Non-Parametric Path Analysis in Structural Causal Models

Junzhe Zhang and Elias Bareinboim
Purdue University, USA
{zhang745, eb}@purdue.edu

Abstract

One of the fundamental tasks in causal inference is to decompose the observed association between a decision \( X \) and an outcome \( Y \) into its most basic structural mechanisms. In this paper, we introduce counterfactual measures for effects along with a specific mechanism, represented as a path from \( X \) to \( Y \) in an arbitrary structural causal model. We derive a novel non-parametric decomposition formula that expresses the covariance of \( X \) and \( Y \) as a sum over unblocked paths from \( X \) to \( Y \) contained in an arbitrary causal model. This formula allows a fine-grained path analysis without requiring a commitment to any particular parametric form, and can be seen as a generalization of Wright’s decomposition method in linear systems (1923,1932) and Pearl’s non-parametric mediation formula (2001).

1 INTRODUCTION

Analyzing the relative strength of different pathways between a decision \( X \) and an outcome \( Y \) is a topic that has interested both scientists and practitioners across disciplines for many decades. Specifically, path analysis allows scientists to explain how Nature’s “black-box” works, and practically, it enables decision analysts to predict how an environment will change under a variety of policies and interventional conditions [Wright, 1923; Baron and Kenny, 1986; Bollen, 1989; Pearl, 2001].

More recently, understanding using causal inference tools how a black-box decision-making system operates has been a target of growing interest in the Artificial Intelligence community, most prominently in the context of Explainability, Transparency, and Fairness [Lu Zhang, 2017; Kusner et al., 2017; Zafar et al., 2017; Kilbertus et al., 2017; Zhang and Bareinboim, 2018a]. For example, consider the standard fairness model described in Fig. 1(a) that is concerned with the relation between a hiring decision \( (Y) \) and an applicant’s religious beliefs \((X)\), which are mediated by the location \((W)\), and confounded by the education background \((Z)\) of the applicant. \(^1\) Directed edges represent functional relations between variables. The relationship between \( X \) and \( Y \) is materialized through four different pathways in the system — the direct path \( l_1 : X \rightarrow Y \), the indirect path \( l_2 : X \rightarrow W \rightarrow Y \), and the spurious paths \( l_3 : X \leftarrow Z \rightarrow Y \) and \( l_4 : X \leftarrow Z \leftarrow W \rightarrow Y \).

Assuming, for simplicity’s sake, that the functional relationships are linear and \( U_i \) is an independent “error” associated with each variable \( V_i \) (called the linear-standard model), Fig. 1(a) shows the structural coefficients corresponding to each edge — i.e., the value of the variable \( Y \) is decided by the structural function \( Y \leftarrow \alpha_{YX}X + \alpha_{YZ}Z + \alpha_{YW}W + U_f \). The celebrated result known as Wright’s method of path coefficients [Wright, 1923, 1934], also known as Wright’s rule, allows one to express the covariance of \( X \) and \( Y \), denoted by \( \text{Cov}(X,Y) \), as the sum of the products of the structural coefficients along the paths from \( X \) to \( Y \) in the underlying causal model. \(^2\) In particular, \( \text{Cov}(X,Y) \) is equal to:

\[
\text{Cov}(X,Y) = \alpha_{YX} + \alpha_{YX}\alpha_{YW} + \alpha_{YZ}\alpha_{YW} + \alpha_{YZ}\alpha_{YW}\alpha_{YW}.
\]

Using the observational covariance matrix, the decomposition above allows one to answer some compelling questions about the relationship between \( X \) and \( Y \) in the underlying model. For instance, the product \( \alpha_{YX}\alpha_{YW} \) explains how much the indirect discrimination through the location (the path \( l_2 \)) accounts for the observed disparities in the religion composition among hired employees.

The path analysis method gained momentum in the so-

\(^1\) This specific setting has been called standard fairness model given its generality to representing a variety of decision-making scenarios [Zhang and Bareinboim, 2018a].

\(^2\) For a survey on linear methods, see [Pearl, 2000, Ch. 5].
special sciences during 1960’s, becoming extremely popular in the form of the mediation formula in which the total effect of $X$ on $Y$ is decomposed into direct and indirect components [Baron and Kenny, 1986; Bollen, 1989; Duncan, 1975; Fox, 1980]. The bulk of this literature, however, required a commitment to a particular parametric form, thus falling short of providing a general method for analyzing natural and social phenomena with nonlinearities and interactions [MacKinnon, 2008].

It took a few decades until this problem could be tackled in higher generality. In particular, the advent of non-parametric structural causal models (SCMs) allowed this leap, and a more fine-grained path-analysis with a much broader scope, including models with nonlinearities and arbitrarily complex interactions [Pearl, 2000, Ch. 7]. In particular, Pearl introduced the causal mediation formula for arbitrary non-parametric models, which decomposes the total effect $TE_{x_0,x_1}(Y) = E[Y_{x_1} | x_0] - E[Y_{x_0}]$ into what is now known as the natural direct (NDE) and indirect (NIE) effects [Pearl, 2001] (see also [Imai et al., 2010, 2011; VanderWeele, 2015]).

In the case of the specific linear-standard causal model,

$$TE_{0,1}(Y) = \alpha_{y|x} + \alpha_{w|x}\alpha_{y|x}$$

for $x_0 = 0$ and $x_1 = 1$ levels. Remarkably, when compared with Eq. 1, NDE and NIE capture the effects along with the direct and indirect paths, but omits the spurious (non-causal) paths between $X$ and $Y$ (in this case, $I_3, I_4$). The mediation formula was recently generalized to account for these spurious paths (more akin to Wright’s rules), which appears under the rubric of the causal explanation formula [Zhang and Bareinboim, 2018a]. This formula decomposes the total variation $TV_{x_0,x_1}(Y) = E[Y|x_1] - E[Y|x_0]$ (difference in conditional distributions) into counterfactual measures of the direct (Ctf-DE), indirect (Ctf-IE), and spurious (Ctf-SE) effects. In the linear-standard model, for $x_0 = 0, x_1 = 1$,

$$TV_{0,1}(Y) = \alpha_{y|x} + \alpha_{w|x}\alpha_{y|x} + \alpha_{xz|x}\alpha_{y|z} + \alpha_{xw|w}\alpha_{yz}$$

Despite the generality of such results, there are still outstanding challenges when performing path analysis in non-parametric models, i.e.: (1) Estimands are defined relative to specific values assigned to the treatment $x_1$ and its baseline $x_0$, which may be difficult to select in some non-linear settings; (2) Mediators and confounders are collapsed and considered en bloc, leading to a coarse decomposition of the relationship between $X$ and $Y$ [Pearl, 2001; Vansteelandt and VanderWeele, 2012; Tchetgen and Shpitser, 2012; VanderWeele et al., 2014; Daniel et al., 2015; Zhang and Bareinboim, 2018a]; (3) Path-specific estimands are well-defined [Pearl, 2001; Avin et al., 2005], but not in a way that they sum up to either the total effect (TE) or variation (TV), precluding the comparison of their relative strengths.

This paper aims to circumvent these problems. In particular, we decompose the covariance of a treatment $X$ and an outcome $Y$ over effects along different mechanisms between $X$ and $Y$. We define a set of novel counterfactual estimands for measuring the relative strength of a specific mechanism represented as a path from $X$ to $Y$ in an arbitrary causal model. These estimands lead to a non-parametric decomposition formula, which expresses the covariance $\text{Cov}(X, Y)$ as a sum of the unblocked paths from $X$ to $Y$ in the causal graph. This formula allows a more fine-grained analysis of the total observed variations of $Y$ due to $X$ (both through causal and spurious mechanisms) when compared to the state-of-art methods. More specifically, our contributions are: (1) counterfactual covariance measures for a specific pathway from $X$ to $Y$ (causal and spurious) in an arbitrary causal model (Defs. 8, 11-12); (2) non-parametric decomposition formulae of the covariance $\text{Cov}(X, Y)$ over paths from $X$ to $Y$ in the causal model (Thm. 5); (3) identification conditions for estimating the proposed path-specific decomposition from the passively-collected data in the standard model (Thms. 6-7).

2 PRELIMINARIES

In this section, we introduce notations used throughout the paper. We will use capital letters to denote variables (e.g., $X$), and small letters for their values ($x$). The abbreviation $P(x)$ represents the probabilities $P(X = x)$. For arbitrary sets $A$ and $B$, let $A - B$ denote their differ-

![Causal diagrams](image-url)
ence, and let $|A|$ be the dimension of set $A$. $V_{i,j}$ stands for a set $\{V_{i}, \ldots, V_{j}\}$ (0 if $i > j$). We use graphical family abbreviations: $An(X)_{G}, De(X)_{G}, Non-De(X)_{G}, Pa(X)_{G}, Ch(X)_{G}$, which stand for the set of ancestors, descendants, non-descendants, parents and children of $X$ in $G$. We omit the subscript $G$ when obvious.

The basic semantical framework of our analysis rests on structural causal models (SCM) [Pearl, 2000, Ch. 7; Bareinboim and Pearl, 2016]. A SCM $M$ consists of a set of endogenous variables $V$ (often observed) and exogenous variables $U$ (often unobserved). The values of each $V_{i} \in V$ are determined by a structural function $f_{i}$ taking as argument a combination of the other endogenous and exogenous variables (i.e., $V_{i} \leftarrow f_{i}(PA_{i}, U_{i}), PA_{i} \subseteq V, U_{i} \subseteq U$). Values of $U$ are drawn from a distribution $P(u)$. A SCM $M$ is called Markovian when the exogenous are mutually independent and each $U_{i} \in U$ is associated with only one endogenous $V_{i} \in V$. If $U_{i}$ is associated with two or more endogenous variables, $M$ is called semi-Markovian.

Each recursive SCM $M$ has an associated causal diagram in the form of a directed acyclic graph (DAG) $G$, where nodes represent endogenous variables and directed edges represent functional relations (e.g., Figs. 1-2). By convention, the exogenous $U$ are not explicitly shown in the graph; a dashed-bidirected arrow between $V_{i}$ and $V_{j}$ indicates the presence of an unobserved confounder (UC) $U_{k}$ affecting both $V_{i}$ and $V_{j}$ (e.g., the path $V_{i} \leftarrow U_{k} \rightarrow V_{j}$).

A path from $X$ to $Y$ is a sequence of edges which does not include a particular node more than once. It may go either along or against the direction of the edges. Paths of the form $X \rightarrow \cdots \rightarrow Y$ are causal (from $X$ to $Y$).

We use d-separation and blocking interchangeably, following the convention in [Pearl, 2000]. Any unblocked path that is not causal is called spurious. The direct link $X \rightarrow Y$ is the direct path and all the other causal paths from $X$ to $Y$ are called indirect. The set of unblocked paths from $X$ to $Y$ given a set $Z$ in a causal diagram $G$ is denoted by $\Pi(X, Y|Z)_{G}$; causal, indirect, and spurious paths are denoted by $\Pi^{c}(X, Y|Z)_{G}$, $\Pi^{i}(X, Y|Z)_{G}$, and $\Pi^{s}(X, Y|Z)_{G}$ ($G$ will be omitted when obvious).

For a causal path $g$ including nodes $V_{1}$, $V_{2}$, we denote $g(V_{1}, V_{2})$ a subpath of $g$ from $V_{1}$ to $V_{2}$.

An intervention on a set of endogenous variables $X$ and exogenous variables $U_{i}$, denoted by $do(x^{*}, u_{i}^{*})$, is an operation where values of $X, U_{i}$ are set to $x^{*}$, $u_{i}^{*}$, respectively, without regard to how they were ordinarily determined ($X$ through $f_{X}$ and $U_{i}$ through $P(U_{i})$). Formally, we can rewrite the definition of potential response [Pearl, 2000, Ch. 7.1] to account for operation on $U_{i}$, namely:

**Definition 1 (Potential Response).** Let $M$ be a SCM, $X, Y$ sets of arbitrary variables in $V$, and $U_{i}$ a set of arbitrary variables in $U$. Let $U_{-i} = U - U_{i}$. The potential response of $Y$ to the intervention $do(x^{*}, u_{i}^{*})$ in the situation $U = u_{i}$, denoted by $Y_{x^{*}, u_{i}^{*}}(u_{i})$, is the solution for $Y$ with $U_{-i} = u_{-i}, U_{i} = u_{i}^{*}$ in the modified submodel $M_{x^{*}}$ where functions $f_{X}$ are replaced by constant functions $X = x^{*}$, i.e., $Y_{x^{*}, u_{i}^{*}}(u_{i}) = Y_{M_{x^{*}}}(u_{i}^{*}, u_{-i}).$

$Y_{x^{*}, u_{i}^{*}}(u_{i})$ can be read as the counterfactual sentence “the value that $Y$ would have obtained in situation $U_{-i} = u_{-i}$, had the treatment $X$ been $x^{*}$ and the situation $U_{i}$ been $u_{i}^{*}$. Averaging $u$ over the distribution $P(u)$, we obtain a counterfactual random variable $Y_{x^{*}, u_{i}^{*}}$. If the values of $x^{*}, u_{i}^{*}$ follow random variables $X^{*}, U_{i}^{*}$, we denote the resulting counterfactual $Y_{X^{*}, U_{i}^{*}}$.

**3 A COARSE COVARIANCE DECOMPOSITION**

In this section, we introduce counterfactual measures that will allow us to non-parametrically decompose the covariance $\text{Cov}(X, Y)$ in terms of direct, indirect and spurious pathways from $X$ to $Y$. Given space constraints, all proofs are included in Appendix 1.

If there exists no spurious path from $X$ to $Y$, then treatment $X$ is independent of the counterfactual $Y_{x^{*}}$, i.e., $(X \perp Y_{x^{*}})_{G}$ [Pearl, 2000, Ch. 11.3.2]. The **spurious covariance** can then be defined as the correlation between the factual variable $X$ and counterfactual $Y_{x^{*}}$.

**Definition 2 (Spurious Covariance).** The spurious covariance between treatment $X = x^{*}$ and outcome $Y$ is:

$$\text{Cov}^{s}_{x^{*}}(X, Y) = \text{Cov}(X, Y_{x^{*}}).$$

**Property 1.** $|\Pi^{s}(X, Y)| = 0 \Rightarrow \text{Cov}^{s}_{x^{*}}(X, Y) = 0.$

The causal covariance can naturally be defined as the difference between the total and spurious covariance.

**Definition 3 (Causal Covariance).** The causal covariance of the treatment $X = x^{*}$ and the outcome $Y$ is:

$$\text{Cov}^{c}_{x^{*}}(X, Y) = \text{Cov}(X, Y - Y_{x^{*}}).$$

Prop. 2 establishes the correspondence between the causal paths and the causal covariance – if there is no causal path from $X$ to $Y$ in the underlying model, the causal covariance equates to zero.

**Property 2.** $|\Pi^{c}(X, Y)| = 0 \Rightarrow \text{Cov}^{c}_{x^{*}}(X, Y) = 0.$

We consider more detailed measures corresponding to the different causal pathways, and first, the direct path:

---

5 Mediators (relative to $X$ and $Y$) are a set of variables $W \subseteq De(X) \cap Non-De(Y)$ such that $|\Pi^{c}(X, Y|W)| = 0$.

6 An alternative way to see the replacement operation relative to $U_{i}$ is to envision a system where $U_{i}$ is observed.
Definition 4 (Direct Covariance). Given a semi-Markovian model $M$, let the set $W$ be the mediators between $X$ and $Y$. The pure ($\text{Cov}_{dp}^{yp}(X,Y)$) and total ($\text{Cov}_{dt}^{yp}(X,Y)$) direct covariance of the treatment $X = x^*$ on the outcome $Y$ are defined respectively as

$$
\text{Cov}_{dp}^{yp}(X,Y) = \text{Cov}(X, Y - Y_{x^*,W}) \quad \text{and} \quad \text{Cov}_{dt}^{yp}(X,Y) = \text{Cov}(X, Y_{W_{x^*}} - Y_{x^*}).
$$

(4)

(5)

By the composition axiom [Pearl, 2000, Ch. 7.3], Eqs. 4 and 5 can be explicitly written as follows:\footnote{Consider Eq. 4 as an example. For any $U = u$, $Y_{X(u),W(u)} = Y_{x^*,W(u)}$ if $X(u) = x^*, W(u) = w$. By the composition axiom, $X(u) = x^*, W(u) = w$ implies $Y(u) = Y_{x^*,W(u)}$, which in turn gives $Y_{X(u),W(u)}(u) = Y(u)$. Averaging $u$ over $P(u)$, we obtain $Y_{X,W} = Y$.}

$$
\text{Cov}(X, Y - Y_{x^*,W}) = \text{Cov}(X, Y, W) - \text{Cov}(X, W, Y_{x^*}).
$$

(6)

The counterfactual pure direct covariance (Eq. 4) is shown graphically in Fig. 2, where (a) corresponds to the $Y$-side, and (b) to the $Y_{x^*,W}$-side. Note that from mediator $W$ perspective, $X$ remains at the level that it would naturally have attained, while the “direct” input from $X$ to $Y$ varies from its natural level (Fig. 2a) to $do(x^*)$ (b). The change of the outcome $Y$ thus measures the effect of the direct path. A similar analysis also applies to the total direct covariance (Eq. 5).

Property 3. $\text{Cov}_{dp}^{yp}(X,Y) = \text{Cov}_{dt}^{yp}(X,Y) = 0$ if $X$ is not a parent of $Y$ (i.e., $X \not\in \text{Pa}(Y)$).

We can turn around the definitions of direct covariance and provide operational estimands for indirect paths.

