Machine Learning (2023) 112:4201–4225
https://doi.org/10.1007/s10994-022-06159-y

Causal discovery with a mixture of DAGs

Eric V. Strobl[1]

Received: 22 February 2021 / Revised: 6 December 2021 / Accepted: 19 February 2022 /
Published online: 30 March 2022

Abstract
Real causal processes may contain cycles, evolve over time or differ between populations. However, many graphical models cannot accommodate these conditions. We propose to model causation using a mixture of directed cyclic graphs (DAGs); each sample follows a joint distribution that factorizes according to a DAG, but the DAG may differ between samples due to multiple independent factors. We then introduce an algorithm called Causal Inference over Mixtures that uses longitudinal data to infer a graph summarizing the causal relations generated from a mixture of DAGs even when cycles, non-stationarity, latent variables or selection bias exist. Experiments demonstrate improved performance in inferring ancestral relations as compared to prior approaches. R code is available at https://github.com/ericstrobl/CIM.

Keywords Causal discovery · Longitudinal data · Directed acyclic graph · Mixture of DAGs

1 Introduction

Causal discovery or causal inference refers to the process of inferring causation from data. Investigators usually perform causal discovery using randomized controlled trials (RCTs). However, RCTs can be impractical or unethical to perform. For example, scientists cannot randomly administer illicit substances or withhold active treatment from critically ill subjects. Many investigators therefore experiment with animals, which in turn raises more ethical questions, knowing that the derived results may not directly apply to humans.

In this paper, we develop an algorithm that discovers causation directly from observational data, or data collected without randomization. Denote the variables in an observational dataset by $\boldsymbol{X}$. We summarize the causal relations between variables in $\boldsymbol{X}$ using a *directed graph*, where the directed edge $X_i \to X_j$ with $X_i, X_j \in \boldsymbol{X}$ means that X_i is a *direct cause* of X_j. Similarly, X_i is a *cause* of X_j if there exists a directed path, or a sequence of

Editor: Annalisa Appice, Grigorios Tsoumakas.

✉ Eric V. Strobl
ericvonstrobl@gmail.com

[1] Vanderbilt University Medical Center, 1601 23rd Ave S, Nashville, TN 37212, USA

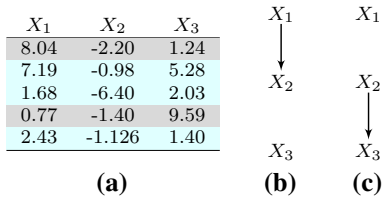


Fig. 1 A dataset containing samples from a distribution modeled by a mixture of two directed graphs in (**a**). The samples in blue arise from the graph in (**b**), while the samples in grey from (**c**) (Color figure online)

directed edges, from X_i to X_j. We want to recover the directed graph as best as possible using the observational dataset.

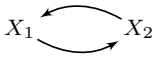
We may however fail to *always* sample from a probability distribution obeying a *single* directed graph in practice. In this paper, we consider the relaxed scenario where each sample follows a single directed graph, but the graph may differ between samples. Consider for example the dataset shown in Fig. 1a. The samples in blue arise from the directed graph shown in Fig. 1b, but the samples in grey arise from the graph in Fig. 1c. If we do not have color coding or labels distinguishing the two different sample types, then the probability distribution obeys a mixture of the two directed graphs. We thus focus on inferring a graph summarizing the relationships encoded in the component graphs using samples from the mixture distribution. Note that we may have many more than two component graphs in practice.[1] We do not assume access to any type of prior knowledge about the number of component graphs.

This approach is particularly powerful for modeling non-equilibriated causal processes with *cycles*. Directed graphs in nature often contain feedback loops, or cycles, where X_i causes X_j and X_j directly causes X_i. For example, Fig. 2a depicts a portion of the thyroid system where X_1 denotes the thyroid stimulating hormone (TSH) and X_2 the T4 hormone (T4). TSH released from the anterior pituitary regulates T4 hormone release from the thyroid gland, while T4 feeds back to inhibit TSH release. Cycles such as these abound in practice, so we must develop algorithms that can accommodate them in order to accurately model causal processes.

Authors have proposed multiple interpretations of cycles in the causal discovery literature. The most popular approach assumes that X_i causes X_j and X_j causes X_i *simultaneously* in an equilibrium distribution (see Appendix 8.1 for a detailed description). This formulation however differs from the standard way cycles are taught in biology, where X_i causes X_j, *then* X_j causes X_i, *then* X_i causes X_j and so forth in a *non-equilibriated* process rarely reaching a stationary point (e.g., Chapters 2, 7 and 15 in (Alberts et al. 2015)); notice that causation occurs in succession and never simultaneously, similar to a discrete switching process.

We therefore propose a different interpretation of cycles based on non-equilibriated distributions, where we model a potentially cyclic causal process using multiple directed acyclic graphs (DAGs), or graphs with directed edges but no cycles. The causal process is represented as a DAG at any single point in time, but the DAG may change across time to accommodate feedback. We illustrate the idea by decomposing the cycle in Fig. 2a into two DAGs: TSH → T4 and T4 → TSH. For each sample, TSH first causes T4 release at time point t_1 and then T4 inhibits TSH release at time point $t_2 > t_1$. Such

[1] The examples in the paper only include two component graphs for ease of presentation and to conserve space; they do not imply an additional assumption.



(a)

X_1	X_2	T
0.21	-0.20	1.29
0.68	-0.47	7.30
1.05	-0.19	4.33
0.72	-1.40	0.10
0.13	-0.56	2.91

(b)

X_1	X_3	X_4	X_7	$\cdots$	T_3
0.31	-1.01	5	0	$\cdots$	1.29
0.89	-0.58	6	0	$\cdots$	7.30
1.11	-0.79	2	1	$\cdots$	4.33
0.14	-1.23	5	0	$\cdots$	0.10
0.21	-0.20	4	1	$\cdots$	2.91
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

(c)

Fig. 2 We decompose the cycle in (**a**) into two DAGs: $X_1 \to X_2$ and $X_2 \to X_1$. The blue samples in (**b**) refer to samples arising from the first DAG and the grey ones to the second. The table in (**c**) depicts a more realistic dataset containing more variables and samples (Color figure online)

a successive relationship is *known* to hold in biology, *and* this process never reaches equilibrium because the hormone levels fluctuate throughout the entire day (Pirahanchi et al. 2021; Lucke et al. 1977). We however can only measure each sample at a single point in time, so the observational dataset in Fig. 2b contains some samples in blue when TSH causes T4 and others in grey when T4 causes TSH. If we do not observe the time variable T, then the observational dataset arises from a *mixture of DAGs* where the mixing occurs over time: $p(X_1, X_2) = \sum_T p(X_1, X_2|T)p(T)$. We interpret the cycle in Fig. 2a as the uncertainty in knowing which samples came from $X_1 \to X_2$ and which from $X_1 \leftarrow X_2$. Cycles therefore arise because of multiple DAGs, each representing an unknown time period in a non-equilibriated process where causation occurs in succession, rather than from a single cyclic graph representing an equilibriated process where causation occurs simultaneously. We now must infer the directed graph in Fig. 2a using the samples from X_1 and X_2 alone. In practice, we observe more than two random variables without color coding and mixing occurs over more than two graphs indexed by multiple variables $\boldsymbol{T}$ denoting entities such as time, gender, income and disease status. Figure 2c therefore depicts a more realistic dataset.

