 shows that variations are introduced by a difference in the posterior distribution over the population. In this way, for instance, Plečko and Bareinboim (2024) consider the x-specific spurious effect x-$\text{SE}_{x_0,x_1}(y)$ defined as

$$\mathbb{E}[y_{x_0} \mid x_1] - \mathbb{E}[y_{x_0} \mid x_0] = \sum_u y_{x_0}(u) \left(P(u \mid x_1) - P(u \mid x_0)\right). \tag{20}$$

The quantity measure the change in Y resulting from a change conditioning on $X = x_0$ to $X = x_1$, while setting $X = x_0$ along all causal pathways.

Importantly, the previous work of Plečko and Bareinboim (2024) considers only first-order contrasts, and in this paper, we generalize this notion to include higher-order contrasts which are particularly useful for quantifying interaction effects. Furthermore, in Sec. 4.1, we investigate how interactions can be analyzed at different levels of granularity of the population.

3. Interaction Analysis

In this section, we introduce the key concepts of interaction analysis. Throughout the section, we will assume that the data we are analyzing is compatible with the causal diagram in Fig. 2 (a variation of this graph was termed the *standard fairness model (SFM)* in the work of Plečko and Bareinboim (2024)). In the diagram, variables Z and W can be thought of as multi-dimensional. We begin by introducing the distinction between first order, second order, and third order contrasts. We then instantiate some well-known first and second-order contrasts, and demonstrate how these contrasts can be used for decomposing the TV measure. Subsequently, we explain how such contrasts can also be used for non-parametric interaction testing. The notion of interaction testing is then paired with the idea that, if interactions are not present in the data, a more parsimonious representation of the TV measure can be obtained, which may allow the practitioner to the interpret the effects in a simpler way.

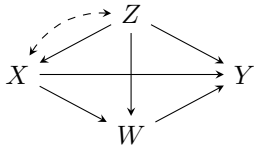


Figure 2: Markovian causal diagram with mediators W and confounders Z.

Definition 8 (First, Second, and Third Order Contrasts) *Let C_1, C_0, C, C' be counterfactual clauses, and E_0, E_1 be observed events. We categorize the following types of contrasts:*

(1) A first order causal contrast $\mathcal{C}^1(C_0, C_1, E, E)$, of the form

$$\sum_u \left[y_{C_1}(u) - y_{C_0}(u)\right] \times P(u \mid E). \tag{21}$$

(2) A first order spurious contrast $\mathcal{C}^1(C, C, E_0, E_1)$, of the form

$$\sum_u y_C(u) \times \left[P(u \mid E_1) - P(u \mid E_0)\right]. \tag{22}$$

(3) A second order causal-causal contrast $\mathcal{C}^2(C_0, C_1, C, C', E, E)$, of the form

$$\sum_u \left[(y_{C_1,C'}(u) - y_{C_0,C'}(u)) - (y_{C_1,C}(u) - y_{C_0,C}(u))\right] \times P(u \mid E). \tag{23}$$

(4) A second order causal-spurious contrast $\mathcal{C}^2(C_0, C_1, E_0, E_1)$, of the form

$$\sum_u \big[y_{C_1}(u) - y_{C_0}(u)\big] \times \big[P(u \mid E_1) - P(u \mid E_0)\big]. \tag{24}$$

(5) A third order causal-causal-spurious contrast $\mathcal{C}^3(C_0, C_1, C, C', E_0, E_1)$, of the form

$$\sum_u \big[(y_{C_1,C}(u) - y_{C_0,C}(u)) - (y_{C_1,C'}(u) - y_{C_0,C'}(u))\big] \times \big[P(u \mid E_1) - P(u \mid E_0)\big]. \tag{25}$$

□

Definition 9 (x-specific Effects) *The x-{total, direct, indirect, spurious} effects are defined as follows:*

$$x\text{-}TE_{x_0,x_1}(y \mid x) = P(y_{x_1} \mid x) - P(y_{x_0} \mid x) \tag{26}$$
$$x\text{-}DE_{x_0,x_1}(y \mid x) = P(y_{x_1,W_{x_0}} \mid x) - P(y_{x_0} \mid x) \tag{27}$$
$$x\text{-}IE_{x_0,x_1}(y \mid x) = P(y_{x_0,W_{x_1}} \mid x) - P(y_{x_0} \mid x) \tag{28}$$
$$x\text{-}SE_{x_0,x_1}(y) = P(y_{x_0} \mid x_1) - P(y_{x_0} \mid x_0). \tag{29}$$

The interactive effects x-direct-indirect and x-total-spurious are defined as:

$$x\text{-}DE\text{-}IE_{x_0,x_1}(y \mid x) = [P(y_{x_1,W_{x_0}} \mid x) - P(y_{x_0} \mid x)] - [P(y_{x_1,W_{x_1}} \mid x) - P(y_{x_0,W_{x_1}} \mid x)] \tag{30}$$
$$x\text{-}TE\text{-}SE_{x_0,x_1}(y) = [P(y_{x_1} \mid x_1) - P(y_{x_0} \mid x_1)] - [P(y_{x_1} \mid x_0) - P(y_{x_0} \mid x_0)]. \tag{31}$$

□

We now look at the structural basis expansion of the interactive effects, to understand what is being computed. The effect x-DE-IE$_{x_0,x_1}(y \mid x)$ can be written as:

$$x\text{-DE-IE}_{x_0,x_1}(y \mid x) = \sum_u [(y_{x_1,W_{x_0}} - y_{x_0,W_{x_0}})(u) - (y_{x_1,W_{x_1}} - y_{x_0,W_{x_1}})(u)] \times P(u \mid x). \tag{32}$$

The unit level difference appearing in Eq. 32 equals

$$\underbrace{y_{x_1,W_{x_0}}(u) - y_{x_0,W_{x_0}}(u)}_{x_0 \to x_1 \text{ DE with } W_{x_0}} - \underbrace{y_{x_1,W_{x_1}}(u) - y_{x_0,W_{x_1}}(u)}_{x_0 \to x_1 \text{ DE with } W_{x_1}}, \tag{33}$$

and can thus be understood as how much changing W_{x_0} to W_{x_1} modifies the size of the direct effect of the transition $x_0 \to x_1$. Due to symmetry, this unit level difference can be also be written as:

$$\underbrace{y_{x_0,W_{x_1}}(u) - y_{x_0,W_{x_0}}(u)}_{x_0 \to x_1 \text{ IE with } X = x_0} - \underbrace{y_{x_1,W_{x_1}}(u) - y_{x_1,W_{x_0}}(u)}_{x_0 \to x_1 \text{ IE with } X = x_1}. \tag{34}$$

In words, the interactive effect can also be understood as how much changing x_0 to x_1 along the direct effect modifies the sizes of the indirect effect of a $x_0 \to x_1$ transition.

We now turn to understanding the total-spurious interaction effect, which can be written using the structural basis expansion as:

$$x\text{-TE-SE}_{x_0,x_1}(y) = \sum_u \big[y_{x_1}(u) - y_{x_0}(u)\big] \times \big[P(u \mid x_1) - P(u \mid x_0)\big]. \tag{35}$$

The unit level difference $y_{x_1}(u) - y_{x_0}(u)$ is integrated against the difference in posterior weighing distributions $P(u \mid x_1) - P(u \mid x_0)$, and thus captures the interaction of the total and spurious effects.