Definition 5 (Indirect Covariance). Given a semi-Markovian model $M$, let the set $W$ be the mediators between $X$ and $Y$. The pure ($\text{Cov}_{ip}^{yp}(X,Y)$) and total ($\text{Cov}_{it}^{yp}(X,Y)$) indirect covariance of the treatment $X = x^*$ on the outcome $Y$ are defined respectively as:

$$
\text{Cov}_{ip}^{yp}(X,Y) = \text{Cov}(X, Y - Y_{W_{x^*}}), \quad \text{Cov}_{it}^{yp}(X,Y) = \text{Cov}(X, Y_{W_{x^*,W}} - Y_{x^*}).
$$

(6)

(7)

Eqs. 6 and 7 correspond to the indirect paths, since they capture the covariance of $X$ and $Y$, but only via paths mediated by $W$. The first argument of $X$ is the same in both halves of the contrast, but this value can either be $x^*$ (Eq. 7) or at the level that $X$ would naturally attain without intervention (Eq. 6).

Property 4. $|\Pi^i(X,Y)| = 0 \Rightarrow \text{Cov}_{ip}^{yp}(X,Y) = \text{Cov}_{it}^{yp}(X,Y) = 0$.

Putting these definitions together, we can prove a general non-parametric decomposition of $\text{Cov}(X,Y)$:

$$
\text{Cov}(X,Y) = \text{Cov}_{dp}^{yp}(X,Y) + \text{Cov}_{ip}^{yp}(X,Y) + \text{Cov}_{it}^{yp}(X,Y).
$$

(8)

Theorem 1. $\text{Cov}(X,Y)$, $\text{Cov}_{ip}^{yp}(X,Y)$ and $\text{Cov}_{it}^{yp}(X,Y)$ obey the following non-parametric relationship:

$$
\text{Cov}(X,Y) = \text{Cov}_{ip}^{yp}(X,Y) + \text{Cov}_{it}^{yp}(X,Y) = \text{Cov}_{dp}^{yp}(X,Y) + \text{Cov}_{ip}^{yp}(X,Y).
$$

(8)

In other words, the covariance between $X$ and $Y$ can be partitioned into its corresponding direct, indirect, and spurious components. In particular, Thm. 1 coincides with Eq. 1 in the linear-standard model.

Corollary 1. In the linear-standard model, for any $x^*$, $\text{Cov}_{ip}^{yp}(X,Y)$, $\text{Cov}_{it}^{yp}(X,Y)$, $\text{Cov}_{dp}^{yp}(X,Y)$ and $\text{Cov}_{ip}^{yp}(X,Y)$ are equal to:

$$
\text{Cov}_{ip}^{yp}(X,Y) = \text{Cov}_{it}^{yp}(X,Y) = \alpha_X Y,
$$

$$
\text{Cov}_{ip}^{yp}(X,Y) = \text{Cov}_{it}^{yp}(X,Y) = \alpha_{W_{x^*}} Y.
$$

Corol. 1 says that the proposed decomposition (Thm. 1) does not depend on the value of $do(x^*)$ in the linear model of Fig. 1(a), which is not achievable in previous value-specific decompositions [Pearl, 2001; Zhang and Bareinboim, 2018a].\footnote{For the nonlinear models, the decomposing terms (e.g., $\text{Cov}_{ip}^{yp}(X,Y)$) are still sensitive to the target level $do(x^*)$. To circumvent the challenges of picking a specific decision value, one could assign a randomized treatment $do(x^* \sim P(X))$, where $P(X)$ is the distribution over the treatment $X$ induced by the underlying causal model.}

4 DECOMPOSING CAUSAL RELATIONS

We now focus on the challenge of decomposing the causal covariance into more elementary components. We use the two-mediators setting (Fig. 1(b)) as example, where $X$ and $Y$ are connected through four causal paths: through both $W_1, W_2$ ($g_1 : X \rightarrow W_1 \rightarrow W_2 \rightarrow Y$), only through $W_1$ ($g_2 : X \rightarrow W_1 \rightarrow Y$), only through $W_2$ ($g_3 : X \rightarrow W_2 \rightarrow Y$), and directly ($g_4 : X \rightarrow Y$). Our goal is to decompose the $\text{Cov}_{ip}^{yp}(X,Y)$ over the paths $g_{1,4}$. Our analysis applies to semi-Markovian models, without loss of generality, and the Markovian example (Fig. 1(b)) is used for simplicity of the exposition.
For a node $S_i \in Pa(Y)$ and a set of causal paths $\pi$, the edge $S_i \rightarrow Y$ defines a funnel operator $\phi_{S_i \rightarrow Y}$, which maps from $\pi$ to the set of paths $\phi_{S_i \rightarrow Y}(\pi)$ obtained from $\pi$ by replacing all paths of the form $X \rightarrow \cdots \rightarrow S_i \rightarrow Y$ with $X \rightarrow \cdots \rightarrow S_i$, and removing all the other paths. As an example, for $\pi = \{g_1, g_2, g_3\}$, $\phi_{S_2 \rightarrow Y}(\pi) = \{g_1(X, W_2), g_3(X, W_2)\}$, where $g_1(X, W_2)$ is the subpath $X \rightarrow W_1 \rightarrow W_2$ and $g_3(X, W_2)$ is the subpath $X \rightarrow W_2$. We next formalize the notion of path-specific interventions, which isolates the influence of the intervention $do(\pi[x^*])$ passing through a subset $\pi$ of causal paths from $X$, denoted by $do(\pi[x^*])$ (a similar notion has been introduced by [Pearl, 2001], and then [Avin et al., 2005; Shpitser and Tchetgen, 2016]).

**Definition 6** (Path-Specific Potential Response). For a SCM $M$ and an arbitrary variable $Y \in V$, let $\pi$ be a set of causal paths. Let $X$ be the source variables of paths in $\pi$. Further, let $X_{\pi \rightarrow Y} = \{X_i : \forall X_i \in X, X_i \rightarrow Y \in \pi\}$ and $S = (Pa(Y) \cap V) \setminus X_{\pi \rightarrow Y}$. The $\pi$-specific potential response of $Y$ to the intervention $do(\pi[x^*])$ in the situation $U = u$, denoted by $Y_{\pi[x^*]}(u)$, is defined as:

$$Y_{\pi[x^*]}(u) = \begin{cases} Y_{\pi \rightarrow Y \setminus S \rightarrow Y \setminus \pi[x^*]}(u) & \text{if } \pi \neq \emptyset \\ Y(u) & \text{otherwise} \end{cases}$$

where $S_{\pi \rightarrow Y \setminus \pi[x^*]}(u)$ is a set of $\pi$-specific potential response $\{S_i_{\pi \rightarrow Y \setminus \pi[x^*]}(u) : S_i \in S\}$.

Despite the non-trivial notation, the $\pi$-specific counterfactual $Y_{\pi[x^*]}$ is simply assigning the treatment $do(\pi[x^*])$ exclusively to the causal paths in $\pi$, while allowing all the other causal paths to behave naturally. This contrasts with the counterfactual $Y_{\pi[x^*]}$, which can be seen as assigning the treatment $do(\pi[x^*])$ to all causal paths from $X$ to $Y$. For instance, repeatedly applying Def. 6 to $g_1 : X \rightarrow W_1 \rightarrow W_2 \rightarrow Y$ (see Appendix 2.1), we obtain the $g_1$-specific potential response $Y_{g_1[x^*]}$ as

$$Y_{g_1[x^*]} = Y_{X, W_1, W_2, W_1:x, W_1:x^*} = YW_{2W_1:x^*}.$$  

The intervention $do(g_1[x^*])$ can be visualized more immediately through its graphical representation (Fig. 3(b)) – the treatment $do(\pi[x^*])$ is assigned throughout $g_1$ while all the other paths are kept at the level that it would have attained “naturally” following $X$. The difference of the outcome $Y$ (induced by $do(g_1[x^*])$) and the unintervened $Y$ (Fig. 3(a)) measures the relative strength of $g_1$ itself, which leads to the following definition.

**Definition 7** (Pure Path-Specific Causal Covariance). For a semi-Markovian model $M$ and an arbitrary causal path $g$ from $X$, the pure $g$-specific causal covariance of the treatment $X = x^*$ on the outcome $Y$ is defined as:

$$\text{Cov}^g_{g[x^*]}(X, Y) = \text{Cov}(X, Y - Y_{g[x^*]}).$$

The following property establishes the correspondence between a causal path and its path-specific estimand.

**Property 5.** $g \not\in \Pi^c(X, Y) \Rightarrow \text{Cov}^g_{g[x^*]}(X, Y) = 0$.

Prop. 5 follows immediately as a corollary of Lem. 1, which implies that the counterfactuals $Y_{\pi(g)[x^*]}$ and $Y_{\pi(g)[x^*]}$ define the same variable over $U$ if $g$ is not a causal path from $X$ to $Y$. 

![Figure 3: Graphical representations of the causal covariation specific to $g_1$ (a-b), $g_{[2,3]}$ (c-d) and $g_4$ (e-f).](image-url)
Lemma 1. \( g \notin \Pi^c(X,Y) \Rightarrow Y_{\pi(g)[x]}(u) = Y_{\pi(g)\cup\{g\}[x]}(u) \).

Considering again the model in Fig. 1(b), let \( g_{[i,j]} = \{g_k \text{ s.t. } i < k < j\} \). Recall that \( g_i = \{X \rightarrow Y\} \), and note that the \( g_i \)-specific causal covariance can be computed using \( \pi(g_i) = g_{[1,i]} \), yielding:

\[
\text{Cov}_{g_i}(X,Y)^{\pi} = \text{Cov}(X, Y_{g_{[1,i]}[x]} - Y_{g_{[1,i]4}[x]}) = \text{Cov}(X, Y_{W_{1x}.W_{2x}} - Y_{x}),(11)
\]

which coincides with the direct effect (Eq. 5 with \( W = \{W_1, W_2\} \)). Fig. 3(c-f) shows a graphical representation of this procedure.

The path-specific quantity given in Def. 8 has another desirable property, namely, the causal covariance \( \text{Cov}_c^c(X,Y) \) can be decomposed as a summation over causal paths from \( X \) to \( Y \). To witness, first let an order over \( \Pi^c(X,Y) \) be \( \mathcal{L}^c : g_1 < \cdots < g_n \). For a path \( g_i \in \Pi^c(X,Y) \), the order \( \mathcal{L}^c \) defines a function \( \mathcal{L}^c_x \) which maps from \( g_i \) to a set of paths \( \mathcal{L}^c_x(g_i) \) that precede \( g_i \) in \( \mathcal{L}^c \), i.e., \( \mathcal{L}^c_x(g_i) = g_{[1,i-1]} \). We derive in the sequel a path-specific decomposition formula for the causal covariance relative to an order \( \mathcal{L}^c \).

**Theorem 2.** For a semi-Markovian model \( M \), let \( \mathcal{L}^c \) be an order over \( \Pi^c(X,Y) \). For any \( x^* \), the following non-parametric relationship hold:

\[
\text{Cov}^c_c(X,Y) = \sum_{g \in \Pi^c(X,Y)} \text{Cov}_{g[x]}^c(X,Y)\mathcal{L}^c_x.
\]

Thm. 2 can be demonstrated in the model of Fig. 1(a). Let an order \( \mathcal{L}^c \) over \( g_{[1,4]} \) be \( g_i < g_j \) if \( i < j \). First note that the path-specific causal covariance of \( g_2 \) \( \text{Cov}_{g_2[x]}^c(X,Y)\mathcal{L}^c_x \) and \( g_3 \) \( \text{Cov}_{g_3[x]}^c(X,Y)\mathcal{L}^c_x \) are equal to, respectively,

\[
\text{Cov}(X, Y_{W_{1x}.W_{2x}} - Y_{W_{1x}.W_{2x}}) \quad (12)
\]

\[
\text{Cov}(X, Y_{W_{1x}.W_{2x}} - Y_{W_{1x}.W_{2x}}) \quad (13)
\]

The causal covariance \( \text{Cov}^c_c(X,Y) \) can then be decomposed as the sum of Eqs. 10-13, respectively, \( g_1, g_4, g_2, g_3 \). Fig. 3 describes this decomposition procedure: we measure the difference of the outcome \( Y \) as the intervention \( do(x^*) \) propagates through paths \( g_1, g_2, g_3, g_4 \). The sum of these differences thus equate to the total influence of the intervention \( do(x^*) \) to the outcome \( Y \), i.e., the causal covariance \( \text{Cov}^c_c(X,Y) \).

5 DECOMPOSING SPURIOUS RELATIONS

We introduce in the sequel a new strategy to decompose the spurious covariance (Def. 2), which will play a central role in the analysis of the spurious relations relative to the pair \( X, Y \). The spurious covariance measures the correlation between the observational \( X \) and the counterfactual \( Y_x \) (Def. 2). We will employ in our analysis the twin network [Balke and Pearl, 1994; Pearl, 2000, Sec. 7.1.4], which is a graphical method to analyzing the relation between observational and counterfactual variables.

Consider the causal model \( M \) in Fig. 4(a), for example, where the exogenous variables \( \{U_1, U_2\} \) are shown explicitly. Its twin network is the union of the model \( M \) (factual) and the submodel \( M_{x*} \) (counterfactual) under intervention \( do(x^*) \), which is shown in Fig. 4(b). The factual \( (M \) and counterfactual \( (M_{x*} \) worlds share only the exogenous variables (in this case, \( U_1, U_2 \)), which constitute the invariances shared across worlds. In this twin network, the observational \( X \) and the counterfactual \( Y_{x*} \) are connected through two paths: one through \( U_1 \) and the other through \( U_2 \). These paths correspond to two pathways from \( X \) to \( Y \) in the original causal diagram:

\[
\tau_1 : X \leftarrow Z_2 \leftarrow Z_1 \leftarrow U_1 \rightarrow Z_1 \rightarrow Z_2 \rightarrow Y ,
\]

\[
\tau_2 : X \leftarrow Z_2 \leftarrow U_2 \rightarrow Z_2 \rightarrow Y.
\]

Note that when considering the corresponding paths in the original graph (Fig. 4(a)), these paths \( (\tau_1, \tau_2) \) are not necessarily simple, i.e., they may contain a particular node more than once. Furthermore, each path can be partitioned into a pair of causal paths (say, \( g_1, g_r \)) from a common source \( U_i \in U \) (e.g., \( \tau_1 \) consists of a pair \( (g_{l_1}, g_{r_1}) \), where \( g_{l_1} : U_1 \rightarrow Z_1 \rightarrow Z_2 \rightarrow X \), and \( g_{r_1} : U_1 \rightarrow Z_1 \rightarrow Z_2 \rightarrow Y \)). Indeed, these non-simple paths are referred to as treks in the causal inference literature, which usually has been studied in the context of linear models [Spirtes et al., 2001; Sullivant et al., 2010].

**Definition 9** (Trek). A trek \( \tau \) in \( G \) (from \( X \) to \( Y \) is an ordered pair of causal paths \( (g_1, g_r) \) with a common exogenous source \( U_i \in U \) such that \( g_1 \in \Pi^c(U_i, X) \) and \( g_r \in \Pi^c(U_i, Y) \). The common source \( U_i \) is called the top of the trek, denoted \( top(g_1, g_r) \). A trek is spurious if \( g_r \in \Pi^c(U_i, Y|X) \), i.e., \( g_r \) is a causal path from \( U_i \) to \( Y \) that is not intercepted by \( X \).
We denote the set of treks from \( X \) to \( Y \) in \( G \) by \( T(X, Y)G \) and spurious treks by \( T^s(X, Y)G \) (\( G \) will be omitted when obvious). We introduce next an estimator for a specific spurious trek. For a spurious trek \( \tau = (g_l, g_r) \) with \( U_i = \text{top}(\tau) \), first let \( X_{g_l} \) denote the path-specific potential response \( X_{g_l|U_i'[\tau]} \), where \( U_i' \) is an i.i.d. draw from the distribution \( P(U_i) \). Similarly, let \( Y_{x^*, g_r} = Y_{x^*, g_r|U_i'[\tau]} \), where \( U_i' \sim P(U_i) \). Pure trek-specific covariance can then finally be defined.