We now develop a method for recovering a directed graph summarizing the causal relations arising from a mixture of DAGs. We do so by first reviewing related work in Sect. 2. We then provide background in Sect. 3. Section 4 introduces the mixture of DAGs framework. In Sect. 5, we explain why existing algorithms fail and then detail a new method called Causal Inference over Mixtures (CIM) to infer causal relations using longitudinal data. We then report experimental results in Sect. 6 highlighting the superiority of CIM compared to prior approaches on both real and synthetic datasets. We finally conclude the paper in Sect. 7. We delegate all proofs to the Appendix.

This paper improves upon a previously published workshop paper (Strobl 2019), where we proposed the mixture of DAGs framework as well as an early version of the CIM algorithm. The paper unfortunately has some limitations, which we corrected herein. We specifically make the following new contributions in this submission:

(a) We simplify and improve the description of the mixture of DAGs framework.
(b) The original global Markov property is incorrect. We prove the correct property without assuming strict positivity, a single latent and discrete variable in $\boldsymbol{T}$, parametric forms or particular variable orderings across the constituent DAGs.

(c) We improve the CIM algorithm by accommodating the new global Markov property and increasing the number of conditioning sets. This substantially improves performance by recovering a sparser causal graph.
(d) We improve the known ground truth for the real datasets. This allows us to run all experiments using sensitivity, fallout and distance from the upper left hand corner of the ROC in order to directly compare algorithms across all three metrics.
(e) We run all algorithms using the GCM conditional independence (CI) test which controls the Type I error rate better than the RCoT test used in the previous paper (Shah et al. 2020; Strobl et al. 2018).

These changes improve the arguments substantially and lead to an even better causal inference algorithm.

2 Related work

Several algorithms perform causal discovery with cycles. Most of these methods assume *stationarity*, or a stable equilibrium distribution over time. The Fast Causal Inference (FCI) algorithm for example infers causal relations by executing CI tests in greedy sequence (Spirtes et al. 2000; Zhang 2008). The algorithm was initially developed for the acyclic case, but it can handle cycles, provided that we can ignore them by transforming the cyclic graph into an acyclic one sharing the same CI relations (Mooij and Claassen 2020; Spirtes 1995). FCI thus cannot recover within-cycle causal relations, but other algorithms can. The Cyclic Causal Discovery (CCD) algorithm for instance works well when no selection bias or latent variables exist (Richardson 1996). The Cyclic Causal Inference (CCI) algorithm extends CCD to handle selection bias and latent variables, but both algorithms require linear or discrete variables for correctness (Strobl 2018; Forré and Mooij 2017, 2018). Investigators have since proposed a variety of extensions based on exhaustive search that can infer causal relations with higher accuracy (Hyttinen et al. 2013, 2014; Lu et al. 2021). These methods however can have trouble scaling to higher dimensions due to the combinatorial search space over directed graphs and the potentially exponential increase in conditioning set sizes of the CI tests.

Another set of methods can handle non-equilibrium distributions, but most of them require a single underlying directed graph either in discrete time with dynamic Bayesian networks (Murphy 2002) or in continuous time with dynamic structural causal models (Dagum et al. 1995; Zhang et al. 2017; Rubenstein et al. 2018; Bellot et al. 2021). Two methods exist for recovering causal processes with multiple graphs (Strobl 2017; Zhang and Glymour 2018), but they assume a mixture of parametric distributions. Saeed et al. (2020) showed that FCI can also handle non-stationarity, provided that a certain variable ordering assumption holds across time. This ordering however can easily be violated with shifting graphical structure or cycles. CIM improves upon all of these methods by allowing arbitrary variable ordering, non-linearity, cycles, non-stationarity, non-parametric distributions, changing graphical structure, latent variables and selection bias.

Finally, several methods can discover causal structure under different known contexts, usually framed in terms of experimental conditions (Mooij et al. 2020; Squires et al. 2020; Ke et al. 2019; Jaber et al. 2020). These algorithms require *observed* variables indexing the contexts. Most also assume that the observational distribution follows a single underlying DAG, from which we can model experiments by removing parents. We instead consider

unknown contexts and violations of acyclicity. The observational distribution therefore follows a mixture of DAGs, where mixing occurs over *latent* context variables and the DAGs obey potentially different partial orderings.

3 Background

We now delve into the background material required to understand the proposed methodology.

3.1 Terminology

In addition to directed edges, we consider other edge types including: $\leftrightarrow$ (bidirected), — (undirected), $\circ\!\!\rightarrow$ (partially directed), $\circ\!\!-$ (partially undirected) and $\circ\!\!-\!\!\circ$ (nondirected). The edges contain three endpoint types: arrowheads, tails and circles. Each circle corresponds to an unknown endpoint thus denoting either an arrowhead or tail. We say that two vertices X_i and X_j are *adjacent* if there exists an edge between the two vertices. We refer to the triple $X_i \ast\!\!\rightarrow X_j \leftarrow\!\!\ast\, X_k$ as a *collider* or *v-structure*, where each asterisk corresponds to an arbitrary endpoint type, when X_i and X_k are non-adjacent. The triple $X_i \ast\!\!-\!\!\ast\, X_j \ast\!\!-\!\!\ast\, X_k$ is conversely a *triangle* if X_i and X_k are adjacent. Unless stated otherwise, a *path* is a sequence of edges without repeated vertices. X_i is an *ancestor* of X_j if there exists a directed path from X_i to X_j or $X_i = X_j$. We write $X_i \in \mathrm{Anc}_{\mathbb{G}}(X_j)$ when X_i is an ancestor of X_j in the graph $\mathbb{G}$. We also apply the definition of an ancestor to a set of vertices $\boldsymbol{Y} \subseteq \boldsymbol{X}$ as follows:

$$\mathrm{Anc}_{\mathbb{G}}(\boldsymbol{Y}) = \{X_i | X_i \in \mathrm{Anc}_{\mathbb{G}}(Y_j) \text{ for some } Y_j \in \boldsymbol{Y}\}.$$

If $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ are disjoint sets of vertices in $\boldsymbol{X}$, then $\boldsymbol{A}$ and $\boldsymbol{B}$ are said to be *d-connected* by $\boldsymbol{C}$ in a directed graph $\mathbb{G}$ if there exists a path Π between some vertex in $\boldsymbol{A}$ and some vertex in $\boldsymbol{B}$ such that, for any collider X_i on Π, X_i is an ancestor of $\boldsymbol{C}$ and no non-collider on Π is in $\boldsymbol{C}$. We also say that $\boldsymbol{A}$ and $\boldsymbol{B}$ are *d-separated* by $\boldsymbol{C}$ if they are not d-connected by $\boldsymbol{C}$. For shorthand, we write $\boldsymbol{A} \perp\!\!\!\perp_d \boldsymbol{B} | \boldsymbol{C}$ to denote d-separation and $\boldsymbol{A} \not\perp\!\!\!\perp_d \boldsymbol{B} | \boldsymbol{C}$ to denote d-connection. The set $\boldsymbol{C}$ is more specifically called a *minimal separating set* if we have $\boldsymbol{A} \perp\!\!\!\perp_d \boldsymbol{B} | \boldsymbol{C}$ but $\boldsymbol{A} \not\perp\!\!\!\perp_d \boldsymbol{B} | \boldsymbol{D}$, where $\boldsymbol{D}$ denotes any proper subset of $\boldsymbol{C}$.