Armed with the structural understanding of different (interaction) effects, we can now decompose the TV measure:

Theorem 1 (TV Decomposition with Interactions) *The total variation (TV) measure can be decomposed as:*

$$TV_{x_0,x_1}(y) = x\text{-}TE_{x_0,x_1}(y \mid x_0) + x\text{-}SE_{x_0,x_1}(y) + x\text{-}TE\text{-}SE_{x_0,x_1}(y). \tag{36}$$

Furthermore, the TV measure can also be decomposed as:

$$\begin{aligned} TV_{x_0,x_1}(y) = {} & x\text{-}DE_{x_0,x_1}(y \mid x_0) + x\text{-}IE_{x_0,x_1}(y \mid x_0) + x\text{-}DE\text{-}IE_{x_0,x_1}(y \mid x_0) \\ & + x\text{-}SE_{x_0,x_1}(y) + x\text{-}TE\text{-}SE_{x_0,x_1}(y). \end{aligned} \tag{37}$$

□

Thm. 1 represents the first major result of the paper. The TV measure is decomposed to feature explicit interaction effects of the direct/indirect pathways, and also the total/spurious pathways. Furthermore, all the transitions appearing for direct, indirect, and spurious effects are $x_0 \to x_1$, and there are no subtractions necessary (i.e., all the effects are added up). This is another major benefit when interpreting the decomposition.

3.1 Interaction Testing

We next move to another important aspect of variation analysis, called interaction testing. We

Definition 10 (Interaction Test) *Let $\mathcal{C}$ be any contrast that measures an interaction effect, and suppose $\mathcal{C}$ can be uniquely computed from the causal diagram in Fig. 2 and observational data $\mathcal{D}$. We say that a hypothesis test of the form*

$$H_0 : \mathcal{C}(\mathcal{D}) = 0, \tag{38}$$

is called an interaction test. □

The reasoning behind interaction testing is rather simple. Generally, interactions make the interpretation of causal effects more difficult. In statistics, there is often a preference for parsimony, whenever such parsimony is justifiable. Therefore, we propose an approach in which an interaction effect $\mathcal{C}$ is tested against 0 based on some available data $\mathcal{D}$. If there is no evidence of the effect being different from 0, we may use a more parsimonious version of the TV decomposition. Before providing an algorithm for doing so, we provide a structural account of why interaction tests.

Definition 11 (Structural Interaction Criteria) *Consider the causal diagram in Fig. 1b, and let $f_y(x, z, u_y)$ be the structural mechanism of the Y variable. We say that there is no structural interaction of total and spurious effects, written Str-TE-SE = 0, if either of the following hold:*

(i) we can write the mechanism $f_y(x, z, u_y)$ as

$$f_y(x, z, u_y) = f_y^{(1)}(x, u_y) + f_y^{(2)}(z, u_y). \tag{39}$$

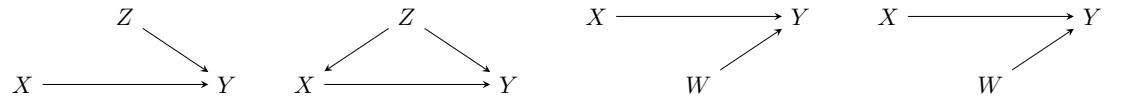
(ii) there is no back-door path between X and Y.

Furthermore, we say that there is no structural interaction of direct and indirect effects in the causal diagram in Fig. 2, written Str-DE-IE = 0, if either of the following hold:

(i) we can write the mechanism $f_y(x, z, w, u_y)$ as

$$f_y(x, z, w, u_y) = f_y^{(1)}(x, z, u_y) + f_y^{(2)}(w, z, u_y). \tag{40}$$

(ii) X is not an input to the mechanism f_w of W.



(a) No back-door path. (b) No SE-TE interaction. (c) No indirect path. (d) No DE-IE interaction.

Figure 3: Causal diagrams used in Ex. 2.

If there are interactions, we say that the corresponding structural criterion is equal to 1. □

The definition of Str-TE-SE = 0, requires that there is no explicit functional term within f_y depending on both x, z (or any unobserved confounder u_c with a back-door path to X), or that there is no back-door path between X and Y. Similarly, the definition of Str-DE-IE requires that there is no functional term within f_y depending simultaneously on both x, w, or that f_w does not depend on the value of x. In Sec. 5, we discuss how the results of this section can be extended to risk ratio and odds ratio scales for binary outcomes. We now provide an example that illustrates the structural notions of interaction:

Example 2 (Structural Criteria for Interactions) *Consider the SCMs $\mathcal{M}_1, \mathcal{M}_2$ given by:*

$$\mathcal{M}_1 = \begin{cases} Z \leftarrow \textit{Bernoulli}(0.5) & (41) \\ X \leftarrow \textit{Bernoulli}(0.5) & (42) \\ Y \leftarrow X + Z + XZ, & (43) \end{cases}, \qquad \mathcal{M}_2 = \begin{cases} Z \leftarrow \textit{Bernoulli}(0.5) & (44) \\ X \leftarrow \textit{Bernoulli}(0.5 + 0.1Z) & (45) \\ Y \leftarrow X + Z, & (46) \end{cases}$$

which are compatible with the diagrams in Fig. 3a, 3b, respectively. In $\mathcal{M}_1$, there is no back-door path between X and Y, and hence there is no interaction between spurious and total effects, despite the term XZ in the f_y mechanism in Eq. 43. In $\mathcal{M}_2$, even though a back-door path between X and Y exists, there is no interaction between spurious and total effects since the f_y mechanism in Eq. 46 does not have a term involving both X and Z. However, if there was an additional term XZ, then the structural interaction criterion Str-TE-SE would evaluate to 1.

Consider further SCMs $\mathcal{M}_3, \mathcal{M}_4$ given by:

$$\mathcal{M}_1 = \begin{cases} X \leftarrow \textit{Bernoulli}(0.5) & (47) \\ W \leftarrow \textit{Bernoulli}(0.5) & (48) \\ Y \leftarrow X + W + XW, & (49) \end{cases}, \qquad \mathcal{M}_2 = \begin{cases} X \leftarrow \textit{Bernoulli}(0.5) & (50) \\ W \leftarrow \textit{Bernoulli}(0.5 + 0.1X) & (51) \\ Y \leftarrow X + W, & (52) \end{cases}$$

which are compatible with the diagrams in Fig. 3c, 3d, respectively. In $\mathcal{M}_3$, there is no indirect path $X \to W \to Y$, since f_w in Eq. 48 does not take X as input. Therefore, regardless of the fact that f_y in Eq. 49 contains a term XW, we still say there is no interaction of direct and indirect paths, written Str-DE-IE = 0. In $\mathcal{M}_4$, there exists an indirect path $X \to W \to Y$ (f_w in Eq. 51 takes X as an input), but the f_y in Eq. 52 mechanism does not contain a term involving both X, and W, and hence again Str-DE-IE = 0. However, if there was an additional XW in f_y in Eq. 52, then the structural interaction criterion Str-DE-IE would evaluate to 1. □

We can now prove an important result relating the structural mechanism f_y with interaction tests:

Proposition 2 (Structural Admissibility of Interaction Tests)

$$\textit{Str-TE-SE} = 0 \implies x\textit{-TE-SE}_{x_0,x_1}(y) = 0, \tag{53}$$

$$\textit{Str-DE-IE} = 0 \implies x\textit{-DE-IE}_{x_0,x_1}(y \mid x) = 0. \tag{54}$$

□

Algorithm 1 Interaction Testing for TV Decomposition

- **Inputs:** Causal Diagram $\mathcal{G}$, Observational Data $\mathcal{D}$

Compute the estimate of the total-spurious interaction effect x-TE-SE$_{x_0,x_1}(y)$, and its 95% confidence interval. Test the hypothesis

$$H_0^{\text{TE-SE}} : x\text{-TE-SE}_{x_0,x_1}(y) = 0. \tag{56}$$

Compute the estimate of the direct-indirect interaction effect x-DE-IE$_{x_0,x_1}(y \mid x_0)$, and its 95% confidence interval. Test the hypothesis

$$H_0^{\text{DE-IE}} : x\text{-DE-IE}_{x_0,x_1}(y \mid x_0) = 0. \tag{57}$$

- if neither $H_0^{\text{TE-SE}}, H_0^{\text{DE-IE}}$ are rejected, return the decomposition

$$\text{TV}_{x_0,x_1}(y) = x\text{-DE}_{x_0,x_1}(y \mid x_0) + x\text{-IE}_{x_0,x_1}(y \mid x_0) + x\text{-SE}_{x_0,x_1}(y) \tag{58}$$