**Definition 10** (Pure Trek-Specific Spurious Covariance).

For a semi-Markovian model \( M \) and a spurious trek \( \tau = (g_l, g_r) \), the pure \( \tau \)-specific spurious covariance of the treatment \( X = x^* \) on the outcome \( Y \) is defined as:

\[
\text{Cov}^{ts}_{\tau[x^*]}(X, Y) = \text{Cov}(X - X_{g_l}, Y_{x^*} - Y_{x^*, g_r}).
\]

In words, the differences \( X - X_{g_l} \) and \( Y_{x^*} - Y_{x^*, g_r} \) are simply measuring the effects of the causal paths \( g_l \) and \( g_r \) (Lem. 1), while the \( \text{Cov}(\cdot) \) operator is in charge of compounding them. (In the extreme case when \( g_l \) or \( g_r \) are disconnected, the pure \( \tau \)-specific spurious covariance will equate to zero.) For example, the pure \( \tau_l \)-specific spurious covariance \( \text{Cov}^{ts}_{\tau_l[x^*]}(X, Y) \) in Fig. 4(a) is

\[
\text{Cov}(X - X_{g_{l1}}, Y_{x^*} - Y_{x^*, g_{r1}}) = \text{Cov}(X - X_{g_{l2}}, Y_{x^*} - Y_{x^*, g_{r2}}).
\]

This quantity can be more easily seen through its graphical representation in Fig. 5 (top). The main idea is to decompose \( U_i \) into two independent components \( U_i'[\tau_l], U_i'[\tau_r] \) (Fig. 5b), which is then contrasted with the world in which \( U_i \) is kept intact (a).\(^{11} \)\(^{12} \)

We note that by Def. 6, \( X = X_0 \) and \( Y_{x^*} = Y_{x^*, \emptyset} \). The pure \( \tau_l \)-spurious specific covariance can be written as:

\[
\text{Cov}^{ts}_{\tau_l[x^*]}(X, Y) = \text{Cov}(X_0 - X_{g_{l1}}, Y_{x^*, \emptyset} - Y_{x^*, g_{r1}}).
\]

More generally, the pure trek-specific spurious covariance for \( \tau = (g_l, g_r) \) measures the covariance of variables \( X_{\pi_l} - X_{\pi_l \cup \{g_l\}} \) and \( Y_{x^*, \pi_l} - Y_{x^*, \pi_l \cup \{g_r\}} \), where \( \pi_l \) (\( \pi_r \)) is an arbitrary set of causal paths from \( U \) that does not contain \( g_l \) (\( g_r \)). This observation will be useful later on, which leads to the trek-specific spurious covariance.

\[
\text{Cov}^{ts}_{\tau_l[x^*]}(X, Y) = \text{Cov}(X_0 - X_{g_l}, Y_{x^*, \emptyset} - Y_{x^*, g_r}).
\]

![Figure 5: The decomposition procedure of the spurious covariance over the spurious treks \( \tau_1, \tau_2 \) (Thm. 3).](image)

**Definition 11** (Trek-Specific Spurious Covariance). For a semi-Markovian model \( M \), let \( \tau \) be a spurious trek \( (g_l, g_r) \) and \( \pi \) be a function mapping \( \tau \) to a pair \( \pi(\tau) = (\pi_l, \pi_r) \) where \( \pi_l \) and \( \pi_r \) are sets of causal paths from \( U \) such that \( g_l \notin \pi_l \) and \( g_r \notin \pi_r \). The \( \tau \)-specific spurious covariance of the treatment \( X = x^* \) on the outcome \( Y \), denoted by \( \text{Cov}^{ts}_{\tau[x^*]}(X, Y) \), is defined as

\[
\text{Cov}(X_{\pi_l} - X_{\pi_l \cup \{g_l\}}, Y_{x^*, \pi_r} - Y_{x^*, \pi_r \cup \{g_r\}}).
\]

The next proposition establishes the relationship between Def. 11 and the corresponding spurious treks. This property can be seen as a necessary condition for any measure of strength for spurious relations.

**Property 6.** \( \tau \notin T^s(X, Y) \) \( \Rightarrow \text{Cov}^{ts}_{\tau[x^*]}(X, Y) \pi = 0 \).

As an example of Def. 11, the trek \( \tau_2 \) in Fig. 4(a) consists of paths \( g_{l2} : U_2 \rightarrow Z_2 \rightarrow X \) and \( g_{r2} : U_2 \rightarrow Z_2 \rightarrow Y \). If we set \( \pi(\tau_2) = (\{g_{l1}\}, \{g_{r1}\}) \), the \( \tau_2 \)-specific spurious covariance can be measured by \( \text{Cov}^{ts}_{\tau_2[x^*]}(X, Y) \pi \), i.e.,

\[
\begin{align*}
\text{Cov}(X_{g_{l1}} - X_{g_{l1}[\tau_2]}, Y_{x^*, g_{r1}} - Y_{x^*, g_{r1}[\tau_2]} & = \text{Cov}(X_{U_1} - X_{U_1'[\tau_2]}, Y_{x^*, U_1} - Y_{x^*, U_1'[\tau_2]}) \quad (15) \\
& (16)
\end{align*}
\]

Eq. 16 is graphically represented in Fig. 5(c-d), where the effect of the trek \( \tau_2 \) is measured. In words, the difference between Fig. 5(c) and (d) is the effect of the causal paths \( g_{l2} \) and \( g_{r2} \) when \( U_2 \) is kept intact versus when divided into two independent components \( U_2', U_2'' \).

Armed with the definition of trek-specific spurious covariance, we can finally study the decomposability of the spurious covariance \( \text{Cov}^{s}_{x^*}(X, Y) \) (Def. 2). First, let \( U^* \subseteq U \) denote the maximal set of exogenous variables that simultaneously affect variables \( X \) and \( Y \) (common exogenous ancestors), and let an order over \( U^* \) be \( U^*_n : U_1 < \cdots < U_n \). For each \( U_i \in U^* \), let \( U_i^* \) be an order \( g^*_1 < \cdots < g^*_n \) over the set \( \Pi^*(U_i, X) \). Similarly, we define \( L^*_r \) for \( \Pi^*(U_i, Y|X) \). The tuple \( L^* = (L^*_a, \{L^*_l, L^*_r\})_{1 \leq i \leq |U^*|} \) thus defines an order

---

\(^{10}\)\(^{11}\)\(^{12}\)
for the spurious treks $T^s(X,Y)$. We denote $L^s$ a function which maps from a trek $\tau$ to sets of paths $L^s(\tau)$ covered by the spurious treks preceding $\tau$ in $L^s$. Formally, given a spurious trek $\tau = (g_l^1, g_r^1)$, $L^s(\tau)$ is equal to 
$$(\Pi^s(U_{[1,i-1]},X) \cup g_l^1_{[1,i-1]}, \Pi^s(U_{[1,i-1]},Y|X) \cup g_r^1_{[1,k-i-1]}).$$

We are now ready to derive the decomposition formula for the spurious covariance $\text{Cov}^s_x(X,Y)$. 

**Theorem 3.** For a semi-Markovian model $M$, let $L^s = \langle L^s, \{L^s_{U,Y}\}_{1 \leq i \in [0..]} \rangle$ be an order over spurious treks $T^s(X,Y)$. For any $x^*$, the following non-parametric relationship hold:

$$\text{Cov}^s_x(X,Y) = \sum_{\tau \in T^s(X,Y)} \text{Cov}^s_{\tau(x^*)}(X,Y) L^s_{\tau}$$

For example, in the model of Fig. 4(a), $U^s = \{U_1, U_2\}$. $t_1$ ($t_2$) is the spurious trek associated with $U_1$ ($U_2$). If we consider the order $L^s$ such that $L^s_{\tau} : U_1 < U_2$, Thm. 3 dictates that $\text{Cov}^s_x(X,Y)$ should be decomposed as the sum of Eqs. 14 and 15. Fig. 5 shows the graphical representation of this decomposition procedure: we measure the change of the covariance between $X$ and $Y$ as we disconnect the relations going through $t_1$ (associated with $U_1$) and $t_2$ ($U_2$), sequentially. The sum of these changes thus equates to the correlations of $X$ and $Y$ along the spurious pathways, i.e., the spurious covariance $\text{Cov}^s_{\tau(x^*)}(X,Y)$. (See Appendix 2 for more examples.)

**6 NON-PARAMETRIC PATH ANALYSIS**

In this section, we put the results of the previous sections together and derive a general path-specific decomposition for the covariance of the treatment $X$ and the outcome $Y$ without assuming any specific parametric form. 

We start by noting that each spurious path from $X$ to $Y$ corresponds to a unique set of spurious treks that start on $X$ and end in $Y$. Recall that a spurious path $l$ can be seen as a pair of causal paths $(g_l, g_r)$, where the only node shared among $g_l$ and $g_r$ is the common source. For example, the spurious path $l : X \leftarrow Z \rightarrow Y$ is a pair $(g_l, g_r)$ such that $g_l : Z \rightarrow X$ and $g_r : Z \rightarrow Y$. We can thus define a rule $f$ which maps a trek $\tau \in T^s(X,Y)$ to a spurious path $l \in \Pi^s(X,Y)$. For $\tau = (g_l, g_r)$, let $V_l$ be the most distant recurring node from top$(g_l, g_r)$ such that $V_l$ is the only node shared among subpaths $g_l(V_l, X)$ and $g_r(V_l, Y)$; the pair $(g_l(V_l, X), g_r(V_l, Y))$ corresponds to a path $l$ in $\Pi^s(X,Y)$. As an example, the trek $t_1$ in Fig. 4(a) has $V_l = Z$, which corresponds to the spurious path $l : X \leftarrow Z \rightarrow Y$, and similarly, $f(t_1) = l$ as well as $f(t_2) = l$. Lem. 2 shows that the rule $f$ forms a valid surjective function.

**Lemma 2.** For a semi-Markovian model $M$, for each spurious trek $\tau \in T^s(X,Y)$, there always exists a unique most distant recurring node $V_l$.

For a spurious path $l$, let $T^s(l) = f^{-1}(l)$ denote its corresponding treks. Specifically, if $l \notin \Pi^s(X,Y)$, then for each $\tau \in T^s(l)$, we must have $\tau \notin T^s(X,Y)$. For instance, if the spurious $l$ in Fig. 4(a) is disconnected, e.g., $Z_2 \neq X$, treks $t_1$, $t_2$ are both disconnected as well. From this observation, we could naturally define the spurious covariance of a path $l$ as a sum over treks in $T^s(l)$.

**Definition 12 (Path-Specific Spurious Covariance).** For a semi-Markovian model $M$ with an associated causal diagram $G$, let $l$ be an arbitrary spurious path in $G$. Let $\pi$ be a function that maps a trek $\tau = (g_l, g_r) \in T^s(l)$ to a pair $\pi(\tau) = (\pi_1, \pi_r)$, where $\pi_1$ and $\pi_r$ are arbitrary sets of causal paths from $U$ such that $g_l \notin \pi_1$ and $g_r \notin \pi_r$. The $l$-specific spurious covariance of the treatment $X = x^*$ on the outcome $Y$ is defined as

$$\text{Cov}_{\pi(\tau)}^s(X,Y) = \sum_{\tau \in T^s(l)} \text{Cov}_{\pi(\tau)}^s(X,Y) \tau$$

**Property 7.** $l \notin \Pi^s(X,Y) \Rightarrow \text{Cov}_{\pi(\tau)}^s(X,Y) = 0$.

The surjectivity of the function $f$ assures that the set $\{T^s(l)\}_{l \in \Pi^s(X,Y)}$ forms a partition over the spurious treks $T^s(X,Y)$. From Thm. 3, it follows immediately that the path-specific spurious covariance (Def. 12) has the property that expresses the spurious covariance $\text{Cov}^s_x(X,Y)$ as a sum over $\Pi^s(X,Y)$.

**Theorem 4.** For a semi-Markovian model $M$, let $L^s = \langle L^s, \{L^s_{U,Y}\}_{1 \leq i \in [0..]} \rangle$ be an order over spurious treks $T^s(X,Y)$. For any $x^*$, the following non-parametric relationship hold:

$$\text{Cov}^s_x(X,Y) = \sum_{l \in \Pi^s(X,Y)} \text{Cov}_{\pi(l)}^s(X,Y) L^s_{\tau}$$

As an example, the path $l : X \leftarrow Z \rightarrow Y$ in Fig. 4(a) corresponds to $T^s(l) = \{t_1, t_2\}$. For an arbitrary order $L^s$, Thm. 4 is applicable and immediately yields $\text{Cov}^s_x(X,Y) = \sum_{l \in \Pi^s(X,Y)} \text{Cov}_{\pi(l)}^s(X,Y) L^s_{\tau}$, which means that the path $l$ accounts for all the spurious relations between $X$ and $Y$. In other words, the spurious joint variability of $X$ and $Y$ is fully explained by the variance of $Z_2$, which is a function of the exogenous variables $U_1$ (through $t_1$) and $U_2$ (through $t_2$).

Thms. 1-4 together lead to a general path-specific decomposition formula, which allows one to non-parametrically decompose the covariance $\text{Cov}(X,Y)$ over all open paths from $X$ to $Y$ in the underlying model.

**Theorem 5 (Path-Specific Decomposition).** For a semi-Markovian model $M$, let $L^s$ be an order over $\Pi^s(X,Y)$.
and \( \mathcal{L}^* = \{ \mathcal{L}_{i}^* \} \) be an order over \( \mathcal{X}^*(X,Y) \). For any \( x^* \), the following non-parametric relationship hold:

\[
\text{Cov}(X, Y) = \sum_{l \in \mathcal{L}^*} \text{Cov}_{l(x^*)}^c(X, Y) \mathcal{L}_{l}^* + \sum_{l \notin \mathcal{L}^*} \text{Cov}_{l(x^*)}^c(X, Y) \mathcal{L}_{l}^*.
\]  

We illustrate the use of Thm. 5 using the model discussed in Sec. 1 (Fig. 1(a)). Recall that \( X \) and \( Y \) are connected through the causal paths \( l_1, l_2 \) and spurious paths \( l_3, l_4 \). Note that \( U^x = \{ U_Z \} \) sparsely affects the treatment \( X \) through the path \( g_l = U_Z \rightarrow Z \rightarrow X \), and the outcome \( Y \) through the paths \( g_{r1} = U_Z \rightarrow Z \rightarrow Y \) and \( g_{r2} = U_Z \rightarrow Z \rightarrow W \rightarrow Y \). Let order \( L^c \) be \( l_1 < l_2 \) and \( L^* \) be \( g_{r1} < g_{r2} \). For any level \( x^* \), Thm. 5 equates the covariance \( \text{Cov}(X, Y) \) to the sum of \( \{ \text{Cov}_{l(x^*)}^c(X, Y) \mathcal{L}_{l}^* \} \) and \( \{ \text{Cov}_{l(x^*)}^c(X, Y) \mathcal{L}_{l}^* \} \) which can be written as

\[
\text{Cov}(X, Y - Y_{x^*}, W) + \text{Cov}(X, Y_{x^*}, W - Y_{x^*}) + \text{Cov}(X - X_{U_l}, Y_{x^*} - Y_{x^*, W, z_{U_l}}) + \text{Cov}(X - X_{U_l}, Y_{x^*}, W_{x^*, z_{U_l}} - Y_{x^*, U_l}) + \text{Cov}(X - X_{U_l}, Y_{x^*}, W_{x^*, z_{U_l}} - Y_{x^*, U_l})
\]  

which are all well-defined, computable from the structural causal model [Def. 1; Pearl, 2000, Sec. 7.1].

7 IDENTIFYING PATH-SPECIFIC DECOMPOSITION

By and large, identifiability is one of the most studied topics in causal inference. It is acknowledged in the literature that obtaining identifiability may be non-trivial even in the context of less granular measures of causal effects, including quantities without nested counterfactual and following the do-calculus analysis.

In this section, we start the study of identifiability conditions for when the path-specific decomposition formula (Thm. 5) can be estimated from data, when the SCM is not fully known. We’ll analyze the causal model discussed in Fig. 1(a) given its generality and potential to encode more complex models. The main assumption encoded in this model is Markovianity, i.e., that all exogenous variables are independent. We show next that identifiability can be obtained under these assumptions.

**Theorem 6.** The path-specific decomposition of Eq. 18 is identifiable if the distributions \( P(x, y_{x^*}) \), \( P(x, y_{x^*, W}) \) and \( P(x, y_{x^*, W, z_{U_l}}) \) are identifiable. Specifically, in the model of Fig. 1(a), \( P(x, y_{x^*}) \), \( P(x, y_{x^*, W}) \), and \( P(x, y_{x^*, W, z_{U_l}}) \) can be estimated, respectively, from the observational distribution \( P(x, y, z, w) \) as follows:
\[
P(x, y_{x^*}) = \sum_{z, w} P(y|x_{x^*}, w)P(w|x_{x^*})P(x, z)
\]
\[
P(x, y_{x^*, W}) = \sum_{z, w} P(y|x_{x^*}, z, w)P(x, z, w)
\]
\[
P(x, y_{x^*, W, z_{U_l}}) = \sum_{z, w, z'} P(y|x_{x^*}, z, w)P(w|x_{x^*}, z')P(x, z')P(z)
\]

Note that all the quantities listed in Thm. 6 are expressible in terms of conditional distributions and do not involve any counterfactual (simple nor nested), which are readily estimable from the observational distribution.

As an example, the \( l_2 \)-specific causal covariance \( \text{Cov}_{l_2(x^*)}^c(X, Y) \mathcal{L}_{l_2}^* \) in Eq. 18 can be written as \( \text{Cov}(X, Y_{x^*, W}) - \text{Cov}(X, Y_{x^*}) \), which is computed from the counterfactual distributions \( P(x, y_{x^*}) \) and \( P(x, y_{x^*, W}) \), respectively. These distributions can be estimated from the observational distribution \( P(x, y, z, w) \) following Thm. 6. Indeed, the path-specific decomposition formula (Thm. 5) is identifiable in the model of Fig. 1(a) regardless of the order \( L^c \) and \( L^* \). (For identifications of other decompositions, see Appendix 1.)

We further considered the identifiability conditions for the path-specific decomposition formula when the more stringent assumption that the underlying structural functions are linear is imposed.

**Theorem 7.** Under the assumption of linearity and the assumption of Fig. 1(a), for any arbitrary orders \( L^c \) and \( L^* \), for any \( x^* \), the path-specific covariance of \( l_1, l_2, l_3 \) and \( l_4 \) are equal to:

\[
\text{Cov}_{l_1(x^*)}^c(X, Y) \mathcal{L}_{l_1}^* = \alpha_{X}, \quad \text{Cov}_{l_2(x^*)}^c(X, Y) \mathcal{L}_{l_2}^* = \alpha_{XY}\sigma_{WY} \quad \text{Cov}_{l_3(x^*)}^c(X, Y) \mathcal{L}_{l_3}^* = \alpha_{XW}\sigma_{ZW} \quad \text{Cov}_{l_4(x^*)}^c(X, Y) \mathcal{L}_{l_4}^* = \alpha_{Z}\sigma_{WXY}
\]

The parameters \( \alpha \) can be estimated from the corresponding (partial) regression coefficients [Pearl, 2000, Ch. 5].

Clearly, after applying Thm. 7 to Eq. 18, the resulting decomposition coincides with Wright’s method of path coefficients in the linear-standard model (Eq. 1).