A *mixed graph* contains edges with only arrowheads or tails, while a *partially oriented mixed graph* may also include circles. We focus on mixed graphs that contain at most one edge between any two vertices. We can associate a mixed graph $\mathbb{G}^*$ with a directed graph $\mathbb{G}$ as follows. We first partition $\boldsymbol{X} = \boldsymbol{O} \cup \boldsymbol{L} \cup \boldsymbol{S}$ denoting observed, latent and selection variables, respectively; the selection variables allow us to model the selection bias frequently present in real data. We then consider a graph over $\boldsymbol{O}$ summarizing the ancestral relations in $\mathbb{G}$ with the following endpoint interpretations: $O_i \ast\!\!\rightarrow O_j$ in $\mathbb{G}^*$ if $O_j \notin \mathrm{Anc}_{\mathbb{G}}(O_i \cup \boldsymbol{S})$, and $O_i \ast\!\!-\!\!O_j$ in $\mathbb{G}^*$ if $O_j \in \mathrm{Anc}_{\mathbb{G}}(O_i \cup \boldsymbol{S})$.

3.2 Probabilistic interpretation

We associate a density $p(\boldsymbol{X})$ to a DAG $\mathbb{G}$ by requiring that the density factorize into the product of conditional densities of each variable given its parents:

$$p(\boldsymbol{X}) = \prod_{i=1}^{p} p(X_i|\text{Pa}_{\mathbb{G}}(X_i)). \tag{1}$$

Any distribution which factorizes as above also satisfies the *global Markov property* w.r.t. $\mathbb{G}$ where, if we have $\boldsymbol{A} \perp\!\!\!\perp_d \boldsymbol{B}|\boldsymbol{C}$ in $\mathbb{G}$, then $\boldsymbol{A}$ and $\boldsymbol{B}$ are conditionally independent given $\boldsymbol{C}$ (Lauritzen et al. 1990). We denote the conditional independence (CI) as $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B}|\boldsymbol{C}$ for short. We refer to the converse of the global Markov property as *d-separation faithfulness*. An algorithm is *constraint-based* if it utilizes CI testing to recover some aspects of $\mathbb{G}^*$ as a consequence of the global Markov property and d-separation faithfulness.

4 Mixture of DAGs

We introduce the framework with univariate T and then generalize to multivariate $\boldsymbol{T}$ because the univariate case is simpler. Note that Spirtes (1994) considered the univariate setting as well, but he also (1) assumed that T is discrete, and (2) described the framework in terms of structural equations rather than densities. We do not impose any type restrictions and detail the density approach. We finally consider the multivariate case which is entirely novel.

4.1 Univariate case

We consider the set of vertices $\boldsymbol{Z} = \boldsymbol{X} \cup T$. We divide $\boldsymbol{Z}$ into three non-overlapping sets $\boldsymbol{O}$, $\boldsymbol{L}$ and $\boldsymbol{S}$ denoting observed, latent and selection variables, respectively. At each time point t, we consider the joint density $p(\boldsymbol{X}, T = t)$ and assume that it factorizes according to a DAG $\mathbb{G}_t$ over $\boldsymbol{Z}$:

$$\begin{aligned} p(\boldsymbol{X}, T = t) &= p(T = t)p(\boldsymbol{X}|T = t) \\ &= p(T = t) \prod_{i=1}^{p} p(X_i|\text{Pa}_t(X_i)), \end{aligned}$$

where $\text{Pa}_t(Z_i)$ refers to $\text{Pa}_{\mathbb{G}_t}(Z_i)$ for shorthand, the parent set of Z_i at time point t. We analyze the following density:

$$p(\boldsymbol{Z}) = \prod_{i=1}^{p+1} p(Z_i|\text{Pa}_T(Z_i)), \tag{2}$$

where $\text{Pa}_T(T) = \emptyset$. The above equation differs from Eq. (1) for a single DAG; the parent set $\text{Pa}_{\mathbb{G}}(Z_i)$ remains constant over time in Eq. (1), but the parent set $\text{Pa}_T(Z_i)$ may vary over time in Eq. (2).

Note that we may have $T \in \text{Pa}_T(Z_i)$ for some $Z_i \in \boldsymbol{Z}$. Let $\boldsymbol{R} \subseteq \boldsymbol{Z}$ correspond to all those variables in $\boldsymbol{Z}$ where T is not in the parent set, so $T \notin \text{Pa}_T(Z_i)$ for all $Z_i \in \boldsymbol{R}$. We also have $T \in \text{Pa}_T(Z_i)$ for all $Z_i \in [\boldsymbol{Z} \setminus \boldsymbol{R}]$. We can then rewrite Eq. (2):

$$\begin{aligned} &\prod_{i=1}^{p+1} p(Z_i|\text{Pa}_T(Z_i)) \\ &\quad = \prod_{Z_i \in \boldsymbol{R}} p(Z_i|\text{Pa}_T(Z_i) \setminus T) \prod_{Z_i \in [\boldsymbol{Z} \setminus \boldsymbol{R}]} p(Z_i|\text{Pa}_T(Z_i) \cup T). \end{aligned} \tag{3}$$

The left hand term corresponds to the stationary component and the right hand to the non-stationary component. We assume that we can sample i.i.d. from the density $p(\boldsymbol{O}|\boldsymbol{S})$:

$$p(\boldsymbol{O}|\boldsymbol{S}) = \sum_{\boldsymbol{L}} p(\boldsymbol{O}, \boldsymbol{L}|\boldsymbol{S}),$$

where mixing occurs over time T in the integration if $T \in \boldsymbol{L}$. We technically do not require $T \in \boldsymbol{L}$, but we refer to the above equation as the *mixture of DAGs* framework because we usually have $T \in \boldsymbol{L}$ in practice.

4.2 Multivariate case

We generalize the mixture of DAGs framework to a multivariate set of mutually independent variables $\boldsymbol{T}$ that may include variables other than time. This step is critical for modeling sparse graphical structure and many independent causes of change. For example, we may let $\boldsymbol{T} = \{T_1, T_2\}$, where T_1 denotes time and T_2 gender. Gender is instantiated independent of time, but the causal process can change over time and differ by gender. The set $\boldsymbol{T}$ can encompass a wide range of variables and will allow the DAG to change according to multiple conditions such as time, location and sub-populations. In contrast, methods like dynamic Bayesian networks and dynamic structural causal models only accommodate changes across a single variable – typically time.

We consider the set of vertices $\boldsymbol{Z} = \boldsymbol{X} \cup \boldsymbol{T}$ instead of the original $\boldsymbol{X} \cup T$. We divide $\boldsymbol{Z}$ into three non-overlapping sets $\boldsymbol{O}$, $\boldsymbol{L}$ and $\boldsymbol{S}$. We assume a joint density $p(\boldsymbol{X}, \boldsymbol{T})$ that factorizes according to a DAG $\mathbb{G}_{\boldsymbol{T}}$ over $\boldsymbol{Z}$:

$$\begin{aligned} p(\boldsymbol{Z}) = p(\boldsymbol{T})p(\boldsymbol{X}|\boldsymbol{T}) &= \prod_{i=1}^{s} p(T_i) \prod_{i=1}^{p} p(X_i|\mathrm{Pa}_{\boldsymbol{T}}(X_i)) \\ &= \prod_{i=1}^{p+s} p(Z_i|\mathrm{Pa}_{\boldsymbol{T}}(Z_i)), \end{aligned} \tag{4}$$

where $\mathrm{Pa}_{\boldsymbol{T}}(\boldsymbol{T}) = \emptyset$. The above equation mirrors Eq. (2).