- if only $H_0^{\text{DE-IE}}$ is rejected, return the decomposition

$$\begin{aligned}\text{TV}_{x_0,x_1}(y) = x\text{-DE}_{x_0,x_1}(y \mid x_0) + x\text{-IE}_{x_0,x_1}(y \mid x_0) + x\text{-DE-IE}_{x_0,x_1}(y \mid x_0) \\ + x\text{-SE}_{x_0,x_1}(y).\end{aligned} \tag{59}$$

- if only $H_0^{\text{TE-SE}}$ is rejected, return the decomposition

$$\text{TV}_{x_0,x_1}(y) = x\text{-DE}_{x_0,x_1}(y \mid x_0) + x\text{-IE}_{x_0,x_1}(y \mid x_0) + x\text{-SE}_{x_0,x_1}(y) + x\text{-TE-SE}_{x_0,x_1}(y).$$

- if both $H_0^{\text{TE-SE}}, H_0^{\text{DE-IE}}$ are rejected, return the TV decomposition in Eq. 37.

- **Output:** TV decomposition with parsimony.

In words, the proposition shows that whenever there is not total-spurious interaction at the structural level, the corresponding interaction effect x-TE-SE is also equal to 0. The same also holds for the direct-indirect interaction and the corresponding x-DE-SE interaction effect. The key of the proposition is in the contrapositive of the mentioned statements. That is, whenever we find that

$$x\text{-TE-SE}_{x_0,x_1}(y) \neq 0 \tag{55}$$

it implies that an interaction at the structural level must exist.

The proposed approach for interaction testing is presented in Alg. 1. We first test the hypothesis x-TE-SE$_{x_0,x_1}(y) = 0$, and if this hypothesis is not rejected, we can use a more parsimonious representation of the TV decomposition, in which the effect x-TE-SE$_{x_0,x_1}(y) = 0$ is removed. We then test the hypothesis x-DE-IE$_{x_0,x_1}(y \mid x_0) = 0$. If the hypothesis is not rejected, we can again use a more parsimonious TV decomposition in which the effect x-DE-IE$_{x_0,x_1}(y \mid x_0)$ is removed. More parsimonious decompositions, naturally, lead to more easily interpretable decompositions for the practitioners.

4. More Granular Interactions

So far, we described an approach for testing the interaction of total and spurious effects, and direct and indirect effects. In Tab. 1 we summarize the approach proposed so far. We considered two types of interactions tests, and we provide the implications of these tests for the TV decomposition.

Researchers, however, may be interested in an even more granular analysis: they may wish to distinguish between direct and spurious, indirect and spurious, or even direct-indirect-spurious

Interaction	Structural Test	Diagram	Interaction Test	Implication for Decomposition
TE $\otimes$ SE	no $f_y(x, z, u_y)$ term in f_y	Fig. 1b	$x\text{-SE}_{x_0,x_1}(y) = -x\text{-SE}_{x_1,x_0}(y)$	$\text{TV}_{x_0,x_1}(y) = x\text{-TE}_{x_0,x_1}(y \mid x_0) + x\text{-SE}_{x_0,x_1}(y)$
DE $\otimes$ IE	no $f_y(x, w, u_y)$ term in f_y	Fig. 1a	$x\text{-DE}_{x_0,x_1}(y \mid x) = -x\text{-DE}_{x_1,x_0}(y \mid x)$	$x\text{-TE}_{x_0,x_1}(y \mid x_0) = x\text{-DE}_{x_0,x_1}(y \mid x_0) + x\text{-IE}_{x_0,x_1}(y \mid x_0)$

Table 1: Summary table of interaction tests and their implications.

interactions. The following example illustrates why such an approach cannot be easily related to the structural mechanism f_y as we did in Prop. 2 for the TE-SE and DE-IE interactions:

Example 3 (DE-SE Interaction) *Consider the following structural causal model $\mathcal{M}$:*

$$Z \leftarrow N(U, 0) \tag{60}$$
$$X \leftarrow U \tag{61}$$
$$W \leftarrow \varepsilon_w + Z \tag{62}$$
$$Y \leftarrow W + XW^2, \tag{63}$$

where $U \sim$ Bernoulli(0.5) and $\varepsilon_w \sim N(0, 1)$. The model is compatible with the causal diagram in Fig. 2. When computing the quantity

$$\mathbb{E}[Y_{x_1,W_{x_0}} - Y_{x_0} \mid x_1] - \mathbb{E}[Y_{x_1,W_{x_0}} - Y_{x_0} \mid x_0], \tag{64}$$

we find that it equals 1. The quantity in Eq. 64 attempts to measure how much the direct effect of a $x_0 \to x_1$ transition, written $Y_{x_1,W_{x_0}} - Y_{x_0}$, changes for units with $X(u) = x_1$ vs. $X(u) = x_0$. Therefore, the quantity measure the interaction of direct and spurious paths, and in $\mathcal{M}$, it is different from 0. However, when looking at the f_y mechanism, we see that there is no term corresponding to the X, Z interaction. □

The reader will note the issue at hand – in the diagram in Fig. 2, the spurious effect of X on Y may be transmitted along two different paths: firstly, it may be transmitted along the $X \leftarrow Z \to Y$ path, but may also be transmitted along the $X \leftarrow Z \to W \to Y$ path. It is the latter path that is active in the SCM $\mathcal{M}$ in the above example. Previously, when testing the TE-SE interaction in the diagram in Fig. 1b, there was a single causal path $X \to Y$, and a single spurious path $X \leftarrow Z \to Y$. Similarly as for the direct and indirect paths in Fig. 1a (or Fig. 2), where there is a single direct $X \to Y$ path and a single indirect $X \to W \to Y$ path. This one-to-one correspondence between an effect (total/spurious or direct/indirect) and a causal path entering Y implied that the existence of an interaction can be easily determined at the structural level, by inspecting the f_y mechanism (see Def. 11) and checking for either the existence of a back-door path between X and Y (in case of TE-SE interaction) or a indirect path (in case of DE-IE interaction). In Ex. 3, since there are two spurious paths from X to Y, examining the mechanism f_y is no longer sufficient. Clearly, an interaction of spurious and direct paths, for instance, may also depend on the functional inputs of the f_w mechanism.

Therefore, understanding the interactions of direct/spurious, indirect/spurious, or even the direct/indirect/spurious effects will require a slightly more involved approach than for total/spurious and direct/indirect interactions. We begin by defining the more granular interactions at the structural level:

Definition 12 (Granular Structural Interaction Criteria) *Consider the causal diagram in Fig. 2, and let $f_y(x, z, w, u_y)$ be the structural mechanism of the Y variable. We say that there is no structural interaction of direct and spurious effects, written Str-DE-SE = 0, if either of the following hold:*

(i) we can write the mechanism $f_y(x, z, w, u_y)$ as

$$f_y(x, z, w, u_y) = f_y^{(1)}(x, u_y) + f_y^{(2)}(z, w, u_y). \tag{65}$$

(ii) there is no back-door path between X and Y.

We say that there is no structural interaction of indirect and spurious effects, written Str-IE-SE = 0, if either of the following hold:

(i) we can write the mechanism $f_y(x, z, w, u_y)$ as

$$f_y(x, z, w, u_y) = f_y^{(1)}(x, z, u_y) + f_y^{(2)}(z, w, u_y), \tag{66}$$

and we can write the mechanism $f_w(x, z, u_w)$ as

$$f_w(x, z, u_w) = f_w^{(1)}(x, u_w) + f_w^{(2)}(z, u_w). \tag{67}$$

(ii) there is no back-door path between X and Y.