8 CONCLUSIONS

We introduced novel covariance-based counterfactual measures to account for effects along with a specific path from a treatment \( X \) to an outcome \( Y \) (Defs. 8, 11-12). We developed machinery to allow, for the first time, the non-parametric decomposition of the covariance of \( X \) and \( Y \) as a summation over different pathways in the underlying causal model (Thm. 5). We further provided identification conditions under which the decomposition formula can be estimated from data (Thm. 6-7).

Acknowledgments

Bareinboim and Zhang are supported in parts by grants from NSF IIS-1704352 and IIS-1750807 (CAREER).
References


1 PROOFS

Proofs build on the exclusion and independence restrictions rules of SCMs [Pearl, 2000, pp. 232], and three axioms of structural counterfactuals: composition, effectiveness, and reversibility [Pearl, 2000, Ch.7.3.1].

Proof of Property 1. If X has no spurious path connecting Y in G, the independence relation Y \perp\!
\perp X must hold for any x* [Pearl, 2000, Ch. 11.3.2], which gives:

\[
\text{Cov}_{x*}(X, Y) = \text{Cov}(X, Y_{x*}) = 0. \quad \square
\]

Proof of Property 2. If X has no causal path connecting Y in G, then for any x*, Y_{x*} = Y. This implies:

\[
\text{Cov}_{x*}(X, Y) = \text{Cov}(X, Y - Y) = 0 \quad \square
\]

Proof of Property 3. We first consider the total direct covariance. To prove \text{Cov}_{x*}^{dt}(X, Y) = 0, it suffices to show that for any x, x*, y,

\[
P(x, y_{x*}, w) = P(x, y). \quad (1)
\]

Let \(PA = Pa(Y)\). Conditioned on PA, W, \(P(x, y_{x*}, W)\) can be written as:

\[
P(x, y_{x*}, W) = \sum_{u, pa} P(x, y_{x*}, u | pa, w)P(pa, w)
= \sum_{u, pa} P(x, y_{x*}, u | pa, w)P(pa, w).
\]

The last step holds by the composition axiom: for any \(u\), if \(W(u) = w\), then \(PA(u) = PA_w(u)\). We will next show that for any \(u, w, x*\),

\[
PA_w(u) = PA_{x*, w}(u). \quad (2)
\]

We will prove this statement by contradictions. If Eq. 2 does not hold, there must exist a unblocked causal path from \(X\) to a node in \(PA\) given \(W\) [Galles and Pearl, 1997]. Since \(PA\) are the parents of \(Y\) and \(X \not\in PA(Y)\), we can find an indirect path from \(X\) to \(Y\) given \(W\), which contradicts the definition of mediators. Eq. 2 implies that:

\[
\sum_{w, pa} P(x, y_{x*}, u | pa, w)P(pa, w)
= \sum_{w, pa} P(x, y_{x*}, u | pa_x, w)P(pa, w)
= \sum_{w, pa} P(x, y_{x*, pa}, u | pa_x, w)P(pa, w)
= \sum_{w, pa} P(x, y_{pa, u} | x, x*)P(pa, w)
\]

The last steps hold by the assumption that \(X \not\rightarrow Y\): since all parents of \(Y\) are fixed, the exclusion restrictions rule gives \(Y_{x*, pa, w}(u) = Y_{pa, w}(u)\) for any \(u, x*, pa, w\). Applying Eq. 2 and the composition axiom again gives:

\[
P(x, y_{x*}, w)
= \sum_{w, pa} P(x, y_{pa, u} | x, x*)P(pa, w)
= \sum_{w, pa} P(x, y_{pa, u} | pa, w)P(pa, w)
= \sum_{w, pa} P(x, y | pa, w)P(pa, w) = P(x, y),
\]

which gives Eq. 1. To prove the pure direct covariance \text{Cov}_{x*}^{dp}(X, Y) = 0, it suffices to show that for any \(x, x*, y\),

\[
P(x, y_{x*}) = P(x, y_{W_{x*}}). \quad (3)
\]

By expanding on \(W_{x*}, PA_{W_{x*}}, P(y_{W_{x*}})\) is equal to:

\[
P(x, y_{W_{x*}}) = \sum_{w, pa} P(x, y_{w} | pa_w, x*)P(pa_w, w_{x*})
= \sum_{w, pa} P(x, y_{pa, w} | pa_w, x*)P(pa_w, w_{x*}).
\]
We will first show that if $PA_w = pa$, then $Y_w = Y_{pa,w}$. Since $X \not\rightarrow Y$ in $G$, we have $Y_{pa,w} = Y_{x^*,pa,w}$, which gives:

$$\sum_{w,pa} P(x, y_{pa,u}|pa_w, w_{x^*})P(pa_w, w_{x^*}) = \sum_{w,pa} P(x, y_{x^*,pa,u}|pa_w, w_{x^*})P(pa_w, w_{x^*}).$$

Applying Eq. 2 gives:

$$\sum_{w,pa} P(x, y_{x^*,pa,u}|pa_w, w_{x^*})P(pa_w, w_{x^*}) = \sum_{w,pa} P(x, y_{x^*,pa,w}|pa_w, w_{x^*})P(pa_w, w_{x^*}) = \sum_{w,pa} P(x, y_{x^*,pa,u}|pa_x, w_{x^*})P(pa_x, w_{x^*}).$$

The last step holds by the composition axiom: if $W_{x^*} = w$, $PA_{x^*}(u) = PA_x(u)$. Apply the composition axiom on $Y_{x^*,pa,w}(u)$ again gives:

$$P(x, y_{W_{x^*}}) = \sum_{w,pa} P(x, y_{x^*,pa,w}|pa_x, w_{x^*})P(pa_x, w_{x^*}) = \sum_{w,pa} P(x, y_{x^*,pa,w}|pa_x, w_{x^*})P(pa_x, w_{x^*}) = P(x, y_{x^*}),$$

which proves Eq. 3. \hfill \blackslug

**Proof of Property 4.** Without loss of generality, we suppose $|W| > 0$. To prove the pure indirect covariance $\text{Cov}^{ip}(X, Y) = 0$, it suffices to show that for any $x^*, u$,

$$Y_{W_{x^*}}(u) = Y(u). \quad (4)$$

We will first show that if $|\Pi^I(X, Y)| = 0$, then for any $x^*, u, w$, one of the following equation must hold

$$Y_w(u) = Y(u), \quad (5)$$

$$W_{x^*}(u) = W(u). \quad (6)$$

Suppose that Eq. 5 and 6 both fail, there must exist a unblocked causal path from $X$ to $W$ and a unblocked causal path from $W$ to $Y$. We then find an indirect path from $X$ to $Y$, which is a contradiction. Either Eq. 5 or 6 imply Eq. 4.

To prove the total indirect covariance $\text{Cov}^{it}(X, Y) = 0$, it suffices to show that for any $x^*, u$,

$$Y_{x^*}(u) = Y_{x^*,W(u)}(u). \quad (7)$$

Similarly, we will show that if $|\Pi^I(X, Y)| = 0$, then for any $x^*, u, w$, Eq. 6 and the following equation cannot both be false:

$$Y_{x^*,w}(u) = Y_{x^*}(u), \quad (8)$$

Suppose Eq. 6 and 8 both fail, there must exist an unblocked causal path from $X$ to $W$ and a unblocked causal path from $W$ to $Y$ given $X$. Since removing conditioning nodes only opens up more causal path, we then find an indirect path from $X$ to $Y$, which is a contradiction. Either Eq. 6 or 8 imply Eq. 7. \hfill \blackslug

**Proof of Theorem 1.** By basic mathematical operations, $\text{Cov}(X, Y)$ can be written as:

$$\text{Cov}(X, Y) = \text{Cov}(X, Y - Y_{x^*}) + \text{Cov}(X, Y_{x^*}) = \text{Cov}^c_{x^*}(X, Y) + \text{Cov}^s_{x^*}(X, Y).$$

$\text{Cov}^c_{x^*}(X, Y)$ can be further decomposed as:

$$\text{Cov}^c_{x^*}(X, Y) = \text{Cov}(X, Y) - \text{Cov}(X, Y_{x^*}) = \text{Cov}(X, Y) - \text{Cov}(X, Y_{x^*,W}) + \text{Cov}(X, Y_{x^*,W}) - \text{Cov}(X, Y_{x^*}) = \text{Cov}^{dp}_{x^*}(X, Y) + \text{Cov}^{it}_{x^*}(X, Y).$$

By replacing the term $\text{Cov}(X, Y_{x^*,W})$ in the above equation with $\text{Cov}(X, Y_{W_{x^*}})$, we have:

$$\text{Cov}^c_{x^*}(X, Y) = \text{Cov}(X, Y) - \text{Cov}(X, Y_{x^*}) = \text{Cov}(X, Y) - \text{Cov}(X, Y_{W_{x^*}}) + \text{Cov}(X, Y_{W_{x^*}}) - \text{Cov}(X, Y_{x^*}) = \text{Cov}^{dp}_{x^*}(X, Y) + \text{Cov}^{dt}_{x^*}(X, Y). \hfill \blackslug

**Proof of Corollary 1.** In the linear-standard model, values of $X, Y, Z, W$ are decided by the following functions:

$$z = u_z, \quad x = \alpha_{xz}z + u_x, \quad w = \alpha_{wx}x + \alpha_{wz}z + u_w,$$

$$y = \alpha_{yx} + \alpha_{yz}z + \alpha_{yw}w + uy.$$

Computing $\text{Cov}^s_{x^*}(X, Y)$ gives:

$$\text{Cov}^s_{x^*}(X, Y) = \text{Cov}(X, Y_{x^*}) = \text{Cov}(X, \alpha_{xx}x^* + \alpha_{yz}z + \alpha_{yw}w_{x^*} + u_y) = \alpha_{yz}\text{Cov}(X, Z) + \alpha_{yw}\text{Cov}(X, W_{x^*}) = \alpha_{yz}\text{Cov}(X, Z) + \alpha_{yw}\text{Cov}(X, \alpha_{wx}x^* + \alpha_{wz}Z + u_w) = (\alpha_{yz} + \alpha_{yw}\alpha_{wz})\text{Cov}(X, Z) = (\alpha_{yz} + \alpha_{yw}\alpha_{wz})\alpha_{xz}\text{Cov}(Z, Z) = (\alpha_{yz} + \alpha_{yw}\alpha_{wz})\alpha_{xz}$$
The last step holds since $\text{Cov}(Z, Z) = \text{Var}(Z) = 1$. We can compute $\text{Cov}_{dx}^{dt}(X, Y)$ as:

$$\text{Cov}_{dx}^{dt}(X, Y) = \text{Cov}(X, Y_{W_x} - Y_{x^*}) = \text{Cov}(X, \alpha_{YX} X + \alpha_{YZ} Z + \alpha_{YW} W_{x^*} + U_T) - \text{Cov}(X, \alpha_{YX} x^* + \alpha_{YZ} Z + \alpha_{YW} W_{x^*} + U_T)$$

$$= \alpha_{YX} \text{Cov}(X, X) = \alpha_{YX}.$$

Similarly, $\text{Cov}_{dx}^{dp}(X, Y)$ is equal to:

$$\text{Cov}_{dx}^{dp}(X, Y) = \text{Cov}(X, Y - Y_{x^*, W}) = \text{Cov}(X, \alpha_{YX} X + \alpha_{YZ} Z + \alpha_{YW} W + U_T) - \text{Cov}(X, \alpha_{YX} x^* + \alpha_{YZ} Z + \alpha_{YW} W + U_T)$$

$$= \alpha_{YX} \text{Cov}(X, X) = \alpha_{YX}.$$

Finally, $\text{Cov}_{dx}^{it}(X, Y)$ and $\text{Cov}_{dx}^{ip}(X, Y)$ can be written as:

$$\text{Cov}_{dx}^{it}(X, Y) = \text{Cov}(X, Y_{x^*, W} - Y_{x^*}) = \text{Cov}(X, \alpha_{YX} x^* + \alpha_{YZ} Z + \alpha_{YW} W - W_{x^*})$$

$$= \alpha_{YX} \text{Cov}(X, W) + \alpha_{YW} \text{Cov}(X, X) = \alpha_{YX} \text{Cov}_{dx}(X, X).$$

$$\text{Cov}_{dx}^{ip}(X, Y) = \text{Cov}(X, Y - Y_{W_{x^*}}) = \text{Cov}(X, \alpha_{YX} X + \alpha_{YZ} Z + \alpha_{YW} W + U_T) - \text{Cov}(X, \alpha_{YX} x^* + \alpha_{YZ} Z + \alpha_{YW} W + U_T)$$

$$= \alpha_{YX} \text{Cov}(X, W - W_{x^*}) = \alpha_{YX} \text{Cov}_{dx}(X, X).$$

Proof of Lemma 1. We will prove this lemma by showing a more general case. Let $\pi_0, \pi_1$ be arbitrary sets of causal paths from $X$ such that $\pi_0 \subseteq \pi_1$. Let $\Pi^c(X, Y)_{\pi_0}$ denote the set of open causal paths in $\pi_0$ which connects $Y$ from $X$, i.e., $\Pi^c(X, Y) \cap \pi_0$, so does $\Pi^c(X, Y)_{\pi_1} = \Pi^c(X, Y) \cap \pi_1$. Let $\Pi^c(X, Y)_{\pi_0, \pi_1}$ denote the difference of sets $\Pi^c(X, Y)_{\pi_1} - \Pi^c(X, Y)_{\pi_0}$. We would like to show that if $|\Pi^c(X, Y)_{\pi_0, \pi_1}| = 0$, then for any $x^*, u$,

$$Y_{\pi_0[x^*]}(u) = Y_{\pi_1[x^*]}(u).$$

We will prove this statement by induction on the length $N$ of the longest causal path in $\pi_1$.

Base Case: If $N = 0$, the means that $\pi_0 = \pi_1 = \emptyset$. By definition, $Y_{\pi_0[x^*]}(u) = Y_{\pi_1[x^*]}(u) = Y(u)$, i.e., Eq. 9 holds.

Inductive Case: Assume that for an arbitrary variable $Y \in V$ and sets of causal paths $\pi_0, \pi_1$ where $\pi_0 \subseteq \pi_1$ and the length of all paths in $\pi_1$ is no greater than $N$, $|\Pi^c(X, Y)_{\pi_0, \pi_1}| = 0$ implies that $Y_{\pi_0[x^*]}(u) = Y_{\pi_1[x^*]}(u)$. We use this assumption to prove that for $\pi_1$ with the length of paths no greater than $N + 1$, if $|\Pi^c(X, Y)_{\pi_0, \pi_1}| = 0$, then for any $x^*, u$, Eq. 9 holds. We will prove its contra-positive statement: if Eq. 9 does not hold for some $x^*, u$, then one of the following cases must hold:

1. There exists a variable $U_i \in U$ such that $U_i \in X_{\pi_1 \rightarrow Y}, U_i \not\in X_{\pi_0 \rightarrow Y}$ and its treatment assignment $u_i^*$ is different from its natural value $u_i$.

2. There exists a variable $X_i \in X_{\pi_1 \rightarrow Y}, X_i \not\in X_{\pi_0 \rightarrow Y}$ and $x_i^* \neq X_i \land X_i \not\in X_{\pi_1 \rightarrow Y}$ such that $S_i \subseteq_X \not\in Y_{\pi_0[x^*]}(u) \not\in X_{\pi_1[x^*]}(u)$.

3. There exists a variable $S_i \in (Pa(Y) \cap V) \land Y_{\pi_1[x^*]}(u) \not\in X_{\pi_1[x^*]}(u)$.

We will next show that for each of the above cases, one can find a causal path $g \in \Pi^c(X, Y)_{\pi_0, \pi_1}$. As for Case 1 and 2, it immediately follows that the direct links $U_i \rightarrow Y$ and $X_i \rightarrow Y$ construct a causal paths $g \in \Pi^c(X, Y)_{\pi_0, \pi_1}$, respectively.

As for Case 3, by the assumption, $S_i \subseteq_X \not\in Y_{\pi_0[x^*]}(u) \not\in X_{\pi_1[x^*]}(u)$ implies that there exists a path $g_s \in \Pi^c(X, Y)_{\pi_0, \pi_1}$ such that $S_i \subseteq_X \not\in Y_{\pi_0[x^*]}(u) \not\in X_{\pi_1[x^*]}(u)$. We can then construct a causal path $g$ from $X$ to $Y$ by appending the edge $S_i \rightarrow Y$ to $g_s$. By the definition of the funnel operator $\not\in X_{\pi_1[x^*]}(u)$, we must have $g \in \Pi^c(X, Y)_{\pi_0, \pi_1}$.