Note that we may have $\boldsymbol{T} \cap \mathrm{Pa}_{\boldsymbol{T}}(Z_i) \neq \emptyset$ for some $Z_i \in \boldsymbol{Z}$. So for each $Z_i \in \boldsymbol{Z}$, let $\boldsymbol{U}_i \subseteq \boldsymbol{T}$ denote the largest set such that $\boldsymbol{U}_i \cap \mathrm{Pa}_{\boldsymbol{T}}(Z_i) = \emptyset$. This implies $\boldsymbol{T} \cap \mathrm{Pa}_{\boldsymbol{T}}(Z_i) = \boldsymbol{T} \setminus \boldsymbol{U}_i \triangleq \boldsymbol{V}_i$. We then rewrite Eq. (4):

$$\prod_{i=1}^{p+s} p(Z_i|\mathrm{Pa}_{\boldsymbol{T}}(Z_i)) = \prod_{i=1}^{p+s} p(Z_i|(\mathrm{Pa}_{\boldsymbol{T}}(Z_i) \setminus \boldsymbol{U}_i) \cup \boldsymbol{V}_i), \tag{5}$$

so that $p(Z_i|\mathrm{Pa}_{\boldsymbol{T}}(Z_i))$ is stationary over $\boldsymbol{U}_i$ but non-stationary over $\boldsymbol{V}_i$. Setting $\boldsymbol{U}_i = T$ and $\boldsymbol{V}_i = \emptyset$ for $Z_i \in \boldsymbol{R}$ and vice versa for $Z_i \in [\boldsymbol{Z} \setminus \boldsymbol{R}]$ recovers Eq. (3). We finally sample i.i.d. from the density $p(\boldsymbol{O}|\boldsymbol{S})$:

$$p(\boldsymbol{O}|\boldsymbol{S}) = \sum_{\boldsymbol{L}} p(\boldsymbol{O}, \boldsymbol{L}|\boldsymbol{S}). \tag{6}$$

where mixing occurs over $\boldsymbol{T} \cap \boldsymbol{L}$ if $\boldsymbol{T} \cap \boldsymbol{L} \neq \emptyset$. We again technically do not require $\boldsymbol{T} \cap \boldsymbol{L} \neq \emptyset$, but this usually holds in practice.

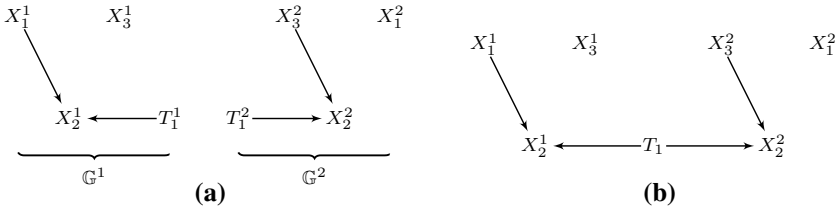


Fig. 3 Construction of a mixed graph. We plot the two DAGs in $\mathcal{G}$ next to each other in (**a**) for Step 1 of Definition 1. Merging the two vertices in T_1' to create T_1 generates $\mathbb{M}$ in (**b**) for Step 2

4.3 Global Markov property

The factorization in Eq. (5) implies certain CI relations. In this section, we will identify the CI relations by deriving a global Markov property similar to the traditional DAG case.

There exists a DAG $\mathbb{G}_{\boldsymbol{T}}$ for each instantiation of $\boldsymbol{T}$ because $\text{Pa}_{\boldsymbol{T}}(Z_i)$ is defined for all $Z_i \in \boldsymbol{Z}$. Consider the collection $\mathcal{G}$ consisting of all DAGs indexed by $\boldsymbol{T}$. The number of DAGs over $\boldsymbol{Z}$ is finite, so $|\mathcal{G}| = q \in \mathbb{N}^+$. Let $\mathcal{T}$ denote the set of all values of $\boldsymbol{T}$ corresponding to members of $\mathcal{G}$, and $\mathcal{T}^j \subseteq \mathcal{T}$ to the set for $\mathbb{G}^j \in \mathcal{G}$. We can then rewrite Eq. (5) as:

$$\prod_{i=1}^{p+s} p(Z_i|\text{Pa}_{\boldsymbol{T}}(Z_i)) = \sum_{j=1}^{q} \mathbb{1}_{\boldsymbol{T}\in\mathcal{T}^j} \prod_{i=1}^{p+s} p(Z_i|\text{Pa}_{\mathbb{G}^j}(Z_i)). \tag{7}$$

We want to find a *single* graph where d-separation between the vertices implies CI in the density that factorizes according to Eq. (7). Clearly, we need to combine the graphs in $\mathcal{G}$ using some procedure. We use the notation $\boldsymbol{A}^j$ to refer to the set of vertices $\boldsymbol{A} \subseteq \boldsymbol{Z}$ associated with $\mathbb{G}^j$ in the resultant graph. We also let $\boldsymbol{A}' = \cup_{j=1}^{q} \boldsymbol{A}^j$ denote the corresponding collection across all DAGs in $\mathcal{G}$. It turns out that the following combination of graphs in $\mathcal{G}$ suffices:

Definition 1 (*Mixture graph*) The *mixture graph* $\mathbb{M}$ is a DAG constructed by combining the graphs in $\mathcal{G}$ using the following procedure:

1. Plot each of the q DAGs in $\mathcal{G}$ adjacent to each other.
2. Merge $T_i' \subseteq \boldsymbol{T}'$ into a single vertex T_i for each $T_i \in \boldsymbol{T}$.

Notice therefore that the DAGs in $\mathcal{G}$ are connected by $\boldsymbol{T}$ in $\mathbb{M}$, so they are statistically dependent in general. We provide an example in Fig. 3. Figure 3a corresponds to the two DAGs in $\mathcal{G}$ plotted next to each other according to Step 1 of Definition 1. We then merge the two vertices in T_1' into a single vertex T_1 according to Step 2 to yield $\mathbb{M}$ in Fig. 3b.

If $\boldsymbol{A} \subseteq \boldsymbol{T}$, then $\boldsymbol{A}' = \boldsymbol{A}$ in $\mathbb{M}$ due to Step 2 above. We can now read off the implied CI relations from $\mathbb{M}$ by utilizing d-separation across groups of vertices rather than just singletons.

Theorem 1 (Global Markov property) *Let* $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ *denote disjoint subsets of* $\boldsymbol{Z}$. *If* $\boldsymbol{A}' \perp\!\!\!\perp_d \boldsymbol{B}' | \boldsymbol{C}'$ *in* $\mathbb{M}$, *then* $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} | \boldsymbol{C}$.