(iii) there is no indirect path $X \to W \to Y$.

We say that there is no structural interaction of direct, indirect and spurious effects, written Str-DE-IE-SE = 0, if either of the following hold:

(i) we can write the mechanism $f_y(x, z, w, u_y)$ as

$$f_y(x, z, w, u_y) = f_y^{(1)}(x, z, u_y) + f_y^{(2)}(z, w, u_y) + f_w^{(3)}(x, w, u_w), \tag{68}$$

and we can write the mechanism $f_w(x, z, u_w)$ as

$$f_w(x, z, u_w) = f_w^{(1)}(x, u_w) + f_w^{(2)}(z, u_w). \tag{69}$$

(ii) we can write the mechanism $f_y(x, z, w, u_y)$ as

$$f_y(x, z, w, u_y) = f_y^{(1)}(x, z, u_y) + f_y^{(2)}(z, w, u_y). \tag{70}$$

(iii) there is no back-door path between X and Y.

(iv) there is no indirect path $X \to W \to Y$.

If there are interactions, we say that the corresponding structural criterion is equal to 1. □

We can now define the quantities that measure the interactions of the above-introduced effects.

Definition 13 (Granular Effect Interactions) *Consider the following effect interactions:*

$$x\text{-DE-SE}_{x_0,x_1}(y) = [P(y_{x_1,W_{x_0}} \mid x_1) - P(y_{x_0} \mid x_1)] - [P(y_{x_1,W_{x_0}} \mid x_0) - P(y_{x_0} \mid x_0)] \tag{71}$$

$$x\text{-IE-SE}_{x_0,x_1}(y) = [P(y_{x_0,W_{x_1}} \mid x_1) - P(y_{x_0} \mid x_1)] - [P(y_{x_0,W_{x_1}} \mid x_0) - P(y_{x_0} \mid x_0)] \tag{72}$$

$$x\text{-DE-IE-SE}_{x_0,x_1}(y) = \big[P(y_{x_1,W_{x_0}} \mid x_1) - P(y_{x_0} \mid x_1)\big] - [P(y_{x_1} \mid x_1) - P(y_{x_0,W_{x_1}} \mid x_1)\big] \tag{73}$$

$$- \big[[P(y_{x_1,W_{x_0}} \mid x_0) - P(y_{x_0} \mid x_0)] - [P(y_{x_1} \mid x_0) - P(y_{x_0,W_{x_1}} \mid x_0)]\big] \tag{74}$$

□

We can again obtain a better insight into the quantities by looking at their structural basis expansion, e.g.,

$$x\text{-DE-SE}_{x_0,x_1}(y) = \sum_u [y_{x_1,W_{x_0}}(u) - y_{x_0}(u)] \times [P(u \mid x_1) - P(u \mid x_0)]. \tag{75}$$

The unit level direct effect, $y_{x_1,W_{x_0}}(u) - y_{x_0}(u)$, is integrated against a posterior difference $P(u \mid x_1) - P(u \mid x_0)$ that induces a difference along the spurious path. In this way, we see that the quantity is a second-order causal-spurious contrast, measuring the interaction of the direct and spurious effects. For indirect-spurious interactions, we have the quantity

$$x\text{-DE-SE}_{x_0,x_1}(y) = \sum_u [y_{x_1,W_{x_0}}(u) - y_{x_0}(u)] \times [P(u \mid x_1) - P(u \mid x_0)]. \tag{76}$$

The interpretation is again very similar, where the unit level indirect effect $y_{x_0,W_{x_1}}(u) - y_{x_0}(u)$ is integrated against the posterior difference $P(u \mid x_1) - P(u \mid x_0)$. The quantity, therefore, captures the indirect-spurious interaction. Finally, for quantifying the three-way interaction, we can use the quantity

$$x\text{-DE-IE-SE}_{x_0,x_1}(y) = \sum_u [(y_{x_1,W_{x_0}} - y_{x_0,W_{x_0}})(u) - (y_{x_1,W_{x_1}} - y_{x_0,W_{x_1}})(u)] \tag{77}$$

$$\times [P(u \mid x_1) - P(u \mid x_0)]. \tag{78}$$

The quantity is therefore a third-order causal-causal-spurious contrast. The unit level interaction effect $(y_{x_1,W_{x_0}} - y_{x_0,W_{x_0}})(u) - (y_{x_1,W_{x_1}} - y_{x_0,W_{x_1}})(u)$ is integrated against the posterior difference $P(u \mid x_1) - P(u \mid x_0)$. In this way, one can quantify the interaction of the direct-indirect interaction and the spurious effect, effectively capturing a three-way interaction effect.

Theorem 2 (Total TV Decomposition) *The TV measure admits the following decomposition:*

$$TV_{x_0,x_1}(y) = x\text{-}DE_{x_0,x_1}(y \mid x_0) + x\text{-}IE_{x_0,x_1}(y \mid x_0) + x\text{-}SE_{x_0,x_1}(y) \tag{79}$$

$$+ x\text{-}DE\text{-}IE_{x_0,x_1}(y \mid x_0) + x\text{-}DE\text{-}SE_{x_0,x_1}(y) + x\text{-}IE\text{-}SE_{x_0,x_1}(y) \tag{80}$$

$$+ x\text{-}DE\text{-}IE\text{-}SE_{x_0,x_1}(y). \tag{81}$$

□

The above theorem is a crown result of the paper. The TV measure can be decomposed into first order direct, indirect, and spurious effects (Line 79), the direct-indirect, direct-spurious, indirect-spurious interaction effects (Line 80), and the three-way interaction of direct, indirect, and spurious effects (Line 81). Once again, one may use an interaction testing procedure as in Alg. 1 to obtain a more parsimonious TV representation in case some of the second or third order effects are not significantly different from 0.

Practical Implications. There are a number of practical implications resulting from the TV decomposition in Eqs. 79-81. Firstly, when the hypothesis

$$H_0^{\text{DE-IE}} : x\text{-DE-IE}_{x_0,x_1}(y \mid x_0) = 0 \tag{82}$$

is rejected, the practitioner can conclude that the mediator value $W = w$ modifies the direct effect, namely that the controlled direct effect

$$\text{CDE}_{x_0,x_1}(y_w) = E[Y_{x_1,w} - Y_{x_0,w}] \tag{83}$$

varies according to levels of $W = w$. Estimating such an effect along different $W = w$ levels may therefore be informative. Similarly, the existence of DE-SE or IE-SE interactions implies that there is heterogeneity in the controlled direct effect

$$\text{CDE}_{x_0,x_1}(y_{w,z}) = E[Y_{x_1,z,w} - Y_{x_0,z,w}], \tag{84}$$

which means that the practitioner may also wish to focus their attention on the heterogeneity of the $\text{CDE}_{x_0,x_1}(y_{w,z})$ according to levels of $Z = z, W = w$ to learn more about the phenomenon under study.

4.1 Population Granularity

In this section, we discuss population granularity, and the testing of interactions along different subpopulations of the data. We begin with a motivating example:

Example 4 (DE-IE Interaction Testing) *Consider the SCM $\mathcal{M}$ given by:*

$$Z \leftarrow \textit{Bernoulli}(0.5) \tag{85}$$
$$X \leftarrow \textit{Bernoulli}(0.5) \tag{86}$$
$$W \leftarrow 1 - X \tag{87}$$
$$Y \leftarrow (2Z - 1)XW. \tag{88}$$

If we test for the interaction of direct and indirect effects, using the measure

$$x\text{-}DE\text{-}IE_{x_0,x_1}(y \mid x) = [\mathbb{E}(y_{x_1,W_{x_0}} \mid x) - \mathbb{E}(y_{x_0,W_{x_0}} \mid x)] - [\mathbb{E}(y_{x_1,W_{x_1}} \mid x) - \mathbb{E}(y_{x_0,W_{x_1}} \mid x)] \tag{89}$$

we find that for both $x \in \{x_0, x_1\}$, we have

$$x\text{-}DE\text{-}IE_{x_0,x_1}(y \mid x) = 0. \tag{90}$$

Therefore, based on the measure x-DE-IE$_{x_0,x_1}(y \mid x)$ we cannot detect an interaction of direct and indirect effects, even though X influences W and there is a term with an interaction XW in f_y.