To prove Lem. 1, let $\pi_0 = \pi(g), \pi_1 = \pi(g) \cup \{g\}$. If $g \not\in \Pi^c(X, Y)$, then $|\Pi^c(X, Y)_{\pi_0, \pi_1}| = 0$. This implies that Eq. 9 holds, i.e.,

$$Y_{\pi(g)[x^*]}(u) = Y_{\pi(g) \cup \{g\}[x^*]}(u).$$

Proof of Property 5. Lem. 1 implies that for any $x^*, u$,

$$g \not\in \Pi^c(X, Y) \Rightarrow Y_{\pi(g)[x^*]}(u) = Y_{\pi(g) \cup \{g\}[x^*]}(u).$$

This gives

$$\text{Cov}_{g}^{ip}(X, Y_{\pi}) = \text{Cov}(X, Y_{\pi(g)[x^*]} - Y_{\pi(g) \cup \{g\}[x^*]}(u) = 0.$$ 

Proof of Theorem 2. By definition,
Following the order $\mathcal{L}^*$, let $g_{[1,n]}$ denote $\Pi^c(X,Y)$. In the model associated with the $g_{[1,n]}$-specific counterfactual $Y_{g_{[1,n]}}^{x^*}(u)$, all variables are under the influence of the intervention $X = x^*$, i.e.,

$$Y_{g_{[1,n]}}^{x^*}(u) = Y_{x^*}(u).$$

Thus, the causal covariance $\text{Cov}^c_{x^*}(X,Y)$ is equal to:

$$\text{Cov}^c_{x^*}(X,Y) = \text{Cov}(X,Y) - Y_{x^*}(u) - Y_{x^*}(u) + Y_{x^*}(u).$$

Reorganizing the above equation gives

$$\text{Cov}^c_{x^*}(X,Y) = \sum_{g \in \Pi^c(X,Y)} \text{Cov}^c_{g|x^*}(X,Y) \mathcal{L}_x^c.$$ 

**Lemma 3.** For a semi-Markovian model $M$, let an order over $U^*$ be $U^*_{\Pi^c} : U_1 < \cdots < U_n$. For any $x^*$, $\text{Cov}^c_{x^*}(X,Y)$ can be expressed as:

$$\sum_{i=1}^n \text{Cov}(X_{U^*_i}, U_{U^*_i} - Y_{x^*,U_{U^*_i}} - Y_{x^*,U_{U^*_i}}).$$

**Proof.** Following the order $\mathcal{L}^*$, let $U_{[1,n]}$ denote $U^*$. We will use $I\{-\}$ to represent an indicator function. Since the exogenous variables $U_i, U_{[1,i]} \cup U_{[i,1]}$ explain all the uncertainties of variables $X_{U_{[1,i]}}$ and $Y_{x^*,U_{[1,i]}}$, we must have:

$$P(x_{U_{[1,i]}}, y_{x^*,U_{[1,i]}}) = \sum_{u_i} \sum_{u_{[1,i]}} \sum_{u_{[i,1]}} I\{X_{u_{[1,i]}} = x\} I\{Y_{x^*,U_{[1,i]}}(u) = y\} \cdot P(u_i)P(u_{[1,i]})P(u_{[i,1]})$$

Let $U_0$ denote the set of exogenous variables which affect $X$ other than $U^*$. Similarly, we define $U_Y$ for $Y_{x^*}$. Since $U^*$ is the maximal set of exogenous variables that affects both $X$ and $Y_{x^*}$, we must have $U_X \cap U_Y = \emptyset$. The above equation can thus be written as:

$$P(x_{U_{[1,i]}}, y_{x^*,U_{[1,i]}}) = \sum_{u_i} \sum_{u_{[1,i]}} \sum_{u_{[i,1]}} I\{X_{u_{[1,i]}} = x\} I\{Y_{x^*,U_{[1,i]}} = y\} P(u_i)P(u_{[1,i]})P(u_{[i,1]})$$

By Lem. 1, Eq. 10 implies that for any $x^*, u, u_{[1]}^*$:

$$X_{\pi_1(u_{[1]})} = X_{\pi_1(u_{[1]})}.$$ 

Eq. 11 implies that in the submodel $M_{x^*}$ with an associated causal diagram $G_{x^*}$ where all incoming edges of $X$ are removed, $g_\tau \in \Pi^c(U_i, Y) \in \mathcal{G}_{x^*}$. By the definition of the submodel $M_{x^*}$ [Pearl, 2000, Ch. 7.1], the counterfactual $Y_{x^*}$ is the outcome $Y$ in the submodel $M_{x^*}$. By Lem. 1, we then have, for any $x^*, u, u_{[1]}^*$:

$$Y_{x^*,\pi_1(u_{[1]})} = Y_{x^*,\pi_1(u_{[1]})}.$$ 

The $\tau$-specific spurious covariance thus equates to:

$$\text{Cov}^s_{x^*}(X,Y) = \text{Cov}(X_{\pi} - X_{\pi \cup (g_\tau)}, Y_{x^*,\pi} - Y_{x^*,\pi \cup (g_\tau)}) = 0.$$

To prove Thm. 3, we first introduce two lemmas.
Replacing distribution since it suffices to prove that for any \( X, Y \), we will next show that

\[
\text{Cov}(X_{U^i}, Y_{U^r}) = \text{Cov}(X_{U^i}, Y_{U^r}) - \text{Cov}(X_{U^i}, Y_{U^r}) - \text{Cov}(X_{U^i}, Y_{U^r}).
\]

(14)

By the basic mathematical operations of covariance,

\[
\text{Cov}(X_{U^i}, Y_{U^r}) = \text{Cov}(X_{U^i}, Y_{U^r}) - \text{Cov}(X_{U^i}, Y_{U^r}) + \text{Cov}(X_{U^i}, Y_{U^r}) - \text{Cov}(X_{U^i}, Y_{U^r}).
\]

It suffices to prove that for any \( x, y \),

\[
P(x_{U^i}, y_{U^r}) = P(x_{U^i}, y_{U^r}) + P(x_{U^i}, y_{U^r}) - P(x_{U^i}, y_{U^r}).
\]

(15)

Let us first consider Eq. 15. From Eq. 12, the distributions \( P(x_{U^i}, y_{U^r}) \) and \( P(x_{U^i}, y_{U^r}) \) can be written as:

\[
P(x_{U^i}, y_{U^r}) = \sum_{u_{i[n+1]} u_{i[n+1]}} P(X(u_{i[n+1]}) = x)
\]

\[
\cdot P(Y_{U^r}(u_{i[n+1]}))
\]

\[
\cdot P(u_{i[n+1]} = y)P(u_{i[n+1]})
\]

\[
P(x_{U^i}, y_{U^r}) = \sum_{u_{i[n+1]} u_{i[n+1]}} P(X(u_{i[n+1]}) = x)
\]

\[
\cdot P(Y_{U^r}(u_{i[n+1]}), y_{U^r})P(u_{i[n+1]})
\]

Since \( U \) and \( U^r \) are i.i.d. draws from the exogenous distribution \( P(u) \), we have for \( u_i = u_i^r \), \( P(u) = P(u_i) \). Replacing \( u_i^r \) with \( u_i \), in the above equations gives Eq. 15. Similarly, we can prove Eq. 16. Eqs. 15-16 together prove Eq. 14.

\[\square\]

**Lemma 4.** For a semi-Markovian model \( M \), let \( T^*(X, Y; U^r) \) denote the set of spurious treks from \( X \) to \( Y \) with a common source \( U^r \). Let \( L_n = \{L_n^l, \{L_n^l, L_n^r\} \leq \leq |U^r| \} \) be an order over spurious treks \( T^*(X, Y) \). For any \( x^* \), the following non-parametric relationships hold:

\[
\text{Cov}(X, Y; U^r) = \sum_{r \in T^*(X, Y; U^r)} \text{Cov}(X, Y; L_n^r)
\]

**Proof.** Let \( (\pi_l, \pi_r) \) denote the pair

\[
(\Pi(U_{1,1-1}), \Pi(U_{1,1-1}, Y|X)).
\]

Following the order \( L_n \), let \( g^1_{1,n} = \Pi(U_i, X) \) and \( g^2_{1,m} = \Pi(U_i, Y|X) \). Since the intervention \( do(U^r) \) assigns a randomized treatment \( U^r(U^r) \) to all causal paths in \( g^1_{1,n} \) (\( g^2_{1,m} \)). The term \( \text{Cov}(X, Y; U^r) \) can thus be written as:

\[
\text{Cov}(X, Y; U^r) = \text{Cov}(X, Y; U^r) - \text{Cov}(X, Y; U^r) - \text{Cov}(X, Y; U^r).
\]

The above equation can be decomposed over causal paths in \( g^1_{1,n} \):

\[
\text{Cov}(X, Y; U^r) = \text{Cov}(X, Y; U^r) - \text{Cov}(X, Y; U^r) - \text{Cov}(X, Y; U^r).
\]

(16)

\[\square\]
\[ + \text{Cov}(X_{\pi_t \cup g'_t[i,j-1]} - X_{\pi_t \cup g'_t[i,j]}),
\]
\[ Y_{x^*, \pi_t \cup g'_t[i,j]} - Y_{x^*, \pi_t \cup g'_t[i,j-1]} \]
\[ + \text{Cov}(X_{\pi_t \cup g'_t[i,j-1]} - X_{\pi_t \cup g'_t[i,j]}),
\]
\[ Y_{x^*, \pi_t \cup g'_t[i,j]} - Y_{x^*, \pi_t \cup g'_t[i,j-1]} \]
\[ : \]
\[ = \sum_{k=1}^{m} \text{Cov}(X_{\pi_t \cup g'_t[i,j-1]} - X_{\pi_t \cup g'_t[i,j]}),
\]
\[ Y_{x^*, \pi_t \cup g'_t[i,j]} - Y_{x^*, \pi_t \cup g'_t[i,k-1]} \cdot \text{Cov}(\cdot) \]

Together, we can obtain

\[ \text{Cov}_{x^*}(X, Y)_{U_t} = \sum_{\tau \in \mathcal{T}^s(X, Y; U_t)} \text{Cov}_{\tau}(X, Y)_{U_t} \]

where \( \tau_{j,k} = (g'_t[j], g'_t[k]) \). Reorganizing the above equation gives:

\[ \text{Cov}_{x^*}(X, Y)_{U_t} = \sum_{\tau \in \mathcal{T}^s(X, Y; U_t)} \text{Cov}_{\tau}(X, Y)_{U_t} \]

We are now ready to prove Thm. 3

**Proof of Theorem 3.** By Lem. 3 and 4, we have:

\[ \text{Cov}_{\tau}(X, Y)_{U_t} = \sum_{\pi_t \cup g'_t[i,j]} \text{Cov}_{\tau}(X, Y)_{U_t} \]

Reorganizing the above equation gives:

\[ \text{Cov}_{\tau}(X, Y)_{U_t} = \sum_{\tau \in \mathcal{T}^s(X, Y; U_t)} \text{Cov}_{\tau}(X, Y)_{U_t} \]

**Proof of Lemma 2. Existence.** We will prove the existence of \( V_t \) by proving a stronger statement: in a semi-Markovian model \( M \), for any non-simple path of the form \((g_t, g_r)\) where \( g_t, g_r \) share a common source \( V_t \) and have sink \( X \) and \( Y \) respectively, there always exists a most distant recurring node \( V_t \) such that \( V_t \) is the only shared common node among subpaths \( g_t(V_t, X) \) and \( g_r(V_t, Y) \). We will use this assumption to prove that for all non-simple path of the form \((g_t, g_r)\) with \( N \) recurring nodes, the most distant recurring \( V_t \) also exists. For a non-simple path \((g_t, g_r)\), we find the next recurring node \( V'_t \) of \( g_t, g_r \) other than the common source \( V_t \). The subpaths \((g_t(V'_t, X), g_r(V'_t, Y))\) forms a non-simple path with \( N \) recurring nodes. By the assumption, for the non-simple path \((g_t(V'_t, X), g_r(V'_t, Y))\), there exists a most distant recurring node \( V'_t \) such that \( V_t \) is the only node shared among subpaths \( g_t(V_t, X), g_r(V_t, Y) \).

We will show that the most distant recurring node \( V_t \) of \((g_t(V'_t, X), g_r(V'_t, Y))\) is also satisfied for \((g_t, g_r)\). Suppose \( V_t \) is not a most distant recurring node of \((g_t, g_r)\), this means that the subpaths \( g_t(V_t, X), g_r(V_t, Y) \) share another common node other than \( V_t \), which contradicts our assumption.

**Uniqueness.** We will prove this lemma by contradictions. Suppose there are two distinct nodes \( V_t^0, V_t^1 \) for a trek \( \tau = (g_t, g_r) \) such that for \( i = 0,1 \), \( V_t^i \) is the only node shared among subpaths \( g_t(V_t^i, X) \) and \( g_r(V_t^i, Y) \). \( V_t^0, V_t^1 \) must satisfy one of the following cases.

1. There exists a causal path from \( V_t^0 \) to \( V_t^1 \) in \( g_t \), denoted by \((V_t^0 \rightarrow V_t^1)_{g_t}\), and a causal path from \( V_t^0 \) to \( V_t^1 \) in \( g_r \), denoted by \((V_t^0 \rightarrow V_t^1)_{g_r}\).
2. \((V_t^0 \rightarrow V_t^1)_{g_t}\) and \((V_t^1 \rightarrow V_t^0)_{g_r}\).
3. \((V_t^0 \rightarrow V_t^1)_{g_t}\) and \((V_t^1 \rightarrow V_t^0)_{g_r}\).
4. \((V_t^0 \rightarrow V_t^1)_{g_t}\) and \((V_t^0 \rightarrow V_t^1)_{g_r}\).

For Case. 1, we must have that \( V_t^1 \) is also a common node shared among the subpaths \( g_t(V_t^0, X) \) and \( g_r(V_t^0, Y) \), which contradicts our assumptions. Similarly, Case. 2 lead to an contradiction, as \( V_t^0 \) is also a common node shared among the subpaths \( g_t(V_t^1, X) \) and \( g_r(V_t^1, Y) \).

For Case. 3, if exists a causal path from \( V_t^0 \) to \( V_t^1 \) and a causal path from \( V_t^1 \) to \( V_t^0 \), the causal diagram \( G \) of the semi-Markovian model \( M \) is not a DAG, which is a contradiction. Similarly, Case. 4 contradicts the assumption that \( G \) is a DAG. Since Cases. 1-4 all lead to contradictions, the most distant recurring node \( V_t \) is unique for each trek \( \tau \in \mathcal{T}^s(X, Y) \).

**Proof of Property 7.** For a spurious path \( l = (g_t, g_r) \) with the common source \( V_t \), if \( l \notin \Pi^*(X, Y) \), then one
Specifically, distributions of the following conditions must hold:

\[ g_t \notin \Pi^r(V_t, X), \quad g_r \notin \Pi^r(V_t, Y|X). \]

For each \( \tau \in T^s(l) \), \( g_t, g_r \) are both its subpaths. This implies that from the above conditions, we must have \( \tau \notin T^s(X, Y) \). By Prop. 6, we have

\[
\text{Cov}^s_{[x*]}(X, Y)_\pi = \sum_{\tau \in T^s(Y)} \text{Cov}^s_{[x*]}(X, Y)_\tau = 0. \]

**Proof of Theorem 4.** Thm. 3 implies

\[
\text{Cov}^s_{x*}(X, Y) = \sum_{\tau \in T^s(X, Y)} \text{Cov}^s_{[x*]}(X, Y)_\tau \quad \text{(17)}
\]

Since the mapping \( f : T^s(X, Y) \rightarrow \Pi^s(X, Y) \) is a surjective function, \( \{T^s(l)\} \in \Pi^s(X, Y) \) is a partition over the set \( T^s(X, Y) \). Eq. 17 could be written as:

\[
\text{Cov}^s_{x*}(X, Y) = \sum_{\ell \in \Pi^s(X, Y)} \sum_{\tau \in T^s(l)} \text{Cov}^s_{[x*]}(X, Y)_\tau = \sum_{\ell \in \Pi^s(X, Y)} \text{Cov}^s_{[x*]}(X, Y)_\ell. \]

**Proof of Theorem 5.** By Thm. 1, we have

\[
\text{Cov}(X, Y) = \text{Cov}^s_{x*}(X, Y) + \text{Cov}^s_{x*}(X, Y). \]

Applying Thm. 2 and 5 to the above equation gives

\[
\text{Cov}(X, Y) = \sum_{\ell \in \Pi^s(X, Y)} \text{Cov}^s_{[x*]}(X, Y)_\ell \quad \text{and} \quad \sum_{\ell \in \Pi^s(X, Y)} \text{Cov}^s_{[x*]}(X, Y)_\ell. \]

We will next prove Thm. 6. Recall in the standard model of Fig. 1(a), \( X \) and \( Y \) are connected with causal paths \( l_1 : X \rightarrow Y, l_2 : X \rightarrow W \rightarrow Y \) and spurious paths \( l_3 : X \leftarrow Z \rightarrow Y \) and \( l_4 : X \leftarrow Z \rightarrow W \rightarrow Y \). \( U^s = \{U_Z\} \) affects the treatment \( X \) through a causal path \( g_t = U_Z \rightarrow Z \rightarrow X \), and the outcome \( Y \) through causal paths \( g_r_1 = U_Z \rightarrow Z \rightarrow Y \) and \( g_r_2 = U_Z \rightarrow Z \rightarrow W \rightarrow Y \). To prove Thm. 7, we will introduce following lemmas.