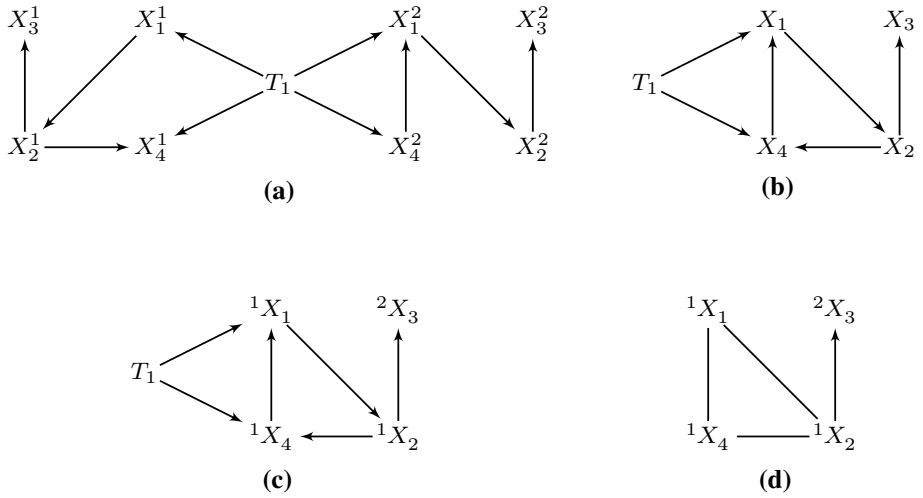


Fig. 4 We have the mixture graph in (**a**) and the fused graph in (**b**). Subfigures (**c**) and (**d**) contain $\mathbb{F}$ and $\mathbb{F}^*$, respectively, with additional time step information

We refer to the reverse direction as d-separation faithfulness with respect to (w.r.t.) $\mathbb{M}$. The result improves upon that of (Spirtes 1994) and (Saeed et al. 2020), where the authors additionally assumed that $\boldsymbol{T}$ is univariate, latent and discrete. Spirtes (1994) also proposed another global Markov property that implies less CI relations even under the univariate assumption (see Definition 2 and then Appendix 8.3 for a detailed discussion).

We provide an example again in Fig. 3. In Fig. 3a, we have $X_1^1 \to X_2^1$ in the first DAG and $X_2^2 \leftarrow X_3^2$ in the second; however, we do not have the v-structure $X_1^j \to X_2^j \leftarrow X_3^j$ in either DAG. We also have the relation $X_1' \perp\!\!\!\perp_d X_3'$ in Fig. 3b, so $\mathbb{M}$ implies $X_1 \perp\!\!\!\perp X_3$ per Theorem 1. In contrast, $X_1' \not\perp\!\!\!\perp_d X_3' | X_2'$ in Fig. 3b, so $\mathbb{M}$ implies $X_1 \not\perp\!\!\!\perp X_3 | X_2$ per d-separation faithfulness w.r.t. $\mathbb{M}$. This holds even though we have $X_1^j \perp\!\!\!\perp_d X_3^j | X_2^j$ in either DAG in Fig. 3a. Variables may therefore be conditionally dependent in the mixture distribution, even though they are d-separated within any single DAG in $\mathcal{G}$, because the variables are connected by $\boldsymbol{T}$ in $\mathbb{M}$.

5 Causal inference over mixtures

5.1 Fused graph

The mixture graph encodes the global Markov property, but we cannot easily visualize cycles in $\mathbb{M}$ because they are spread across different sub-DAGs. We therefore construct a *fused graph*, first introduced in (Spirtes 1994), that contains cycles but does not necessarily encode the best global Markov property.

Definition 2 The fused graph $\mathbb{F}$ is a directed graph (potentially cyclic) constructed by merging the graphs in $\mathcal{G}$. In other words:

1. Plot each of the q DAGs in $\mathcal{G}$ adjacent to each other.
2. Merge $Z'_i \subseteq \boldsymbol{Z}'$ into a single vertex Z_i for each $Z_i \in \boldsymbol{Z}$, so that $\mathbb{F}$ may contain cycles.

Intuitively, the fused graph combines each set of vertices Z'_i in $\mathbb{M}$ into a single vertex Z_i. $\mathbb{F}$ therefore summarizes cycles in one directed graph, so it is more intuitive than $\mathbb{M}$, where cycles are spread across multiple sub-graphs.

We provide an example of a mixture graph and its associated fused graph in Fig. 4. We focus on the mixture graph in Fig. 4a, where we have a cycle involving $\{X_1, X_2, X_4\}$, but we do not observe the full cycle in either sub-DAG. We have also drawn out $\mathbb{F}$ in Fig. 4b. X_2 is an ancestor of X_1 in $\mathbb{F}$ even though X'_2 is not an ancestor of X'_1 in $\mathbb{M}$.

We will utilize the global Markov property of $\mathbb{M}$ in order to recover (parts of) a mixed graph $\mathbb{F}^*$ summarizing the ancestral relations in $\mathbb{F}$, because $\mathbb{F}^*$ allows us to visualize cycles that are not present within $\mathbb{M}$ but exist once the DAGs are combined in $\mathbb{F}$. This differs from the work in (Spirtes 1994), where the author proposed to use a global Markov property based directly on $\mathbb{F}$. This property unfortunately implies less CI relations even for univariate $\boldsymbol{T}$, so we cannot use it to infer as many ancestral relations in $\mathbb{F}$ as compared to the proposed global Markov property based on $\mathbb{M}$ (details to come in Sect. 5.4).

5.2 Longitudinal data

We have unfortunately identified an instance, where it is impossible to even detect a v-structure under acyclicity using a CI oracle alone (Appendix 8.4). We therefore rely on additional information to orient arrowheads (which encode non-ancestral relations) using *longitudinal data*, where we assume access to multiple *time steps* of variables.[2] Note that we differentiate between discrete time *steps* and discrete or continuous time *points* because each time step may include a mixture of different time points corresponding to instantiations of a time variable in $\boldsymbol{T}$. We assume access to w time steps, so that we can partition $\boldsymbol{O}$ into w disjoint subsets denoted by ${}^1\boldsymbol{O}, \ldots, {}^w\boldsymbol{O}$. We thus have $\cup_{k=1}^{w} {}^k\boldsymbol{O} = \boldsymbol{O}$. We then consider the following density:

Definition 3 (*Longitudinal density*) A longitudinal density is a density $p(\cup_{k=1}^{w} {}^k\boldsymbol{O}, \boldsymbol{L}, \boldsymbol{S})$ that factorizes according to Eq. (5) such that no variable in time step a is an ancestor of a variable in time step $b < a$ and $w \geq 2$.

Causation proceeds forward in time, so no variable in time step a can be an ancestor of a variable in time step $b < a$.

It is important to take some time and decompose Definition 3, because it can be confused with some other concepts in causal discovery, such as those used in dynamic Bayesian networks or equilibrium distributions. The set $\boldsymbol{Z} = {}^1\boldsymbol{O} \cup \boldsymbol{L} \cup \boldsymbol{S}$ with $w = 1$ corresponds to a standard variable set in causal discovery, where we assume access to only one time step for the observable variables. Under the mixture of DAGs framework, the density of $\boldsymbol{Z}$ factorizes as $\prod_{i=1}^{p+s} p\left(Z_i | \text{Pa}_{\boldsymbol{T}}(Z_i)\right)$ just like in Eq. (5). Mixing over DAGs then occurs according to $\boldsymbol{T} \cap \boldsymbol{L}$ with $\sum_{\boldsymbol{L}} p({}^1\boldsymbol{O}, \boldsymbol{L} | \boldsymbol{S}) = p({}^1\boldsymbol{O} | \boldsymbol{S})$ as in Eq. (6). The time *step* $w = 1$ therefore can

[2] Time steps are also commonly known as *waves* in the applied literature (Taris 2000).