However, when we consider the difference $[Y_{x_1,W_{x_0}} - Y_{x_0,W_{x_0}}](u) - [Y_{x_1,W_{x_1}} - Y_{x_0,W_{x_1}}](u)$ averaged across $Z = 0$ and $Z = 1$, we find that

$$[\mathbb{E}(y_{x_1,W_{x_0}} \mid z_1) - \mathbb{E}(y_{x_0,W_{x_0}} \mid z_1)] - [\mathbb{E}(y_{x_1,W_{x_1}} \mid z_1) - \mathbb{E}(y_{x_0,W_{x_1}} \mid z_1)] = 1 \tag{91}$$
$$[\mathbb{E}(y_{x_1,W_{x_0}} \mid z_0) - \mathbb{E}(y_{x_0,W_{x_0}} \mid z_0)] - [\mathbb{E}(y_{x_1,W_{x_1}} \mid z_0) - \mathbb{E}(y_{x_0,W_{x_1}} \mid z_0)] = -1 \tag{92}$$

Therefore, an interaction that was not visible in the x-DE-IE$_{x_0,x_1}(y \mid x) = 0$ effect, becomes visible once the conditioning on $X = x$ is replaced by the conditioning on $Z = z$. □

The above example illustrates an important concept of *power*, previously discussed in (Plečko and Bareinboim, 2024). In particular, x-DE-IE$_{x_0,x_1}(y \mid x)$ measures the direct-indirect effect interaction across all units compatible with $X(u) = x$. It can be thus be expanded as:

$$x\text{-DE-IE}_{x_0,x_1}(y \mid x) = \mathbb{E}\Big([y_{x_1,W_{x_0}}(u) - y_{x_0,W_{x_0}}(u)] - [y_{x_1,W_{x_1}}(u) - y_{x_0,W_{x_1}}(u)] \mid X = x\Big), \tag{93}$$
$$= \sum_z \mathbb{E}\Big([y_{x_1,W_{x_0}}(u) - y_{x_0,W_{x_0}}(u)] - [y_{x_1,W_{x_1}}(u) - y_{x_0,W_{x_1}}(u)] \mid X = x, Z = z\Big) \tag{94}$$
$$\times P(Z = z \mid X = x).$$

In Ex. 4, we have that $Y_{x,W_{x'}} \perp\!\!\!\perp X \mid Z = z$ for any choice of x, x', known as conditional ignorability, which is a consequence of the fact that there are no back-door paths between X, Y. In Ex. 4 we also have $X \perp\!\!\!\perp Z$. Therefore, for the SCM $\mathcal{M}$ in Eqs. 85-88, we have that

$$x\text{-DE-IE}_{x_0,x_1}(y \mid x) = \sum_z \mathbb{E}\Big([y_{x_1,W_{x_0}}(u) - y_{x_0,W_{x_0}}(u)] - [y_{x_1,W_{x_1}}(u) - y_{x_0,W_{x_1}}(u)] \mid Z = z\Big) \tag{95}$$
$$\times P(Z = z)$$
$$= \sum_z z\text{-DE-IE}_{x_0,x_1}(y \mid z) P(Z = z), \tag{96}$$

where z-DE-IE$_{x_0,x_1}(y \mid z) = \mathbb{E}\Big([y_{x_1,W_{x_0}}(u) - y_{x_0,W_{x_0}}(u)] - [y_{x_1,W_{x_1}}(u) - y_{x_0,W_{x_1}}(u)] \mid Z = z\Big)$. Therefore, x-DE-IE$_{x_0,x_1}(y \mid x)$ is a mixture of the z-specific direct-indirect interaction effects, which

happen to cancel out in Ex. 4. However, when performing interaction testing, one can also directly compute the more granular, z-specific effects z-DE-IE$_{x_0,x_1}(y \mid z)$ instead of the x-specific x-DE-IE$_{x_0,x_1}(y \mid x)$. In the language of interaction testing, we can say that the test

$$H_0^{z\text{-DE-IE}} : z\text{-DE-IE}_{x_0,x_1}(y \mid z) = 0, \tag{97}$$

is a valid interaction test for the DE-IE interaction for any fixed choice of $Z = z$, and may be superior to testing x-DE-IE$_{x_0,x_1}(y \mid x)$ against 0.

Further population granularity. At the conceptual level, it is possible to consider even more granular DE-IE interactions, for instance by conditioning on x, z, w, or even x, z, w, y, e.g., based on the measures:

$$(x,z,w)\text{-DE-IE}_{x_0,x_1}(y \mid x,z,w) = \mathbb{E}\Big([y_{x_1,W_{x_0}}(u) - y_{x_0,W_{x_0}}(u)] - [y_{x_1,W_{x_1}}(u) - y_{x_0,W_{x_1}}(u)] \mid x,z,w\Big) \tag{98}$$

$$(x,z,w,y)\text{-DE-IE}_{x_0,x_1}(y \mid x,z,w,y) = \mathbb{E}\Big([y_{x_1,W_{x_0}}(u) - y_{x_0,W_{x_0}}(u)] - [y_{x_1,W_{x_1}}(u) - y_{x_0,W_{x_1}}(u)] \mid x,z,w,y\Big) \tag{99}$$

These measures quantify the DE-IE interaction for the set of units compatible with $X = x, Z = z, W = w$, and $X = x, Z = z, W = w, Y = y$, respectively. This provides even more granular quantification of the interaction effects, for an increasingly smaller population. It is even possible to consider interactions at the unit-level

$$u\text{-DE-IE}_{x_0,x_1}(y \mid u) = [y_{x_1,W_{x_0}}(u) - y_{x_0,W_{x_0}}(u)] - [y_{x_1,W_{x_1}}(u) - y_{x_0,W_{x_1}}(u)], \tag{100}$$

which quantifies the effect for a single unit u, and can thus be seen as the most granular possible quantification of the DE-IE interaction. However, we remark that the measures in Eqs. 98-100 are not identifiable from observational data even in the causal diagram in Fig. 2, since they would require the evaluation of the joint distribution of counterfactual outcomes, which cannot be done without much stronger (and often untestable) assumptions.

5. Risk Ratio and Odds Ratio Scales

The discussion so far was concerned with the existence of interaction on a difference scale. Oftentimes, especially when dealing with binary outcomes, different scales may be interesting, such as the risk ratio (RR) scale or the odds ratio (OR) scale. In this section, we describe how the results of the manuscript can be extended to different scales, with a particular focus on binary outcomes.