**Lemma 5.** In the standard model (Fig. 1(a)), for an order \( t^c : l_1 < l_2 \), the path-specific decomposition of the causal covariance \( \text{Cov}^s_{x*}(X, Y) \) (Thm. 2) are identifiable if \( P(x, y_{x*}, w) \) and \( P(x, y_{x*}) \) are identifiable. Specifically, distributions \( P(x, y_{x*}, w) \) and \( P(x, y_{x*}) \) can be estimated from the observational distribution \( P(x, y, z, w) \) as following:

\[
P(x, y_{x*}) = \sum_{z, w} P(y_{x*}|z, w)P(w|x, z)P(x, z),
\]

\[
P(x, y_{x*}, w) = \sum_{z, w} P(y_{x*}|z, w)P(x, z, w).
\]

**Proof.** By Thm. 2, the causal covariance \( \text{Cov}^s_{x*}(X, Y) \) equates to

\[
\text{Cov}^s_{x*}(X, Y) = \text{Cov}^s_{[x*]}(X, Y) \mathbb{L}_x + \text{Cov}^s_{[x*]}(X, Y) \mathbb{L}_z.
\]

We will show that each quantity on the right-hand side of the above equation is identifiable from \( P(x, y, z, w) \). For the order \( t^c : l_1 < l_2 \),

\[
\text{Cov}^s_{[x*]}(X, Y) \mathbb{L}_x = \text{Cov}^s_{[x*]}(X, Y) \mathbb{L}_x
\]

\[
= \text{Cov}(X, Y - Y_{x*}, W)
\]

\[
= \text{Cov}(X, Y) - \text{Cov}(X, Y_{x*}, W).
\]

\[
\text{Cov}^s_{[x*]}(X, Y) \mathbb{L}_z = \text{Cov}^s_{[x*]}(X, Y) \mathbb{L}_z
\]

\[
= \text{Cov}(X, Y_{x*}, W - Y_{x*})
\]

\[
= \text{Cov}(X, Y_{x*}, W) - \text{Cov}(X, Y_{x*}).
\]

It suffices to show that distributions \( P(x, y_{x*}, w) \) and \( P(x, y_{x*}) \) are identifiable. By expanding on \( Z, W_{x*}, P(x, y_{x*}) \) can be written as:

\[
P(x, y_{x*})
\]

\[
= \sum_{z, w} P(y_{x*}|x, z, w)P(w_{x*}|x, z)P(x, z)
\]

\[
= \sum_{z, w} P(y_{x*}, z, w|x, z, w)P(w_{x*}|x, z)P(x, z).
\]

The last step holds due to the following reasons: (1) by the exclusion restrictions rule, since \( Z \) has no parent node in the model of Fig. 1(a), \( Z = Z_{x*} \) for any \( x^* \); (2) by the composition axiom, we have:

\[
Z = z \Rightarrow X = X_{z},
\]

\[
Z_{x*} = z \Rightarrow W_{x*} = W_{x*, z},
\]

\[
Z_{x*} = z, W_{x*} = w \Rightarrow Y_{x*} = Y_{x*, z, w}.
\]

By the independence exclusions rule, for any \( x^*, x, z, w, \)

\[
W_{x*, z} \perp X_{z}, Z, \quad \text{and} \quad Y_{x*, z, w} \perp X_{z}, Z, W_{x*, z}.
\]

We thus have:

\[
\sum_{z, w} P(y_{x*}, z, w|x, z, w)P(w_{x*}|x, z)P(x, z)
\]

\[
= \sum_{z, w} P(y_{x*}, z, w|w_{x*}, z)P(w_{x*}|x, z)P(x, z)
\]
Since the standard model is Markovian,
\[ P(w_{x^*,z}) = P(w|x^*), \quad (22) \]
\[ P(y_{x^*,z,w}) = P(y|x^*, z, w). \quad (23) \]

Thus,
\[ P(x, y_{x^*}) = \sum_{z,w} P(y|x^*, z, w) P(w|x^*, z) P(x, z). \]

By expanding on \( Z, W, P(x, y_{x^*}, W) \) can be written as:
\[ P(x, y_{x^*}, W) = \sum_{z,w} P(y_{x^*, w}|x, z, w) P(x, z, w) \]
\[ = \sum_{z,w} P(y_{x^*, w}|x, z, w, w_{x^*, z}) P(x, z, w) \]

The last step holds due to following reasons: (1) By the composition axiom, \( W = W_{x,z} \) if \( X = x, Z = z \); (2) By the exclusion restrictions rules, \( Z = Z_{x^*, w} \) if \( Z \) has no parent node. Applying the composition axiom again gives:
\[ Z_{x^*, w} = z \Rightarrow Y_{x^*, w} = Y_{x^*, z,w}. \]

We thus have:
\[ \sum_{z,w} P(y_{x^*, w}|x, z_{x^*, w}, w, w_{x^*, z}) P(x, z, w) \]
\[ = \sum_{z,w} P(y_{x^*, w}|x, z_{x^*, w}, w_{x^*, z}) P(x, z, w) \]

The independence relation 21 gives:
\[ P(x, y_{x^*}, W) = \sum_{z,w} P(y_{x^*, z,w}|x, z_{x^*, w}, w, w_{x^*, z}) P(x, z, w) \]
\[ = \sum_{z,w} P(y_{x^*, z,w}|x, z_{x^*, w}) P(x, z, w) \]
\[ = \sum_{z,w} P(y|x^*, z, w) P(x, z, w). \]

The last step holds by Eq. 23.

\[ \square \]

**Lemma 6.** In the standard model (Fig. 1(a)), for a order \( L^s \) where \( L^s_r : g_{r_1} < g_{r_2} \), the path-specific decomposition of the spurious covariance \( \text{Cov}^s_{x^*}(X,Y) \) (Thm. 5) is identifiable if \( P(x, y_{x^*}) \), \( P(x, y_{x^*, W_{x^*, z}}, U_{x^*, z}) \) are identifiable. Specifically, distributions \( P(x, y_{x^*}) \), \( P(x, y_{x^*, W_{x^*, z}}, U_{x^*, z}) \) are identifiable can be estimated from the observational distribution \( P(x, y, z, w) \) as following:
\[ P(x, y_{x^*}) = \sum_{z,w} P(y|x^*, z, w) P(w|x^*, z) P(x, z), \]
\[ P(x, y_{x^*, W_{x^*, z}}, U_{x^*, z}) = \sum_{z,w,z'} P(y|x^*, w, z) P(w|x^*, z') \times P(x, z') P(z) \]

**Proof.** By Thm. 5, the spurious covariance \( \text{Cov}^s_{x^*}(X,Y) \) equates to
\[ \text{Cov}^s_{x^*}(X,Y) = \text{Cov}^t_{x^*}(X,Y) \xi^s_2 + \text{Cov}^t_{y^*}(X,Y) \xi^s_2. \]

We will next show that each quantity on the right-hand side of the above equation is identifiable from \( P(x, y, z, w) \). In the standard model, Considering the order \( L^s \) of \( g_{r_1} < g_{r_2} \),
\[ \text{Cov}^s_{x^*}(X,Y) = \text{Cov}^t_{x^*}(X,Y) \xi^s_2 + \text{Cov}^t_{y^*}(X,Y) \xi^s_2. \]

The last step holds since \( U_{x^*, z} \) is an independent counterfactual variable: the variable \( X \) is function over \( U_X, U_{x^*}; \) the exogenous variables \( U_X, U_{x^*} \) are independent of all the other variables in the domain. Similarly,
\[ \text{Cov}^s_{y^*}(X,Y) \xi^s_2 = \text{Cov}(X - X_{U^*_x}, Y_{x^*} - Y_{x^*, g_{r_1}(1,2)}) = \text{Cov}(X - X_{U^*_x}, Y_{x^*} - Y_{x^*, g_{r_1}(1,2)}) \]
\[ = \text{Cov}(X, Y_{x^*}, W_{x^*, z}, U_{x^*, z}) - Y_{x^*, U_{x^*, z}} \]
\[ = \text{Cov}(X, Y_{x^*}, W_{x^*, z}) \]

The last two steps holds since \( X_{U^*_x} \) and \( Y_{x^*, U_{x^*, z}} \) are independent counterfactual variables. It will suffice to show that the distributions \( P(x, y_{x^*}) \), \( P(x, y_{x^*, W_{x^*, z}}, U_{x^*, z}) \) are identifiable. \( P(x, y_{x^*}) \) can be identified using Lem. 5. By conditioning on \( U_Z \), \( P(x, y_{x^*, W_{x^*, z}}, U_{x^*, z}) \) can be written as:
\[ P(x, y_{x^*, W_{x^*, z}}, U_{x^*, z}) = \sum_{u_Z} P(x, y_{x^*, W_{x^*, z}}, U_{x^*, z}|u_Z) P(u_Z) \]

With \( U_Z \) fixed, variables \( X \) and \( Y_{x^*, W_{x^*, z}} \) are functions of the exogenous variable \( U \), which is independent of \( U_Z \). We thus have the following independence relation
\[ U_Z \perp X, Y_{x^*, W_{x^*, z}} \]

which gives:
\[ P(x, y_{x^*, W_{x^*, z}}, U_{x^*, z}) = \sum_{u_Z} P(x, y_{x^*, W_{x^*, z}}, U_{x^*, z}) P(u_Z) \]
\[ (26) \]

By expanding on \( Z, U_{x^*, z}, W_{x^*}, P(x, y_{x^*, W_{x^*, z}}, U_{x^*, z}) \) can be written as:
\[ P(x, y_{x^*, W_{x^*, z}}, U_{x^*, z}) = \sum_{z,w,z'} P(x, y_{x^*, w, z}, w_{x^*, z'}) \]
\[ \times P(x, z') P(z) \]
Since the function $f_Z$ takes only $U_{Z}$ as an argument, the variables $Z_{\vec{z}}$ are deterministic, i.e.,

$$P(x, y_{x^*, w, z}, w_{x^*, z', z'}, z_{u_Z}) = P(x, y_{x^*, w, z}, w_{x^*, z'}) I\{Z_{u_Z} = z\}$$

where $I\{\cdot\}$ is an indicator function. The above equation, together with Eq. 26, gives:

$$P(x, y_{x^*, w, z}, w_{x^*, z'}) = \sum_{z, z', w} P(x, y_{x^*, w, z}, w_{x^*, z'}) I\{Z_{u_Z} = z\} P(u_Z^z)$$

$$= \sum_{z, z', w} P(x, y_{x^*, w, z}, w_{x^*, z'}) P(z).$$

(27)

By the composition axiom and the exclusion restrictions rule [Pearl, 2000, Ch. 7.3], in the model of Fig. 1(a), for any $z, x^*$,

$$Z = z \Rightarrow X = X_z,$$

$$Z = Z_{x^*},$$

$$Z_{x^*} = z \Rightarrow W_{x^*} = W_{x^*, z},$$

(28)

The above relations imply that:

$$P(x, y_{x^*, w, z}, w_{x^*, z'}) = P(x, y_{x^*, w, z}, w_{x^*, z'}, z')$$

The independence restrictions rule [Pearl, 2000, Ch. 7.3] implies that in the model of Fig. 1(a), counterfactuals $X_{z'}, Y_{x^*, w, z}, W_{x^*, z'}, Z$ are mutually independent. We thus obtain:

$$P(x_{z'}, y_{x^*, w, z}, w_{x^*, z'}, z') = P(x_{z'}) P(y_{x^*, w, z}) P(w_{x^*, z'}) P(z')$$

(29)

Since the standard model is Markovian,

$$P(x_{z'}) = P(x|z'),$$

$$P(y_{x^*, w, z}) = P(y|x^*, w, z),$$

$$P(w_{x^*, z'}) = P(w|x^*, z'),$$

(30)

Eqs. 27, 29 and 30 together give:

$$P(x, y_{x^*, w, z}, z_{u_Z}) = \sum_{z, z', w} P(x|z') P(y|x, w, z) P(w|x^*, z') P(z') P(z).$$

$$= \sum_{z, z', w} P(y|x, w, z) P(w|x^*, z') P(x, z') P(z).$$

We are now ready to prove Thm. 6.

Proof of Theorem 6. Recall the target path-specific decomposition of $\text{Cov}(X, Y)$ is:

$$\text{Cov}(X, Y - Y_{x^*, w}) + \text{Cov}(X, Y_{x^*, w} - Y_{x^*})$$

$$+ \text{Cov}(X - X_{U_Z^z}, Y_{x^*, w} - Y_{x^*}, Z_{u_Z})$$

$$+ \text{Cov}(X - X_{U_Z^z}, Y_{x^*, w}, Z_{u_Z} - Y_{x^*, U_Z^z}).$$

This decomposition is induced by the order $L^c : l_1 < l_2$ and $L^s$ where $L^c_r : g_{r_1} < g_{r_2}$. Thm. 6 immediately follows from Lems. 5 and 6.

We next consider the identification of other decompositions of $\text{Cov}(X, Y)$ in the model of Fig. 1(a). Indeed, one could show that the decomposition of $\text{Cov}(X, Y)$ (Thm. 5) are always identifiable in the standard model regardless of the order $L^c$ and $L^s$.

Lemma 7. In the standard model (Fig. 1(a)), for an order $L^c : l_2 < l_1$, the path-specific decomposition of the causal covariance $\text{Cov}^c_{l_2}(X, Y)$ (Thm. 2) are identifiable if $P(x, y_{x^*})$ and $P(x, y_{W_{x^*}})$ are identifiable. Specifically, distributions $P(x, y_{x^*})$ and $P(x, y_{W_{x^*}})$ can be estimated from the observational distribution $P(x, y, w)$ as follows:

$$P(x, y_{x^*}) = \sum_{z, w} P(y|x^*, z, w) P(w|x^*, z) P(x, z),$$

$$P(x, y_{W_{x^*}}) = \sum_{z, w} P(y|x, z, w) P(w|x^*, z) P(x, z).$$

Proof. Consider the order $L^c : l_2 < l_1$. The path-specific causal covariance of $l_1, l_2$ are equal to:

$$\text{Cov}^c_{l_1[x^*]}(X, Y)_{L^c_{x^*}} = \text{Cov}^c_{l_1}(X, Y)$$

$$= \text{Cov}(X, Y_{W_{x^*}} - Y_{x^*})$$

$$= \text{Cov}(X, Y_{W_{x^*}}) - \text{Cov}(X, Y_{x^*}),$$

$$\text{Cov}^c_{l_2[x^*]}(X, Y)_{L^c_{x^*}} = \text{Cov}^c_{l_2}(X, Y)$$

$$= \text{Cov}(X, Y - Y_{W_{x^*}})$$

$$= \text{Cov}(X, Y) - \text{Cov}(X, Y_{W_{x^*}}).$$

It suffices to show that distributions $P(x, y_{W_{x^*}})$ and $P(x, y_{x^*})$ are identifiable. $P(x, y_{x^*})$ can be identified using Lem. 5. By expanding on $Z, W_{x^*}, P(x, y_{W_{x^*}})$ can be written as:

$$P(x, y_{W_{x^*}}) = \sum_{z, w} P(y_w|x, z, w_{x^*}) P(w_{x^*}|x, z) P(x, z)$$

$$= \sum_{z, w} P(y_w|x, z, w_{x^*}) P(w_{x^*}|x, z) P(x, z).$$

In the last step, since $Z$ is a non-descendant node of $X, W$ and $X$ is a non-descendant node of $W$, we have $Z = Z_{x^*} = Z_w$ and $X = X_w$. By the composition axiom,

$$Z = z \Rightarrow X = X_z,$$

$$Z_{x^*} = z \Rightarrow W_{x^*} = W_{x^*, z},$$

$$X_w = x, Z_w = z \Rightarrow Y_w = Y_{x, z, w}. $$
which gives:
\[
\sum_{z,w} P(y_{w} | x_{w}, z_{w}, w_{x^*}) P(w_{x^*} | x_{x^*}) P(x, z)
\]
\[
= \sum_{z,w} P(y_{x,z,w} | x_{w}, z_{w}, w_{x^*}) P(w_{x^*,z} | x_{x^*}, z_{z^*}) P(x, z)
\]
\[
= \sum_{z,w} P(y_{x,z,w} | x_{z}, z_{z^*}, w_{x^*}) P(w_{x^*,z} | x_{z}, z_{z^*}) P(x, z)
\]
\[
= \sum_{x,z,w} P(y_{x,z,w} | x_{z}, z_{z^*}, w_{x^*}) P(w_{x^*,z} | x_{z}, z_{z^*}) P(x, z)
\]
\[
= \sum_{x,z,w} P(y_{x,z,w} | x_{z}, z_{z^*}, w_{x^*}) P(w_{x^*,z} | x_{z}, z_{z^*}) P(x, z)
\]

The last step holds since \( Z = Z_{x^*} \), and \( Z_{x^*} = z \Rightarrow W_{x^*} = W_{x^*,z} \). Applying Eqs. 20 and 21 gives:
\[
P(x, y_{w^*}) = \sum_{z,w} P(y_{x,z,w}) P(w_{x^*,z}) P(x, z)
\]
\[
= \sum_{z,w} P(y_{x,z,w} | x_{z}, z_{z^*}) P(w_{x^*,z} | x_{z}, z_{z^*}) P(x, z).
\]

The last step holds by Eqs. 22 and 23.