Table 1 An example of a longitudinal dataset containing three time steps, 9 variables in $\boldsymbol{O}$ and two DAGs in $\mathcal{G}$. The light blue samples correspond to $\mathbb{G}^1$ and the grey samples to $\mathbb{G}^2$. Notice that each time step is still a mixture of DAGs

Time Step 1			Time Step 2			Time Step 3		
${}^1O_1$	${}^1O_2$	${}^1O_3$	${}^2O_4$	${}^2O_5$	${}^2O_6$	${}^3O_7$	${}^3O_8$	${}^3O_9$
2.28	−1.27	1.61	0.15	30.18	0.25	0.39	−0.27	1.80
0.48	−0.59	0.23	0.55	−0.45	0.58	0.05	−0.41	0.25
0.16	−0.82	0.13	0.61	−0.26	0.41	0.12	−0.59	0.57
0.61	−1.18	0.70	1.38	−1.28	0.05	1.38	−0.36	0.91
1.10	−0.42	1.14	0.24	−2.22	0.25	0.76	−1.24	0.54
0.76	−0.47	0.66	0.34	−0.97	1.10	0.78	−1.06	2.47
0.10	−0.37	1.01	0.12	−0.33	0.98	0.35	−0.09	0.40
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

include a mixture of DAGs representing cycles as well as different time *points*, sub-populations and distributions indexed by $\boldsymbol{T}$.

With longitudinal data, we consider multiple time steps ${}^1\boldsymbol{O}, \ldots, {}^w\boldsymbol{O}$ and consider the set $\boldsymbol{Z} = \cup_{k=1}^{w} {}^k\boldsymbol{O} \cup \boldsymbol{L} \cup \boldsymbol{S}$ with $w \geq 2$. We then assume that the entire joint density factorizes according to a mixture of DAGs just like in Eq. (5): $\prod_{i=1}^{p+s} f\left(Z_i | \text{Pa}_{\boldsymbol{T}}(Z_i)\right)$. We of course require the additional constraint that ${}^aO_i \in \text{Pa}_{\boldsymbol{T}}({}^bO_j)$ implies $a \leq b$ as indicated in Definition 3 because causation must proceed forward in time. Mixing over DAGs then occurs over $\boldsymbol{T} \cap \boldsymbol{L}$ such that $\sum_{\boldsymbol{L}} p(\cup_{k=1}^{w} {}^k\boldsymbol{O}, \boldsymbol{L} | \boldsymbol{S}) = p(\cup_{k=1}^{w} {}^k\boldsymbol{O} | \boldsymbol{S})$. Thus, although we partition the variables across time steps, each time step can still include a mixture of DAGs representing cycles as well as different time points, sub-populations and distributions indexed by $\boldsymbol{T}$ just like the original case where $w = 1$. The time steps are also statistically dependent in general because the factorization in Eq. (5) includes variables in different time steps. Our setup is similar to the problem discussed in (Rubenstein et al. 2018), where they highlight the indeterminacies that arise when discretizing a continuous time causal model into a few discrete time steps because each discrete time step can contain samples from multiple models.

We provide an example of a longitudinal dataset in Table 1. The dataset is derived using 2 DAGs in $\mathcal{G}$, 9 variables in $\boldsymbol{O}$ and three time steps. Each time step contains blue and grey rows corresponding to samples obeying either the first or second DAG, respectively. The variables are statistically dependent between time steps in general and do not contain missing values. Further observe that O_i^k in $\mathbb{M}$ is *not* equivalent to ${}^kO_i \in {}^k\boldsymbol{O}$; the first notation refers to O_i in $\mathbb{G}^k \in \mathcal{G}$, while the other refers to O_i measured at time step k which may arise from any graph in $\mathcal{G}$ because each time step is a mixture of DAGs; the pre-super script and the post-super script therefore denote different concepts.

5.3 Output target

If $\boldsymbol{Y} \subseteq \boldsymbol{O}$, then let ${}^a\boldsymbol{Y}$ and ${}^a\boldsymbol{Y}'$ denote $\boldsymbol{Y} \cap {}^a\boldsymbol{O}$ and $[{}^a\boldsymbol{Y}]'$, respectively. We write ${}_d^c\text{Adj}_{\mathbb{F}^*}({}^aO_i)$ to mean those variables between time steps c and d inclusive that are adjacent to ${}^aO_i$ in $\mathbb{F}^*$. We will specifically construct $\mathbb{F}^*$ with the following adjacencies:

List 1 (Adjacency Interpretations)

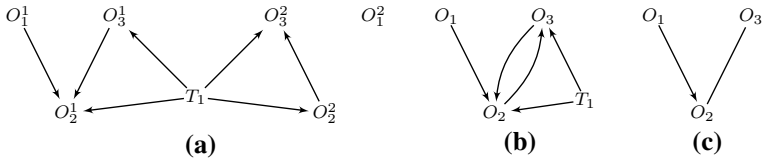


Fig. 5 An example where both FCI and CCI fail. We have a mixture graph in (**a**) and its fused graph in (**b**). Subfigure (**c**) contains the correct $\mathbb{F}^*$, but FCI and CCI infer the incorrect collider $O_1 \ast\!\!\rightarrow O_2 \leftrightarrow\!\!\ast\, O_3$

1. If we have ${}^aO_i \ast\!\!-\!\!\ast\, {}^bO_j$ (with possibly $a = b$), then ${}^aO'_i \not\perp\!\!\!\perp_d {}^bO'_j | \boldsymbol{W}' \cup \boldsymbol{S}'$ in $\mathbb{M}$ for all $\boldsymbol{W} \subseteq {}^a_b\text{Adj}_{\mathbb{F}^*}({}^aO_i) \setminus {}^bO_j$ and all $\boldsymbol{W} \subseteq {}^a_b\text{Adj}_{\mathbb{F}^*}({}^bO_j) \setminus {}^aO_i$.
2. If we do not have ${}^aO_i \ast\!\!-\!\!\ast\, {}^bO_j$ (with possibly $a = b$), then ${}^aO'_i \perp\!\!\!\perp_d {}^bO'_j | \boldsymbol{W}' \cup \boldsymbol{S}'$ in $\mathbb{M}$ for some $\boldsymbol{W} \subseteq \boldsymbol{O} \setminus \{{}^aO_i, {}^bO_j\}$.

The endpoints of $\mathbb{F}^*$ have the following modified interpretations:

List 2 (Endpoint Interpretations)

1. If ${}^aO_i \ast\!\!\rightarrow {}^bO_j$, then ${}^bO_j \notin \text{Anc}_{\mathbb{F}}({}^aO_i)$; in other words, merging the graphs in $\mathcal{G}$ does not create a directed path from ${}^bO_j$ to ${}^aO_i$.
2. If ${}^aO_i \ast\!\!-\!\!{}^bO_j$, then ${}^bO_j \in \text{Anc}_{\mathbb{F}}({}^aO_i \cup \boldsymbol{S})$; in other words, merging the graphs in $\mathcal{G}$ creates a directed path from ${}^bO_j$ to ${}^aO_i \cup \boldsymbol{S}$.