The key idea in case of binary outcomes will be to replace the structural mechanism f_y with its probability counterpart, defined by:

$$p_y(x,z,w) := \mathbb{E}_{u_y}[Y_{x,z,w}] = P_{u_y}(Y_{x,z,w} = 1), \tag{101}$$

The structural mechanism f_y returns a binary value of Y, while p_y returns a probability of Y belonging to class 1 given covariates $X = x, Z = z, W = w$. When considering another scale apart from the difference scale, we may consider suitable transformations of the "mechanism" $p_y(x,z,w)$. In fact, when the data corresponds to the causal diagram in Fig. 2, the outcome Y can be thought of as being simply replaced by $P = \mathbb{E}_{u_y}[Y_{x,z,w}]$, and the same causal diagram is still valid with P replacing the Y outcome. For the risk ratio scale, we consider the transformation

$$\log p_y(x,z,w), \tag{102}$$

whereas for the odds ratio scale we consider the transformation

$$\text{logit } p_y(x,z,w), \tag{103}$$

where $\text{logit}(a) = \log \frac{a}{1-a}$. Analogously to the notion of no interaction on a difference scale in Def. 11, we can consider no interaction criteria on RR and OR scales, which can be defined bas on the structural interaction criteria on the difference scale:

Definition 14 (Structural Interaction Criteria for RR and OR Scales) *Consider the causal diagram in Fig. 1b, and let $p_y(x,z)$ be the probability $P(Y=1 \mid x,z)$. We say that*

(a) There is no structural interaction of total and spurious effects on the risk ratio scale, written Str-TRR-SRR = 0, if the mechanism $\log p_y(x,z)$ satisfies Str-TE-SE = 0. Explicitly, Str-TRR-SRR = 0 if either we can write:

$$p_y(x,z) = \exp\{f_y^{(1)}(x) + f_y^{(2)}(z)\}, \tag{104}$$

or if there is no back-door path between X, Y.

(b) There is no total-spurious interaction on the odds ratio scale, written Str-TOR-SOR = 0, if the mechanism $\text{logit}\, p_y(x,z)$ satisfies Str-TE-SE = 0. Explicitly, Str-TOR-SOR = 0 if either we can write

$$p_y(x,z) = \text{expit}\{f_y^{(1)}(x) + f_y^{(2)}(z)\}, \tag{105}$$

or if there is no back-door path between X, Y.

Next, consider the causal diagram in Fig. 2, and let $p_y(x,z,w)$ be the probability $P(Y=1 \mid x,z,w)$. We say that

(c) There is no direct-indirect structural interaction for the risk ratio scale, written Str-DRR-IRR = 0, if $\log p_y(x,z,w)$ satisfies Str-DE-IE. Explicitly, Str-DRR-IRR = 0 if either we can write

$$p_y(x,z,w) = \exp\{f_y^{(1)}(x,z) + f_y^{(2)}(z,w)\}, \tag{106}$$

or if there is no indirect path $X \to W \to Y$.

(d) There is no direct-indirect structural interaction for the odds ratio scale, written Str-DOR-IOR = 0, if $\text{logit}\, p_y(x,z,w)$ satisfies Str-DE-IE. Explicitly, Str-DOR-IOR = 0 if either we can write

$$p_y(x,z,w) = \text{expit}\{f_y^{(1)}(x,z) + f_y^{(2)}(z,w)\}, \tag{107}$$

or if there is no indirect path $X \to W \to Y$.

If interactions exist, we say that the corresponding structural criterion is equal to 1. □

The no interaction notions, on a structural level, are very similar (analogous) for the RR and OR scales can be written using the structural interaction notions on the difference scale. The key subtlety is that we need to consider appropriate transformations of the probability of a positive outcome, $\log P(y \mid x,z,w)$ for the RR scale, and $\text{logit}\, P(y \mid x,z,w)$ for the OR scale. We remark that the more granular interactions (DE-SE, SE-IE, and DE-IE-SE) can also be written for OR and RR scales, but we omit these definitions in the interest of brevity.

Naturally, the transformations that are used also have implications for how we can test for interactions, and how the TV decompositions are affected. For the remainder of this section, we describe the approach for the RR scale, but the analogous results can be obtained for the OR scale with only slight modifications.

We first introduce the risk ratio measures of direct, indirect, and spurious effects, and the relevant interactions:

Definition 15 (x-specific TRR, DRR, IRR, and SRR) *The x-{total, direct, indirect, spurious} risk ratios are defined as follows:*

$$x\text{-}TRR_{x_0,x_1}(y \mid x) = \frac{P(y_{x_1} \mid x)}{P(y_{x_0} \mid x)} \tag{108}$$

$$x\text{-}DRR_{x_0,x_1}(y \mid x) = \frac{P(y_{x_1,W_{x_0}} \mid x)}{P(y_{x_0} \mid x)} \tag{109}$$

$$x\text{-}IRR_{x_0,x_1}(y \mid x) = \frac{P(y_{x_0,W_{x_1}} \mid x)}{P(y_{x_0} \mid x)} \tag{110}$$

$$x\text{-}SRR_{x_0,x_1}(y) = \frac{P(y_{x_0} \mid x_1)}{P(y_{x_0} \mid x_0)}. \tag{111}$$

The x-specific TRR-SRR and DRR-IRR interaction measures are defined as

$$x\text{-}DRR\text{-}IRR_{x_0,x_1}(y \mid x) = \frac{P(y_{x_1} \mid x)}{P(y_{x_0,W_{x_1}} \mid x)} \times \frac{P(y_{x_0} \mid x)}{P(y_{x_1,W_{x_0}} \mid x)} \tag{112}$$

$$x\text{-}TRR\text{-}SRR_{x_0,x_1}(y) = \frac{P(y_{x_1} \mid x_1)}{P(y_{x_0} \mid x_1)} \times \frac{P(y_{x_0} \mid x_0)}{P(y_{x_1} \mid x_0)}. \tag{113}$$

□

The interpretation of interaction measures at the RR scale can be understood as follows, again in analogy with the interpretation for the difference scale. The x-DRR-IRR$_{x_0,x_1}(y \mid x)$ looks at the increase in risk from a $x_0 \to x_1$ transition along the direct path while having W_{x_1} for the group of units $X(u) = x$, divided by the increase in risk of the same transition while having W_{x_0}. In other words, the quantity is trying to compute how much changing $W_{x_0} \to W_{x_1}$ modifies the increase in the risk associated with a direct $x_0 \to x_1$ transition. Similarly, the x-TRR-SRR$_{x_0,x_1}(y)$ quantifies how much changing $X = x_0$ to $X = x_1$ along the spurious path modifies the change in risk of a $x_0 \to x_1$ transition along the causal paths.

Armed with the above understanding of RR scale measure, we can investigate the analogue of the TV decomposition, offering the formal result that mirrors Thm. 1:

Theorem 3 (RR Total Variation Decomposition) *Let the complete risk ratio $CRR_{x_0,x_1}(y)$ be defined as $\frac{P(y|x_1)}{P(y|x_0)}$. The $CRR_{x_0,x_1}(y)$ can be decomposed as follows:*

$$CRR_{x_0,x_1}(y) = x\text{-}TRR_{x_0,x_1}(y \mid x_0) \times x\text{-}SRR_{x_0,x_1}(y) \times x\text{-}TRR\text{-}SRR_{x_0,x_1}(y). \tag{114}$$

The CRR measure can also be decomposed as:

$$\begin{aligned} CRR_{x_0,x_1}(y) = {} & x\text{-}DRR_{x_0,x_1}(y \mid x_0) \times x\text{-}IRR_{x_0,x_1}(y \mid x_0) \times x\text{-}DRR\text{-}IRR_{x_0,x_1}(y \mid x_0) \\ & \times x\text{-}SRR_{x_0,x_1}(y) \times x\text{-}TRR\text{-}SRR_{x_0,x_1}(y). \end{aligned} \tag{115}$$

□

Once again, the TRR-SRR and DRR-IRR interaction measure can be related to the structural level, through the following admissibility result:

Proposition 3 (Structural Admissibility of RR Interaction Tests)

$$Str\text{-}TRR\text{-}SRR = 0 \implies x\text{-}TRR\text{-}SRR_{x_0,x_1}(y) = 1, \tag{116}$$

$$Str\text{-}DRR\text{-}IRR = 0 \implies x\text{-}DRR\text{-}IRR_{x_0,x_1}(y \mid x) = 1. \tag{117}$$