**Lemma 8.** In the standard model (Fig. 1(a)), for a order \( \mathcal{L}^* \) where \( \mathcal{L}^* \) : \( g_{r_2} < g_{r_1} \), the path-specific decomposition of the spurious covariance \( \text{Cov}^*_{s} (X, Y) \) (Thm. 5) is identifiable if \( P(x, y_{x^*}), P(x, y_{x^*}, W_{x^*,u^*_Z}) \) are identifiable. Specifically, distributions \( P(x, y_{x^*}), P(x, y_{x^*}, W_{x^*,u^*_Z}) \) are identifiable can be estimated from the observational distribution \( P(x, y, z, w) \) as following:

\[
P(x, y_{y^*}) = \sum_{z,w} P(y_{x,z,w}) P(w_{x^*,z}) P(x, z),
\]
\[
P(x, y_{y^*,v^*,u^*_Z}) = \sum_{z,z',w} P(y_{x,z,w}, z') P(w_{x^*,z}) P(x, z).
\]

**Proof.** Considering the order \( \mathcal{L}^* \) where \( \mathcal{L}^* \) : \( g_{r_2} < g_{r_1} \),
\[
\text{Cov}^*_{s} (X, Y)_{\mathcal{L}^*} = \text{Cov}(X - X_{U^*_Z}, Y_{y^*} - Y_{y^*,g_{r_2} = 1, z^*})
\]
\[
= \text{Cov}(X - X_{U^*_Z}, Y_{y^*,v^*,u^*_Z} - Y_{y^*,u^*_Z})
\]
\[
= \text{Cov}(X, Y_{y^*,v^*,u^*_Z} - Y_{y^*,u^*_Z})
\]
\[
= \text{Cov}(X, Y_{y^*,v^*,u^*_Z} - Y_{y^*,u^*_Z}).
\]

Similarly,
\[
\text{Cov}^*_{s} (X, Y)_{\mathcal{L}^*} = \text{Cov}(X - X_{U^*_Z}, Y_{y^*}, Z_{y^*}, - Y_{y^*,g_{r_2}})
\]
\[
= \text{Cov}(X - X_{U^*_Z}, Y_{y^*}, Z_{y^*}, - Y_{y^*,W_{x^*,u^*_Z}})
\]
\[
= \text{Cov}(X, Y_{y^*}, Z_{y^*}, - Y_{y^*,W_{x^*,u^*_Z}})
\]
\[
= \text{Cov}(X, Y_{y^*}) - \text{Cov}(X, Y_{y^*,W_{x^*,u^*_Z}}).
\]

The last step holds since \( X_{U^*_Z} \) and \( Y_{y^*,u^*_Z} \) are independent counterfactual variables. It will suffice to show that the distributions \( P(x, y_{x^*}), P(x, y_{x^*, W_{x^*,u^*_Z}}) \) are identifiable. By conditioning on \( U^*_Z \),
\[
P(x, y_{y^*,W_{x^*,u^*_Z}}) = \sum_{u^*_Z} P(x, y_{x^*, W_{x^*,u^*_Z}} | u^*_Z) P(u^*_Z)
\]

With \( U^*_Z \) fixed, variables \( X \) and \( Y_{y^*,W_{x^*,u^*_Z}} \) are functions of the exogenous variable \( U \), which is independent of \( U^*_Z \). We thus have the independence relation
\[
U^*_Z \perp \!\!\!\!\!\perp X, Y_{y^*, W_{x^*,u^*_Z}}.
\]

which gives:
\[
P(x, y_{y^*,W_{x^*,u^*_Z}}) = \sum_{u^*_Z} P(x, y_{y^*, W_{x^*,u^*_Z}} | u^*_Z) P(u^*_Z) (35)
\]

By expanding on \( Z, Z_{u^*_Z}, W_{x^*,u^*_Z} \),
\[
P(x, y_{y^*,W_{x^*,u^*_Z}}) = \sum_{z,z',w} P(x, y_{y^*, w, w_{x^*,u^*_Z}, z', z_{u^*_Z}})
\]

By the composition axiom and the exclusion restrictions rule [Pearl, 2000, Ch. 7.3] (treating \( U^*_Z \) as an endogenous variable), for any \( z, x^*, w, u_{z} \),
\[
Z_{u^*_Z} = Z_{x^*,u^*_Z},
\]
\[
Z_{x^*,u^*_Z} = z \Rightarrow W_{x^*,u^*_Z} = W_{x^*,z,u^*_Z} = W_{x^*,z},
\]
\[
Z = Z_{x^*,w},
\]
\[
Z_{x^*,w} = z \Rightarrow Y_{x^*,w} = Y_{x^*,w,z}.
\]

Eqs. 28 and 36 imply
\[
P(x, y_{y^*, w, w_{x^*,u^*_Z}, z', z_{u^*_Z}}) = P(x, y_{y^*, w, w_{x^*,u^*_Z}, z', z_{u^*_Z}})
\]

Since the function \( f_{Z} \) takes only \( U^*_Z \) as an argument. The variables \( Z_{u^*_Z} \) are thus deterministic, i.e.,
\[
P(x, y_{y^*, w, w_{x^*,u^*_Z}, z', z_{u^*_Z}}) = P(x, y_{y^*, w, w_{x^*,u^*_Z}, z', z_{u^*_Z}})
\]

The above equation, together with Eq.35, gives
\[
P(x, y_{y^*,W_{x^*,u^*_Z}})
\]
\[
= \sum_{z,z',w} P(x, y_{y^*, w, z', w_{x^*,z}, z'} | u^*_Z) I\{Z_{u^*_Z} = z\} P(u^*_Z)
\]

The independence restrictions rule [Pearl, 2000, Ch. 7.3] implies that in the model of Fig. 1(a), counterfactuals
Since the domain is normalized, together with Eq. 30, the above equation is equal to:

\[
P(x, y|x^*, w_z^*, u_z^*) = \sum_{z, z', w} P(x|z') P(y|x^*, w, z') P(w|x^*, z) P(z') P(z)
\]

We will next consider the path-specific spurious covariance together with Eq. 30, the above equation is equal to:

\[
P(x, z, z, z|z, z) = \sum_{z, z', w} P(x|z') P(y|x^*, w, z') P(w|x^*, z) P(z') P(z)
\]

Since Lem. 5-8 cover all possible orders \(\mathcal{L}^c, \mathcal{L}^s\), the decompositions Thm. 5 are always identifiable in the standard fairness model.

**Proof of Theorem 7.** By Eqs. 18, 19, 31 and 32, we have for order \(\mathcal{L}_1^c < l_2 \) and \(\mathcal{L}_2^s: l_2 < l_1\),

\[
\text{Cov}_{l_1}^c(X, Y)|I_{l_3,l_4} = \text{Cov}_{l_2}^s(X, Y),
\]

\[
\text{Cov}_{l_1}^c(X, Y)|I_{l_3,l_4} = \text{Cov}_{l_2}^s(X, Y),
\]

\[
\text{Cov}_{l_1}^c(X, Y)|I_{l_3,l_4} = \text{Cov}_{l_2}^s(X, Y),
\]

\[
\text{Cov}_{l_1}^s(X, Y)|I_{l_3,l_4} = \text{Cov}_{l_2}^c(X, Y).
\]

Applying Thm. 1 to the above equations implies that for an arbitrary order \(\mathcal{L}^c\) over \(l_1, l_2\),

\[
\text{Cov}_{l_1}^c(X, Y)|I_{l_3,l_4} = \alpha_{XY}, \text{Cov}_{l_2}^s(X, Y)|I_{l_3,l_4} = \alpha_{WX}\alpha_{YW}.
\]

We will next consider the path-specific spurious covariance of \(l_3, l_4\). As for \(\mathcal{L}^s\) where \(\mathcal{L}_1^s: g_{r_1} < g_{r_2}\), by Eqs. 24-25,

\[
\text{Cov}_{l_1}^s(X, Y)|I_{l_3,l_4} = \alpha_{XY}, \text{Cov}_{l_2}^c(X, Y)|I_{l_3,l_4} = \alpha_{WX}\alpha_{YW},
\]

\[
\text{Cov}_{l_1}^c(X, Y)|I_{l_3,l_4} = \alpha_{XY}, \text{Cov}_{l_2}^s(X, Y)|I_{l_3,l_4} = \alpha_{WX}\alpha_{YW}.
\]

\[
\text{Cov}(X, Z) = \text{Cov}(\alpha_{XZ}Z + U_X, Z) = \alpha_{XZ}\text{Var}(Z) + \alpha_{xz} (38)
\]

\[
\text{Cov}(X, Z_U^z) \text{ equates to:}
\]

\[
\text{Cov}(X, Z_U^z) = \text{Cov}(\alpha_{XZ}Z + U_X, Z_U^z) = \alpha_{XZ}\text{Cov}(U_Z, Z_U^z) + \text{Cov}(U_X, Z_U^z) = 0.
\]

The last step holds since \(U_Z, U_X\) and \(U_Z\) are mutually independent. Eqs. 37-39 together give:

\[
\text{Cov}_{l_1}^c(X, Y)|I_{l_3,l_4} = \alpha_{XZ}\alpha_{YZ}.
\]

Similarly,

\[
\text{Cov}_{l_1}^s(X, Y)|I_{l_3,l_4} = \alpha_{XZ}\alpha_{YZ},
\]

\[
\text{Cov}_{l_1}^c(X, Y)|I_{l_3,l_4} = \alpha_{XZ}\alpha_{YZ},
\]

\[
\text{Cov}_{l_1}^s(X, Y)|I_{l_3,l_4} = \alpha_{XZ}\alpha_{YZ}.
\]

Eqs. 40 - 43 combined imply that for an arbitrary order \(\mathcal{L}^s\),

\[
\text{Cov}_{l_1}^s(X, Y)|I_{l_3,l_4} = \alpha_{XZ}\alpha_{YZ},
\]

\[
\text{Cov}_{l_1}^c(X, Y)|I_{l_3,l_4} = \alpha_{XZ}\alpha_{YZ}.
\]

Specifically, Parameters \(\alpha\) can be estimated from the partial regression coefficients [Pearl, 2000, Ch. 5] as following:

\[
\alpha_{XY} = \gamma_{YX,ZW}, \alpha_{YZ} = \gamma_{YZ,WX}, \alpha_{YW} = \gamma_{YW,ZX}, \alpha_{WX} = \gamma_{WX,Z}, \alpha_{WZ} = \gamma_{WZ,X}, \alpha_{XZ} = \alpha_{XZ}.
\]

2 EXAMPLES

In this section, we will illustrate the results presented in this paper with more detailed examples.

2.1 PATH-SPECIFIC POTENTIAL RESPONSE

Consider the standard model of Fig. 1(a). Recall the path \(g_1: X \rightarrow W_1 \rightarrow W_2 \rightarrow Y\). We next show, step by step, the derivation of the \(g_1\)-specific potential response \(Y_{g_1[x^*]}\). Since the edge \(X \rightarrow Y \notin \{g_1\}\), the set \(\mathcal{X} \rightarrow \mathcal{Y} \neq \emptyset\). We thus have \(\mathcal{S} = (\mathcal{Pa}(Y)|X) \cap \mathcal{V} - X_{\mathcal{Y} \rightarrow \mathcal{X}} = \{X, W_1, W_2\}.\) By Def. 6,

\[
Y_{g_1[x^*]} = Y_{X \rightarrow Y \rightarrow (g_1[x^*]), W_1 \rightarrow (g_1[x^*]), W_2 \rightarrow Y \rightarrow (g_1[x^*])}.
\]
Since the edges $X \rightarrow Y$ and $W_1 \rightarrow Y$ are not subpaths of $g_1$, 
\[ \langle X \rightarrow Y \rangle (g_1) = \langle W_1 \rightarrow Y \rangle (g_1) = \emptyset. \]
By Def. 6, the above equation implies 
\[ X_{\langle X \rightarrow Y \rangle (g_1)[x^*]} = X, \quad W_{1_{\langle W_1 \rightarrow Y \rangle (g_1)[x^*]}} = W_1. \]
$Y_{g_1[x^*]}$ can thus be written as: 
\[ Y_{g_1[x^*]} = Y_{X,W_1,W_2 \langle W_1 \rightarrow Y \rangle (g_1)[x^*]}, \]
Since \( \langle W_2 \rightarrow Y \rangle (g_1) \) returns the subpath \( \{ g_1(X,W_2) \} \), 
\[ Y_{g_1[x^*]} = Y_{X,W_1,W_2 \langle W_2 \rightarrow Y \rangle (g_1[x^*])}. \] (44)
where $W_{2_{\langle W_2 \rightarrow Y \rangle (g_1)[x^*]}}$ is the path-specific potential response of $W_2$. Since $X \rightarrow Y \notin \{ g_1(X,W_2) \}$, the set $S$ for $W_{2_{\langle W_2 \rightarrow Y \rangle (g_1)[x^*]}} \{ X, W_1 \}$. Applying Def. 6 again, 
\[ W_{2_{\langle W_2 \rightarrow Y \rangle (g_1)[x^*]}} = W_{2_{\langle X \rightarrow W_2 \rangle (g_1(X,W_2)[x^*])}} W_{1_{\langle W_1 \rightarrow W_2 \rangle (g_1(X,W_2)[x^*])}}. \]
Since 
\[ \langle X \rightarrow W_2 \rangle (g_1(X,W_2)) = \emptyset, \]
\[ \langle W_1 \rightarrow W_2 \rangle (g_1(X,W_2)) = \{ g_1(X,W_1) \}, \]
$W_{2_{\langle X \rightarrow W_2 \rangle (g_1(X,W_2)[x^*])}}$ can be written as: 
\[ W_{2_{\langle X \rightarrow W_2 \rangle (g_1(X,W_2)[x^*])}} = W_{2_{\langle X \rightarrow W_1 \rangle (g_1(X,W_1)[x^*])}} \] (45)
where $W_{1_{\langle X \rightarrow W_1 \rangle (g_1(X,W_1)[x^*])}}$ is the path-specific potential response of $W_1$. Since the edge $X \rightarrow W_1 = g_1(X,W_1)$, the set $S = (Pa(W_1) \cap V) - X_{\rightarrow W_1} = \emptyset$. By Def. 6, 
\[ W_{1_{\langle X \rightarrow W_1 \rangle (g_1(X,W_1)[x^*])}} = W_{1_{\rightarrow W_1}}. \] (46)
Eqs. 44-46 above give: 
\[ Y_{g_1[x^*]} = Y_{X,W_1,W_2,W_1,W_1} = Y_{W_2W_1}. \]

2.2 DECOMPOSING CAUSAL RELATIONS

We will consider the model in Fig. 6 where causal effects from $X$ and $Y$ are mediated by $W_1, W_2, W_3$, and all directed edges are confounded. There are eight causal paths from $X$ to $Y$:
\begin{align*}
g_1 &: X \rightarrow Y, \\
g_2 &: X \rightarrow W_1 \rightarrow Y, \\
g_3 &: X \rightarrow W_2 \rightarrow Y, \\
g_4 &: X \rightarrow W_3 \rightarrow Y, \\
g_5 &: X \rightarrow W_1 \rightarrow W_2 \rightarrow Y, \\
g_6 &: X \rightarrow W_1 \rightarrow W_3 \rightarrow Y, \\
g_7 &: X \rightarrow W_2 \rightarrow W_3 \rightarrow Y, \\
g_8 &: X \rightarrow W_1 \rightarrow W_2 \rightarrow W_3 \rightarrow Y.
\end{align*}

![Figure 6: Causal diagram for the three-mediators setting where causal paths from $X$ and $Y$ are mediated by $W_1, W_2, W_3$.](image)

Let an order $\mathcal{L}^c$ be $g_i < g_j$ if $i < j$. Thm. 2 is applicable and express the causal covariance $\text{Cov}_{\mathcal{L}^c}(X,Y)$ as:
\[ \text{Cov}^c_{\mathcal{L}^c}(X,Y) = \sum_{i=1}^{8} \text{Cov}^c_{g_i[x^*]}(X,Y) \mathcal{L}^c. \]

The path-specific causal covariance $\{ \text{Cov}^c_{g_i[x^*]}(X,Y) \mathcal{L}^c \}_{i=1,\ldots,8}$ are equal to:
\[ \text{Cov}^c_{g_1[x^*]}(X,Y) \mathcal{L}^c = \text{Cov}(X,Y) - \text{Y}_{g_1[x^*]} \]
\[ = \text{Cov}(X,Y) - Y_{x^*,W_1,W_2,W_3}, \]
\[ \text{Cov}^c_{g_2[x^*]}(X,Y) \mathcal{L}^c = \text{Cov}(X,Y_{g_1[x^*]} - Y_{g_1[x^*]} \)
\[ = \text{Cov}(X,Y_{x^*,W_1,W_2,W_3} - Y_{x^*,W_1,W_2,W_3}), \]
\[ \text{Cov}^c_{g_3[x^*]}(X,Y) \mathcal{L}^c = \text{Cov}(X,Y_{g_1[x^*]} - Y_{g_1[x^*]} \)
\[ = \text{Cov}(X,Y - Y_{x^*,W_1,W_2,W_3} - Y_{x^*,W_1,W_2,W_3}), \]
\[ \text{Cov}^c_{g_4[x^*]}(X,Y) \mathcal{L}^c = \text{Cov}(X,Y_{g_1[x^*]} - Y_{g_1[x^*]} \)
\[ = \text{Cov}(X,Y_{x^*,W_1,W_2,W_3} - Y_{x^*,W_1,W_2,W_3}), \]
\[ \text{Cov}^c_{g_5[x^*]}(X,Y) \mathcal{L}^c = \text{Cov}(X,Y_{g_1[x^*]} - Y_{g_1[x^*]} \)
\[ = \text{Cov}(X,Y_{x^*,W_1,W_2,W_3} - Y_{x^*,W_1,W_2,W_3}), \]
\[ \text{Cov}^c_{g_6[x^*]}(X,Y) \mathcal{L}^c = \text{Cov}(X,Y_{g_1[x^*]} - Y_{g_1[x^*]} \)
\[ = \text{Cov}(X,Y_{x^*,W_1,W_2,W_3} - Y_{x^*,W_1,W_2,W_3}), \]
\[ \text{Cov}^c_{g_7[x^*]}(X,Y) \mathcal{L}^c = \text{Cov}(X,Y_{g_1[x^*]} - Y_{g_1[x^*]} \)
\[ = \text{Cov}(X,Y_{x^*,W_1,W_2,W_3} - Y_{x^*,W_1,W_2,W_3}), \]
\[ \text{Cov}^c_{g_8[x^*]}(X,Y) \mathcal{L}^c = \text{Cov}(X,Y_{g_1[x^*]} - Y_{g_1[x^*]} \)
\[ = \text{Cov}(X,Y_{x^*,W_1,W_2,W_3} - Y_{x^*,W_1,W_2,W_3}). \]

2.3 DECOMPOSING SPURIOUS RELATIONS

We will consider the generalized two-confounders setting described in Fig. 7(a) where $X$ and $Y$ are confounded by $Z_1, Z_2$. The exogenous variables $U_1, U_2$
associated with $Z_1, Z_2$ are represented explicitly in the causal diagram. In the model of Fig. 7(a), $U^s = \{U_1, U_2\}$ which affects the observational $X$ and the counterfactuals $Y^s$ through causal paths shown in Fig. 7(b). There are thus five spurious treks:

$$\begin{align*}
\tau_1 &= (g_1^1, g_1^1), \quad \tau_2 = (g_1^1, g_1^2), \quad \tau_3 = (g_1^1, g_1^1) \\
\tau_4 &= (g_2^1, g_2^2), \quad \tau_5 = (g_2^2, g_2^2).
\end{align*}$$

The treatment $X$ and the outcome $Y$ are connected through four spurious paths:

$$\begin{align*}
l_1 : X \leftarrow Z_1 \rightarrow Y, & \quad l_2 : X \leftarrow Z_2 \rightarrow Y, \\
l_3 : X \leftarrow Z_1 \rightarrow Z_2 \rightarrow Y, & \quad l_4 : X \leftarrow Z_2 \leftarrow Z_1 \rightarrow Y. 
\end{align*}$$

Let an order $L_i^s$ be $U_1 < U_2$. Let an order $L_i^s$ be $g_i^j < g_i^k$ if $j < k$. The order $L_i$, is similarly defined. Thm. 3 decomposes the spurious covariance $\text{Cov}^s_{x^s}(X, Y)$ over the spurious paths $l_1, \ldots, l_4$:

$$\text{Cov}^s_{x^s}(X, Y) = \sum_{i=1}^4 \text{Cov}^s_{\tau_i}(X, Y) L_i^s.$$ 

The path-specific spurious covariance

$$\{\text{Cov}^s_{\tau_i}(X, Y) L_i^s\}_{i=1, \ldots, 4}$$

are equal to:

$$\begin{align*}
\text{Cov}^s_{l_1}(X, Y) L_1^s &= \text{Cov}^s_{l_1}(X, Y) L_1^s \\
&= \text{Cov}(X - X_{Z_1} - Z_1, z_{2, Y^s} - Y^s, g_{l_2}) \\
&= \text{Cov}(X - X_{Z_1} - Z_1, z_{2, Y^s} - Y^s, g_{l_1}).
\end{align*}$$

$$\begin{align*}
\text{Cov}^s_{l_2}(X, Y) L_2^s &= \text{Cov}^s_{l_2}(X, Y) L_2^s \\
&= \text{Cov}(X - X_{Z_1} - Z_1, z_{2, Y^s} - Y^s, g_{l_2}) \\
&= \text{Cov}(X - X_{Z_1} - Z_1, z_{2, Y^s} - Y^s, g_{l_1}).
\end{align*}$$

$\text{Cov}^s_{l_3}(X, Y) L_3^s = \text{Cov}^s_{l_3}(X, Y) L_3^s$ 

$\text{Cov}^s_{l_4}(X, Y) L_4^s = \text{Cov}^s_{l_4}(X, Y) L_4^s$

Figure 7: (a) Causal diagram for the two-confounders setting where $X$ to $Y$ are confounded by $Z_1, Z_2$; (b) Causal paths through which the exogenous variables $U_1, U_2$ affect $X$ and $Y^s$ in the two-confounders setting.