The arrowheads do not take into account selection variables because we often cannot a priori specify whether a variable is an ancestor of $\boldsymbol{S}$ in $\mathbb{F}$ using either time step information or other prior knowledge in practice. We re-emphasize that O_i^k in $\mathbb{M}$ is not equivalent to ${}^kO_i$ in $\mathbb{F}$. We draw an example of $\mathbb{M}$ in Fig. 4a, its fused graph $\mathbb{F}$ *with time step notation* in Fig. 4c and the corresponding mixed graph $\mathbb{F}^*$ in Fig. 4d, where ${}^1\boldsymbol{O} = \{{}^1X_1, {}^1X_2, {}^1X_4\}$, ${}^2\boldsymbol{O} = {}^2X_3$, $\boldsymbol{L} = T_1$, $\boldsymbol{S} = \emptyset$ and $w = 2$.

5.4 Algorithm

We cannot apply an existing constraint-based algorithm like FCI on data arising from a mixture of DAGs and expect to recover a partially oriented $\mathbb{F}^*$. For example, FCI and CCI can make incorrect inferences if $\mathcal{G}$ contains more than one DAG. Consider the mixture graph in Fig. 5a, where all variables lie in the same time step. O_2 is an ancestor of O_3 in $\mathbb{F}$ drawn in Fig. 5b, but we have $O'_1 \perp\!\!\!\perp_d O'_3$ in $\mathbb{M}$, so O_1 and O_3 are independent by Theorem 1. FCI and CCI therefore infer the incorrect collider $O_1 \ast\!\!\rightarrow O_2 \leftrightarrow\!\!\ast\, O_3$ in $\mathbb{F}^*$ during v-structure discovery. We thus require an alternative algorithm to correctly recover a partially oriented $\mathbb{F}^*$.

We now propose a new algorithm called Causal Inference over Mixtures (CIM) which correctly recovers causal relations. We summarize the procedure in Algorithm 1. The CIM algorithm works as follows. First, CIM runs a variant of PC-stable's skeleton discovery procedure in order to discover adjacencies as well as minimal separating sets in Step 1 (Colombo and Maathuis 2014). This step is summarized in Algorithm 2. The skeleton discovery procedure attempts to find a minimal set that renders ${}^aO_i$ and ${}^bO_j$ conditionally

independent using variables adjacent to ${}^aO_i$ and between time steps a and b inclusive in Lines 8 and 9. If the algorithm succeeds in doing so, then it removes the edge between ${}^aO_i$ and ${}^bO_j$ in Line 10. Algorithm 2 therefore recovers the adjacencies with interpretations listed in List 1. The algorithm stores the minimal separating sets in the array Sep in Line 11 so that $\text{Sep}({}^aO_i, {}^bO_j)$ contains a minimal separating set of ${}^aO_i$ and ${}^bO_j$, if such a set exists.

Algorithm 1: Causal Inference over Mixtures (CIM)

Data: CI oracle, time steps $\mathcal{W}$, other prior information $\mathcal{P}$
Result: partially oriented mixed graph $\widehat{\mathbb{F}}^*$

1 Run Algorithm 2, a variant of PC-stable's skeleton discovery procedure.
2 If we have ${}^aO_i \ast\!\!-\!\!\circ {}^bO_j$ and $a < b$ according to $\mathcal{W}$, or ${}^bO_j$ cannot be an ancestor of ${}^aO_i$ according to $\mathcal{P}$, then orient ${}^aO_i \ast\!\!-\!\!\circ {}^bO_j$ as ${}^aO_i \ast\!\!\rightarrow {}^bO_j$ in $\widehat{\mathbb{F}}^*$.
3 If we have ${}^aO_i \ast\!\!\rightarrow {}^bO_j \ast\!\!-\!\!\ast {}^cO_k$ with ${}^aO_i$ and ${}^cO_k$ non-adjacent, ${}^bO_j \notin \text{Sep}({}^aO_i, {}^cO_k)$ and there exists another minimal separating set $\boldsymbol{W} \subseteq {}^a_c\text{Adj}_{\widehat{\mathbb{F}}^*}({}^aO_i) \setminus {}^cO_k$ or $\boldsymbol{W} \subseteq {}^a_c\text{Adj}_{\widehat{\mathbb{F}}^*}({}^cO_k) \setminus {}^aO_i$ containing ${}^bO_j$, then record $\boldsymbol{W}$ into $\text{Sep2}({}^aO_i, {}^bO_j, {}^cO_k)$.
4 If we have ${}^aO_i \ast\!\!\rightarrow {}^bO_j \circ\!\!-\!\!\ast {}^cO_k$ with ${}^aO_i$ and ${}^cO_k$ non-adjacent, and either ${}^bO_j \in \text{Sep}({}^aO_i, {}^cO_k)$ or $\text{Sep2}({}^aO_i, {}^bO_j, {}^cO_k)$ is non-empty, then orient ${}^bO_j \circ\!\!-\!\!\ast {}^cO_k$ as ${}^bO_j -\!\!\ast {}^cO_k$ in $\widehat{\mathbb{F}}^*$.
5 Execute the following orientation rule until no more edges can be oriented: if we have the sequence of vertices $\langle {}^{a_1}O_1, \ldots, {}^{a_n}O_n \rangle$ such that ${}^{a_i}O_i -\!\!\ast {}^{a_{i+1}}O_{i+1}$ with $1 \le i \le n-1$, and we have ${}^{a_1}O_1 \circ\!\!-\!\!\ast {}^{a_n}O_n$, then orient ${}^{a_1}O_1 \circ\!\!-\!\!\ast {}^{a_n}O_n$ as ${}^{a_1}O_1 -\!\!\ast {}^{a_n}O_n$ in $\widehat{\mathbb{F}}^*$.