□

In words, whenever we can write the quantity $\log p_y(x, z)$ without a functional term $f(x, z)$ taking both x, z as inputs, then the interaction x-TRR-SRR$_{x_0,x_1}(y)$ must equal 1. Similarly, whenever we can write $\log p_y(x, z, w)$ without a functional term $f(x, w)$ taking both x, w as inputs, then the interaction x-DRR-IRR$_{x_0,x_1}(y \mid x)$ must equal 1. The admissibility result from Prop. 3 again leads to an interaction testing procedure, shown in Alg. 2. For instance, the hypothesis $H_0^{\text{TRR-SRR}}$ tests

Algorithm 2 Interaction Testing for CRR Decomposition

- **Inputs:** Causal Diagram $\mathcal{G}$, Observational Data $\mathcal{D}$

  Compute the estimate of the total-spurious interaction risk ratio x-TRR-SRR$_{x_0,x_1}(y)$, and its 95% confidence interval. Test the hypothesis

$$H_0^{\text{TRR-SRR}} : x\text{-TRR-SRR}_{x_0,x_1}(y) = 1. \tag{118}$$

  Compute the estimate of the direct-indirect interaction risk ratio x-DRR-IRR$_{x_0,x_1}(y \mid x_0)$, and its 95% confidence interval. Test the hypothesis

$$H_0^{\text{DRR-IRR}} : x\text{-DRR-IRR}_{x_0,x_1}(y \mid x_0) = 1. \tag{119}$$

  - if neither $H_0^{\text{TRR-SRR}}, H_0^{\text{DRR-IRR}}$ are rejected, return the decomposition

$$\text{CRR}_{x_0,x_1}(y) = x\text{-DRR}_{x_0,x_1}(y \mid x_0) \times x\text{-IRR}_{x_0,x_1}(y \mid x_0) \times x\text{-SRR}_{x_0,x_1}(y) \tag{120}$$

  - if only $H_0^{\text{DRR-IRR}}$ is rejected, return the decomposition

$$\begin{aligned}\text{CRR}_{x_0,x_1}(y) = x\text{-DRR}_{x_0,x_1}(y \mid x_0) \times x\text{-IRR}_{x_0,x_1}(y \mid x_0) \times x\text{-DRR-IRR}_{x_0,x_1}(y \mid x_0) \\ \times x\text{-SRR}_{x_0,x_1}(y).\end{aligned} \tag{121}$$

  - if only $H_0^{\text{TRR-SRR}}$ is rejected, return the decomposition

$$\begin{aligned}\text{CRR}_{x_0,x_1}(y) = x\text{-DRR}_{x_0,x_1}(y \mid x_0) \times x\text{-IRR}_{x_0,x_1}(y \mid x_0) \\ \times x\text{-SRR}_{x_0,x_1}(y) \times x\text{-TRR-SRR}_{x_0,x_1}(y).\end{aligned} \tag{122}$$

  - if both $H_0^{\text{TRR-SRR}}, H_0^{\text{DRR-IRR}}$ are rejected, return the CRR decomposition in Eq. 115.
- **Output:** CRR decomposition with parsimony.

whether

$$\frac{P(y_{x_1} \mid x_1)}{P(y_{x_0} \mid x_1)} = \frac{P(y_{x_1} \mid x_0)}{P(y_{x_0} \mid x_0)}, \tag{123}$$

or in words, if the change in risk ratio of a causal $x_0 \to x_1$ transition is equal for the $X(u) = x_1$ and $X(u) = x_0$ groups. The test can, by symmetry, also be understood as testing if

$$\frac{P(y_{x_1} \mid x_1)}{P(y_{x_1} \mid x_0)} = \frac{P(y_{x_0} \mid x_1)}{P(y_{x_0} \mid x_0)}. \tag{124}$$

In words, the test is comparing the change in risk ratio of a spurious $x_0 \to x_1$ transition while having $X = x_1$ along causal paths vs. the change in risk of the same transition while having $X = x_0$ along causal paths. Naturally, the $H_0^{\text{DRR-IRR}}$ hypothesis is amenable to similar interpretations.

Finally, we discuss what happens when looking at more granular effects, such as the interactions of spurious/direct or spurious/indirect risk ratios. We begin by defining the granular notions of interaction in this context:

Definition 16 (Granular Risk Ratio Interactions) *Consider the following effect interactions:*

$$x\text{-}DRR\text{-}SRR_{x_0,x_1}(y) = \frac{P(y_{x_1,W_{x_0}} \mid x_1)}{P(y_{x_0} \mid x_1)} \times \frac{P(y_{x_0} \mid x_0)}{P(y_{x_1,W_{x_0}} \mid x_0)} \tag{125}$$

$$x\text{-}IRR\text{-}SRR_{x_0,x_1}(y) = \frac{P(y_{x_0,W_{x_1}} \mid x_1)}{P(y_{x_0} \mid x_1)} \times \frac{P(y_{x_0} \mid x_0)}{P(y_{x_0,W_{x_1}} \mid x_0)} \tag{126}$$

$$x\text{-}DRR\text{-}IRR\text{-}SRR_{x_0,x_1}(y) = \frac{P(y_{x_1} \mid x_1)}{P(y_{x_0,W_{x_1}} \mid x_1)} \times \frac{P(y_{x_0} \mid x_1)}{P(y_{x_1,W_{x_0}} \mid x_1)} \tag{127}$$

$$\times \left(\frac{P(y_{x_1} \mid x_0)}{P(y_{x_0,W_{x_1}} \mid x_0)} \times \frac{P(y_{x_0} \mid x_0)}{P(y_{x_1,W_{x_0}} \mid x_0)} \right)^{-1}. \tag{128}$$

□

Using the more granular notion of interaction at risk ratio scale, we can prove the most granular decomposition of the CRR measure:

Theorem 4 (Total CRR Decomposition) *The $CRR_{x_0,x_1}(y)$ can be decomposed as follows:*

$$CRR_{x_0,x_1}(y) = x\text{-}DRR_{x_0,x_1}(y \mid x_0) \times x\text{-}IRR_{x_0,x_1}(y \mid x_0) \times x\text{-}SRR_{x_0,x_1}(y) \tag{129}$$

$$\times\, x\text{-}DRR\text{-}IRR_{x_0,x_1}(y \mid x_0) \times x\text{-}DRR\text{-}SRR_{x_0,x_1}(y) \times x\text{-}IRR\text{-}SRR_{x_0,x_1}(y) \tag{130}$$

$$\times\, x\text{-}DRR\text{-}IRR\text{-}SRR_{x_0,x_1}(y). \tag{131}$$

□

The decomposition of the CRR appearing in the above theorem is the risk ratio analogue of the decomposition of the TV measure from Thm. 2.

6. Experiments

In this section, we perform an extensive evaluation of the proposed method. The goal of the evaluation is to understand how often we detect different types of interactions. The experimental part is intended to answer the question of "how often do causal pathways interact?". For answering this question, we use a selection of 10 datasets including data in criminal justice, economics, marketing, banking, medical treatments, epidemiology, and public health. For each dataset, we first construct the graphical model shown in Fig. 2. This is done by selecting the X, Z, W, and Y variable sets. There are 5 types of interactions we consider. First, we consider the interactions of spurious and total effects, and direct and indirect effects, appearing in the decompositions in Thm. 1 and Thm. 3. Recall that these interactions were closely related to the existence of different terms in the structural mechanism f_y (Prop. 2, Prop. 3). Furthermore, we also consider the more granular interactions, including spurious-direct, spurious-indirect, and direct-indirect-spurious. For each interaction, we use a corresponding interaction test to test for the existence of the interaction at: (i) difference scale; (ii) risk ratio scale; (iii) odds ratio scale (only for binary outcomes Y). Below we provide a short description of the datasets for experiments, including the description of how variables X, Z, W, and Y were chosen, while in Appendix A we provide further details including the full list of variables used:

(i) **COMPAS:** (Larson et al., 2016) The COMPAS dataset contains information used for predicting recidivism risk among individuals seeking parole. Variables include demographic information (Z), recidivism outcome (Y), prior criminal history (W) and race (Z). The focus is on understanding the association of race and the recidivism risk.