### 2.4 Path-Specific Decomposition

Considering the model of Fig. 8, the treatment $X$ and the outcome $Y$ are connected by the causal paths:

$$\begin{align*}
l_1 : X \rightarrow Y, & \quad l_2 : X \rightarrow W_1 \rightarrow Y, \\
l_3 : X \rightarrow W_2 \rightarrow Y, & \quad l_4 : X \rightarrow W_1 \rightarrow W_2 \rightarrow Y, 
\end{align*}$$

and the spurious paths:

$$\begin{align*}
l_5 : X \leftarrow Z_1 \rightarrow Y, & \quad l_6 : X \leftarrow Z_1 \rightarrow W_1 \rightarrow Y, \\
l_7 : X \leftarrow Z_1 \rightarrow W_2 \rightarrow Y, & \quad l_8 : X \leftarrow Z_1 \rightarrow W_1 \rightarrow W_2 \rightarrow Y, \\
l_9 : X \leftarrow Z_1 \rightarrow Z_2 \rightarrow Y, & \quad l_{10} : X \leftarrow Z_1 \rightarrow Z_2 \rightarrow W_1 \rightarrow Y, \\
l_{11} : X \leftarrow Z_1 \rightarrow Z_2 \rightarrow W_2 \rightarrow Y, & \quad l_{12} : X \leftarrow Z_1 \rightarrow Z_2 \rightarrow W_1 \rightarrow W_2 \rightarrow Y, \\
l_{13} : X \leftarrow Z_2 \leftarrow Z_1 \rightarrow Y, & \quad l_{14} : X \leftarrow Z_2 \leftarrow Z_1 \rightarrow W_1 \rightarrow Y, \\
l_{15} : X \leftarrow Z_2 \leftarrow Z_1 \rightarrow W_2 \rightarrow Y, & \quad l_{16} : X \leftarrow Z_2 \leftarrow Z_1 \rightarrow W_1 \rightarrow W_2 \rightarrow Y, \\
l_{17} : X \leftarrow Z_2 \rightarrow Y, & \quad l_{18} : X \leftarrow Z_2 \rightarrow W_1 \rightarrow Y, \\
l_{19} : X \leftarrow Z_2 \rightarrow W_2 \rightarrow Y, & \quad l_{20} : X \leftarrow Z_2 \rightarrow W_1 \rightarrow W_2 \rightarrow Y.
\end{align*}$$

Let $U_1, U_2$ denote the independent errors associated with the confounders $Z_1, Z_2$ respectively. In this model, $U^s = \{U_1, U_2\}$ where the causal paths $\Pi^s(U_1, X)$ and
Figure 8: Causal diagram for the two-mediators-two-confounders setting where $X$ to $Y$ are confounded by $Z_1, Z_2$ and mediated by $W_1, W_2$.

\[ \Pi^c(U_1, X | Y) \] are:

- $g_{11}: U_1 \rightarrow Z_1 \rightarrow X$,
- $g_{12}: U_1 \rightarrow Z_1 \rightarrow Z_2 \rightarrow X$,
- $g_{21}^1 : U_1 \rightarrow Z_1 \rightarrow Y$,
- $g_{22}^1 : U_1 \rightarrow Z_1 \rightarrow W_1 \rightarrow Y$,
- $g_{12}^1 : U_1 \rightarrow Z_1 \rightarrow Z_2 \rightarrow W_2 \rightarrow Y$,
- $g_{13}^1 : U_1 \rightarrow Z_1 \rightarrow Z_2 \rightarrow W_1 \rightarrow W_2 \rightarrow Y$,

and the causal paths $\Pi^c(U_2, X)$ and $\Pi^c(U_2, X | Y)$ are:

- $g_{22}^2 : U_2 \rightarrow Z_2 \rightarrow X$,
- $g_{23}^2 : U_1 \rightarrow Z_2 \rightarrow Y$,
- $g_{24}^2 : U_2 \rightarrow Z_2 \rightarrow W_1 \rightarrow Y$,
- $g_{25}^2 : U_1 \rightarrow Z_2 \rightarrow W_2 \rightarrow Y$,
- $g_{26}^2 : U_1 \rightarrow Z_2 \rightarrow W_1 \rightarrow W_2 \rightarrow Y$.

There are thus twenty spurious treks from $X$ to $Y$:

- $\tau_1 = (g_{41}^1, g_{r_1}^1)$, $\tau_2 = (g_{r_1}^1, g_{42}^1)$, $\tau_3 = (g_{42}^1, g_{r_3}^1)$,
- $\tau_4 = (g_{41}^1, g_{r_4}^1)$, $\tau_5 = (g_{r_1}^1, g_{43}^1)$, $\tau_6 = (g_{42}^1, g_{r_6}^1)$,
- $\tau_7 = (g_{41}^1, g_{r_7}^1)$, $\tau_8 = (g_{r_1}^1, g_{44}^1)$, $\tau_9 = (g_{42}^1, g_{r_9}^1)$,
- $\tau_{10} = (g_{43}^1, g_{r_{10}}^1)$, $\tau_{11} = (g_{r_1}^1, g_{45}^1)$, $\tau_{12} = (g_{42}^1, g_{r_{12}}^1)$,
- $\tau_{13} = (g_{43}^1, g_{r_{13}}^1)$, $\tau_{14} = (g_{r_1}^1, g_{46}^1)$, $\tau_{15} = (g_{42}^1, g_{r_{15}}^1)$,
- $\tau_{16} = (g_{44}^1, g_{r_{16}}^1)$, $\tau_{17} = (g_{45}^1, g_{r_{17}}^1)$, $\tau_{18} = (g_{46}^1, g_{r_{18}}^1)$,
- $\tau_{19} = (g_{47}^1, g_{r_{19}}^1)$, $\tau_{20} = (g_{48}^1, g_{r_{20}}^1)$.

The set \( \mathcal{T}^* (l_i) \) is a partition over the spurious treks $\tau_{1-20}$:

\[
\begin{align*}
\mathcal{T}^* (l_5) &= \{ \tau_1 \}, \\
\mathcal{T}^* (l_7) &= \{ \tau_3 \}, \\
\mathcal{T}^* (l_9) &= \{ \tau_5 \}, \\
\mathcal{T}^* (l_{11}) &= \{ \tau_7 \}, \\
\mathcal{T}^* (l_{13}) &= \{ \tau_9 \}, \\
\mathcal{T}^* (l_{15}) &= \{ \tau_{11} \}, \\
\mathcal{T}^* (l_{17}) &= \{ \tau_{13}, \tau_{17} \}, \\
\mathcal{T}^* (l_{19}) &= \{ \tau_{15}, \tau_{19} \}, \\
\mathcal{T}^* (l_{20}) &= \{ \tau_{16}, \tau_{20} \}.
\end{align*}
\]

Let an order $\mathcal{L}_n$ be $l_1 < l_2 < l_3 < l_4$, and an order $\mathcal{L}_u$ be $U_1 < U_2$. Let an order $\mathcal{L}_l$ ($\mathcal{L}_u$) follow the rule $g_{ij} < g_{ij}$ ($g_{ij} < g_{ij}$) if $j < k$. Thm. 5 is applicable and decomposes the covariance $\text{Cov}(X, Y)$ over paths $\{l_{1-20}\}$:

\[
\text{Cov}(X, Y) = \sum_{i=1}^{20} \text{Cov}_{\mathcal{L}_l}(X, Y) + \sum_{i=5}^{20} \text{Cov}_{\mathcal{L}_u}(X, Y).
\]

The path-specific causal covariance $\{ \text{Cov}_{\mathcal{L}_l}(X, Y) \} = 1, \ldots, 4$ are equal to:

\[
\begin{align*}
\text{Cov}_{\mathcal{L}_l}(X, Y) &= \text{Cov}(X, Y - Y_{l_1[x^*]}) \\
&= \text{Cov}(X, Y - Y_{l_2[x^*], W_2}), \\
\text{Cov}_{\mathcal{L}_u}(X, Y) &= \text{Cov}(X, Y, W_1 - Y_{l_1[x^*], W_2}), \\
\text{Cov}_{\mathcal{L}_l}(X, Y) &= \text{Cov}(X, Y, W_1, W_2 - Y_{l_2[x^*], W_2}), \\
\text{Cov}_{\mathcal{L}_u}(X, Y) &= \text{Cov}(X, Y, W_1, W_2, W_{2^*} - Y_{l_1[x^*], W_2}), \\
\text{Cov}_{\mathcal{L}_l}(X, Y) &= \text{Cov}(X, Y, W_1, W_2, W_{2^*} - Y_{l_2[x^*], W_2}).
\end{align*}
\]

The path-specific spurious covariance $\{ \text{Cov}_{\mathcal{L}_u}(X, Y) \} = 5, \ldots, 20$ are equal to:

\[
\begin{align*}
\text{Cov}_{\mathcal{L}_u}(X, Y) &= \text{Cov}(X, Y - Y_{l_1[x^*], W_1}) \\
&= \text{Cov}(X, Y, Z_{l_1[x^*], Z_2}, Y_{l_2[x^*], W_1, W_{2^*}, Z_{l_1[x^*], Z_2}}), \\
\text{Cov}_{\mathcal{L}_u}(X, Y) &= \text{Cov}(X, Y, Z_{l_1[x^*], Z_2}, Y_{l_2[x^*], W_1, W_{2^*}, Z_{l_1[x^*], Z_2}}),
\end{align*}
\]
\[
\text{Cov}_{t_1[x^*]}(X, Y)_{t_1} = \text{Cov}_{t_0[x^*]}(X, Y)_{t_0} \\
= \text{Cov}(X - X_{g_{t_1}}, Y_{x^*} - g_{[1,2]} - Y_{x^*} - g_{[1,3]}) \\
= \text{Cov}(X - X_{Z_{U_{t_1}}}, Y_{x^*} - W_{1_{t_1}} - Z_{V_{t_1}} - Z_{t_1}, \\
- Y_{x^*}, W_{1_{t_1}}, Z_{U_{t_1}}, Z_{t_1}, Z_{V_{t_1}} - Z_{t_1}), \\
\text{Cov}_{t_2[x^*]}(X, Y)_{t_2} = \text{Cov}_{t_1[x^*]}(X, Y)_{t_1} \\
= \text{Cov}(X - X_{g_{t_1}}, Y_{x^*} - g_{[1,3]} - Y_{x^*} - g_{[1,5]}) \\
= \text{Cov}(X - X_{Z_{U_{t_1}}}, Y_{x^*} - W_{1_{t_1}} - Z_{V_{t_1}} - Z_{t_1}, \\
- Y_{x^*}, W_{1_{t_1}}, Z_{U_{t_1}}, Z_{t_1}, Z_{V_{t_1}} - Z_{t_1}), \\
\text{Cov}_{t_3[x^*]}(X, Y)_{t_3} = \text{Cov}_{t_2[x^*]}(X, Y)_{t_2} \\
= \text{Cov}(X - X_{g_{t_1}}, Y_{x^*} - g_{[1,5]} - Y_{x^*} - g_{[1,7]}) \\
= \text{Cov}(X - X_{Z_{U_{t_1}}}, Y_{x^*} - W_{1_{t_1}} - Z_{V_{t_1}} - Z_{t_1}, \\
- Y_{x^*}, W_{1_{t_1}}, Z_{U_{t_1}}, Z_{t_1}, Z_{V_{t_1}} - Z_{t_1}), \\
\text{Cov}_{t_4[x^*]}(X, Y)_{t_4} = \text{Cov}_{t_3[x^*]}(X, Y)_{t_3} \\
= \text{Cov}(X - X_{g_{t_1}}, Y_{x^*} - g_{[1,7]} - Y_{x^*} - g_{[1,9]}) \\
= \text{Cov}(X - X_{Z_{U_{t_1}}}, Y_{x^*} - W_{1_{t_1}} - Z_{V_{t_1}} - Z_{t_1}, \\
- Y_{x^*}, W_{1_{t_1}}, Z_{U_{t_1}}, Z_{t_1}, Z_{V_{t_1}} - Z_{t_1}), \\
\text{Cov}_{t_5[x^*]}(X, Y)_{t_5} = \text{Cov}_{t_4[x^*]}(X, Y)_{t_4} \\
= \text{Cov}(X - X_{g_{t_1}}, Y_{x^*} - g_{[1,9]} - Y_{x^*} - g_{[1,11]}) \\
= \text{Cov}(X - X_{Z_{U_{t_1}}}, Y_{x^*} - W_{1_{t_1}} - Z_{V_{t_1}} - Z_{t_1}, \\
- Y_{x^*}, W_{1_{t_1}}, Z_{U_{t_1}}, Z_{t_1}, Z_{V_{t_1}} - Z_{t_1}).
\]
\[
\text{Cov}_{t_{10}[x^*]}(X, Y) \mathcal{L}_x^s = \text{Cov}_{t_{10}[x^*]}(X, Y) \mathcal{L}_x^s + \text{Cov}_{t_{10}[x^*]}(X, Y) \mathcal{L}_x^s
\]

\[
= \text{Cov}(X_{g_1^t} - X_{g_{1,2}^t}, Y_{x^*, g_{1,16}^t} - Y_{x^*, g_{1,7}^t}) + \text{Cov}(X_{g_{1,2}^t} - X_{g_{1,2}^t} \cup g_{15}^t, Y_{x^*, g_{1,16}^t} \cup g_{1,12}^t - Y_{x^*, g_{1,8}^t} \cup g_{1,3}^t)
\]

\[
= \text{Cov}(X_{Z_{U_1}}, Y_{x^*, U_1}, W_{x^*, U_1}^r, W_{x^*, U_1}^r, Z_{U_1}, Z_{2U_1} - X_{U_1}^r, \nonumber
\]

\[
- Y_{x^*, U_1}, W_{x^*, U_1}^r, W_{x^*, U_1}^r, Z_{U_1}, Z_{2U_1}^r, Z_{U_1}, Z_{2U_1} - X_{U_1}^r
\]

\[
\text{Cov}_{t_{20}[x^*]}(X, Y) \mathcal{L}_x^s = \text{Cov}_{t_{20}[x^*]}(X, Y) \mathcal{L}_x^s + \text{Cov}_{t_{20}[x^*]}(X, Y) \mathcal{L}_x^s
\]

\[
= \text{Cov}(X_{g_1^t} - X_{g_{1,2}^t}, Y_{x^*, g_{1,16}^t} - Y_{x^*, g_{1,7}^t}) + \text{Cov}(X_{g_{1,2}^t} - X_{g_{1,2}^t} \cup g_{15}^t, Y_{x^*, g_{1,16}^t} \cup g_{1,12}^t - Y_{x^*, g_{1,8}^t} \cup g_{1,3}^t)
\]

\[
= \text{Cov}(X_{Z_{U_1}}, Y_{x^*, U_1}, W_{x^*, U_1}^r, W_{x^*, U_1}^r, Z_{U_1}, Z_{2U_1} - X_{U_1}^r, \nonumber
\]

\[
- Y_{x^*, U_1}, W_{x^*, U_1}^r, W_{x^*, U_1}^r, Z_{U_1}, Z_{2U_1}^r, Z_{U_1}, Z_{2U_1} - X_{U_1}^r
\]

References