(ii) **Census 2018:** (Cevid et al., 2022; Plečko and Bareinboim, 2024) The dataset contains detailed demographic information from the 2018 US Census. Variables include sex (X), income level in dollars per year (Y), employment status and educational background (W), and other demographic variables (Z). The focus is on understanding the association of sex and income level.

(iii) **MIMIC-IV:** (Johnson et al., 2023) The MIMIC-IV dataset contains electronic health records from a large tertiary hospital center in Boston, Massachusetts. We are interested in patients admitted to the intensive care unit (ICU). Variables include patients' race (X), mortality outcome (Y), comorbidities and demographic information (Z), and illness severity and treatment information (W). The focus is on understanding the association of race and mortality in intensive care unit outcomes.

(iv) **HELOC:** (Fair Isaac Corporation (FICO), 2016) The Home Equity Line of Credit (HELOC) dataset is provided by the Fair Isaac Corporation (FICO) and includes credit applications made by homeowners. Variables include proportion of transaction delinquencies (X), credit risk performance (Y), financial background and debt history (Z), and recent credit utilization (W). The focus is on understanding the association of high number of transactions delinquencies and the estimate of the credit risk performance.

(v) **UCI Credit:** (Yeh, 2016) The UCI Credit dataset contains information about default of credit card payments of bank customers in Taiwan. Variables include applicant sex (X), default on payment (Y), age (Z), and education, marital status, and recent financial behavior (W). The focus is on understanding the association of customer gender and default on payment.

(vi) **UCI Adult:** (Becker and Kohavi, 1996) This dataset includes census data from the United States for predicting whether the individual's income exceeds \$50K/yr. Variables include education level (X), income level (Y), demographic factors (Z), and family/employment characteristics (W). The focus is on understanding the association of high levels of education and and income level.

(vii) **UCI Wine Quality:** (Cortez et al., 2009) This dataset consists of physicochemical tests of wine samples and their quality ratings. Variables include alcohol content (X), wine quality (Y), chemical composition (Z), and physical/fermentation properties (W). The focus is on understanding the association of high alcohol content and wine quality.

(viii) **UCI Bank Marketing:** (Moro et al., 2012) This dataset includes marketing campaign data from a Portuguese banking institution. Variables include housing loan status (X), agreement for a term deposit subscription (Y), demographic information (Z), and campaign details (W). The focus is on understanding the association of housing loan status and term deposit subscription.

(ix) **UCI Estimation of Obesity:** (Palechor and De la Hoz Manotas, 2019) This dataset contains information on Colombian individuals' eating habits and physical attributes to determine obesity levels. Variables include family history of overweight (X), body mass index (BMI, labeled Y), demographic factors (Z), and lifestyle habits (W). The focus is on understanding the association of family history of overweight and BMI.

(x) **BRFSS Diabetes Health Indicators:** This dataset was collected by Behavioral Risk Factor Surveillance System (BRFFSS), and contains health-related survey data to assess diabetes risk factors. Variables include physical activity (X), diabetes status (Y), demographic and lifestyle factors (Z), and health conditions (W). The focus is on understanding the association of physical activity and diabetes status.

Dataset / Interaction	SE ⊗ TE	DE ⊗ IE	SE ⊗ DE	SE ⊗ IE	DE ⊗ IE ⊗ SE
COMPAS	•	•			•
Census 2018	•	•	•	•	•
UCI Credit			•	•	•
MIMIC-IV	•			•	•
HELOC	•	•	•	•	•
UCI Adult	•		•	•	•
UCI Wine Quality	•	•	•		•
UCI Bank Marketing	•	•	•		•
UCI Obesity	•		•	•	•
BRFSS Diabetes	•				

Table 2: Interaction testing experimental results.

The summary of the experimental results is shown in Tab. 2. Overall, we found that 39/50 of the investigated interactions seemed to be significant. We note that the proportion of about 80% is quite high, and we remark that the interaction tests performed here are not the most powerful possible (i.e., more granular tests could be used, as discussed in Sec. 4.1). Therefore, our findings demonstrate that interactions frequently need to be assessed in various research domains.

7. Conclusion

In this paper, we argued that a new concept of *variation analysis* should be adopted, to reflect recent methodological advances in which confounded/spurious effects are considered when analying co-variations between a treatment X and an outcome Y. Traditionally, in mediation analysis, only the causal variations (direct, indirect) originating from X and entering Y are considered, and the quantity that is under study is often the total effect (TE, which encompasses direct and indirect effects). In variation analysis, however, the focus is on the total variation measure (TV, which encompasses all co-variations between X, Y), thereby representing a more general approach than mediation analysis. We further introduced the concept of structural interactions (Defs. 11, 12), and discussed how different causal pathways (direct, indirect, and spurious) between treatment X and outcome Y can interact, and such interactions can be quantified (Def. 9). We then proved a first decomposition of the TV measure that includes and quantifies all the different interactions among causal paths (Thms. 1, 2), and requires only $x_0 \to x_1$ transitions in all the effects, allowing for easier interpretation of the decomposition. We then defined the concept of interaction testing, which is a hypothesis test aimed at testing the existence of interaction between different causal paths (Def. 10). Subsequently, we demonstrated that whenever a hypothesis is not rejected, it implies that the decomposition of the TV measure can be made more parsimonious, by omitting some of the interaction terms (Alg. 1). Once again, this provides a practical approach that allows the user to more easily interpret the decomposition of the TV measure. Further, we extended the above results to include the risk ratio and odds ratio scales (Sec. 5). Finally, in Sec. 6, we performed an extensive analysis of 10 well-known and commonly used datasets – in an attempt to discover how often there are significant interactions among causal pathways. Over 10 datasets and 5 types of interactions, we found that 39 out of 50 possible interactions were significant – indicating that, in practice, interactions of causal pathways are of major importance when analyzing co-variations between a treatment X and an outcome Y.

References

Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, page 507–556. Association for Computing Machinery, New York, NY, USA, 1st edition, 2022.

Reuben M Baron and David A Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6):1173, 1986.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

Domagoj Cevid, Loris Michel, Jeffrey Näf, Peter Bühlmann, and Nicolai Meinshausen. Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. *Journal of Machine Learning Research*, 23(333):1–79, 2022.

Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision support systems*, 47(4): 547–553, 2009.

Fair Isaac Corporation (FICO). Home equity line of credit, 2016. URL `https://community.fico.com/s/explainable-machine-learning-challenge`.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Benjamin Moody, Brian Gow, Li-wei H Lehman, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9, 2016.

S. Moro, P. Rita, and P. Cortez. Bank Marketing. UCI Machine Learning Repository, 2012. DOI: https://doi.org/10.24432/C5K306.

Fabio Mendoza Palechor and Alexis De la Hoz Manotas. Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from colombia, peru and mexico. *Data in brief*, 25:104344, 2019.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000. 2nd edition, 2009.

Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, page 411–420, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

Drago Plečko and Elias Bareinboim. Causal fairness analysis: a causal toolkit for fair machine learning. *Foundations and Trends® in Machine Learning*, 17(3):304–589, 2024.

James M Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, pages 143–155, 1992.

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

Tyler VanderWeele. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press, 2015.

I-Cheng Yeh. Default of Credit Card Clients. UCI Machine Learning Repository, 2016. DOI: https://doi.org/10.24432/C55S3H.

Junzhe Zhang and Elias Bareinboim. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.