Algorithm 2: CIM's skeleton discovery procedure

Data: CI oracle, $\mathcal{W}$
Result: $\widehat{\mathbb{F}}^*$, Sep

1 Form a complete graph $\widehat{\mathbb{F}}^*$ over $\boldsymbol{O}$ with edges $\circ\!\!-\!\!\circ$
2 $l \leftarrow -1$
3 **repeat**
4 $l = l + 1$
5 **repeat**
6 Select a new ordered pair of vertices $({}^aO_i, {}^bO_j)$ that are adjacent in $\widehat{\mathbb{F}}^*$ and satisfy $|{}^a_b\text{Adj}_{\widehat{\mathbb{F}}^*}({}^aO_i) \setminus {}^bO_j| \ge l$
7 **repeat**
8 Choose a new set $\boldsymbol{W} \subseteq {}^a_b\text{Adj}_{\widehat{\mathbb{F}}^*}({}^aO_i) \setminus {}^bO_j$ with $|\boldsymbol{W}| = l$
9 **if** ${}^aO_i \perp\!\!\!\perp {}^bO_j | \boldsymbol{W} \cup \boldsymbol{S}$ **then**
10 Delete the edge ${}^aO_i \circ\!\!-\!\!\circ {}^bO_j$ from $\widehat{\mathbb{F}}^*$
11 $\text{Sep}({}^aO_i, {}^bO_j) \leftarrow \text{Sep}({}^bO_j, {}^aO_i) \leftarrow \boldsymbol{W}$
12 **end**
13 **until** *${}^aO_i$ and ${}^bO_j$ are no longer adjacent in $\widehat{\mathbb{F}}^*$ or all $\boldsymbol{W} \subseteq {}^a_b\text{Adj}_{\widehat{\mathbb{F}}^*}({}^aO_i) \setminus {}^bO_j$ with $|\boldsymbol{W}| = l$ have been considered*;
14 **until** *all ordered pairs of adjacent vertices $({}^aO_i, {}^bO_j)$ in $\widehat{\mathbb{F}}^*$ with $|{}^a_b\text{Adj}_{\widehat{\mathbb{F}}^*}({}^aO_i) \setminus {}^bO_j| \ge l$ have been considered*;
15 **until** *all pairs of adjacent vertices $({}^aO_i, {}^bO_j)$ in $\widehat{\mathbb{F}}^*$ satisfy $|{}^a_b\text{Adj}_{\widehat{\mathbb{F}}^*}({}^aO_i) \setminus {}^bO_j| \le l$*;

CIM next adds arrowheads in Step 2 using time step information from a longitudinal dataset with the list $\mathcal{W}$. If we have ${}^aO_i \circ\!\!-\!\!\circ {}^bO_j$ with $a < b$, then CIM orients ${}^aO_i \circ\!\!\rightarrow {}^bO_j$ because ${}^bO_j \notin \text{Anc}_{\mathbb{F}}({}^aO_i)$ according to List 2. We can orient additional arrowheads using other prior knowledge $\mathcal{P}$. Step 2 orients many arrowheads in practice, so long as we have at least two time steps of data.

For every triple ${}^aO_i \ast\!\!\rightarrow {}^bO_j \ast\!\!-\!\!\ast {}^cO_k$ with ${}^aO_i$ and ${}^cO_k$ non-adjacent, CIM then attempts to find a minimal separating set that contains ${}^bO_j$ in Step 3. These sets are important due to the following lemma which allows us to infer tails in Step 4:

Lemma 1 *Suppose* ${}^aO'_i \perp\!\!\!\perp_d {}^bO'_j | \boldsymbol{W}' \cup \boldsymbol{S}'$ *in* $\mathbb{M}$ *but* ${}^aO'_i \not\perp\!\!\!\perp_d {}^bO'_j | \boldsymbol{V}' \cup \boldsymbol{S}'$ *for every* $\boldsymbol{V} \subset \boldsymbol{W}$. *If* ${}^cO_k \in \boldsymbol{W}$, *then* ${}^cO_k \in \text{Anc}_{\mathbb{F}}({}^aO_i \cup {}^bO_j \cup \boldsymbol{S})$.

Theorem 1 therefore allows us to infer more ancestral relations via the above lemma – as compared to the global directed Markov property based directly on the fused graph $\mathbb{F}$ (Spirtes 1994) – because $\mathbb{M}$ has more d-separation relations than $\mathbb{F}$. CIM finally adds some additional tails in Step 5 due to transitivity of the tails.

We now formally claim that Algorithm 1 is sound:

Theorem 2 *Suppose the longitudinal density* $p(\cup_{k=1}^{w} {}^k\boldsymbol{O}, \boldsymbol{L}, \boldsymbol{S})$ *factorizes according to Eq.* (5). *Assume that all prior information* $\mathcal{P}$ *is correct. Then, under d-separation faithfulness w.r.t.* $\mathbb{M}$, *the CIM algorithm returns* $\widehat{\mathbb{F}}^*$ *– the mixed graph* $\mathbb{F}^*$ *partially oriented.*

Thus if ${}^aO_i *\!\!-\!\!* {}^bO_j$ for any two vertices in $\widehat{\mathbb{F}}^*$, then ${}^bO_j \in \text{Anc}_{\mathbb{F}}({}^aO_i \cup \boldsymbol{S})$; in other words, merging the graphs in $\mathcal{G}$ creates a directed path from ${}^bO_j$ to ${}^aO_i \cup \boldsymbol{S}$ per List 2. Moreover, CIM completes in $O(r^s)$ time where r denotes the number of variables in $\cup_{k=1}^{w} {}^k\boldsymbol{O}$ and s the maximum number of vertices adjacent to any vertex in $\mathbb{F}^*$ due to Steps 1 and 3 of Algorithm 1. We can therefore predict that CIM will take about the same amount of time to complete as PC.

6 Experiments

We had two overarching goals: (1) evaluate the performance of CIM against other constraint-based algorithms using real data, and (2) determine if we can reconstruct the real data results using synthetic data sampled from a mixture of DAGs. We utilized the setup described below.

6.1 Algorithms

CIM is a constraint-based algorithm that executes CI tests in greedy sequence. We therefore compared CIM against similar greedy constraint-based algorithms in recovering the ancestral relations in $\mathbb{F}$: PC, FCI, RFCI and CCI. FCI covers the recent proposal by Saeed et al. (2020). We equipped all algorithms with a nonparametric CI test called GCM (Shah et al. 2020) and fixed $\alpha = 0.01$ across all experiments. We gave all algorithms the same time step information during skeleton discovery in order to orient arrowheads between the time steps. The algorithms perform much worse without the additional knowledge. As a result, we more specifically compared CIM against the time series versions of the algorithms (Entner and Hoyer 2010; Malinsky and Spirtes 2018; Runge et al. 2019).

6.1.1 Metrics

Let tails refer to positives and arrowheads to negatives. Recall that the output of CIM $\widehat{\mathbb{F}}^*$ includes arrowheads and tails, but the arrowheads are oriented by time steps and prior knowledge according to Step 2. CIM therefore only infers tails using CI tests.

Since CIM only infers tails, we compared the algorithms on their ability to infer ancestral relations according to List 2. We specifically evaluated the algorithms using *sensitivity* and *fallout*. The sensitivity is defined as TP/P, where TP refers to true positives and P to positives. The fallout is defined as FP/N, where FP refers to false positives and N to negatives. A tail in place of an arrowhead corresponds to a false positive.

The receiver operating characteristic (ROC) curve plots sensitivity against the fallout. If an algorithm does not orient any tails, then the sensitivity is zero. On the other hand, if an algorithm just orients all tails, then the fallout is one. If an algorithm achieves a perfect balance by orienting all true tails as tails and no true arrowheads as tails, then sensitivity is one and the fallout is zero. Perfect accuracy therefore corresponds to a sensitivity of one and a fallout of zero at the upper left hand corner of the ROC curve. Constraint-based algorithms do not output a continuous score required to compute the area under the ROC curve, but we can assess *overall performance* using the Euclidean distance from the upper left hand corner (Perkins and Schisterman 2006).

6.2 Real data

6.2.1 Framingham heart study

We first evaluated the algorithms on real data. We considered the Framingham Heart Study (FHS), where investigators measured cardiovascular changes across time in residents of Framingham, Massachusetts (Mahmood et al. 2014). The dataset contains three time steps of data with 8 variables in each time step. We obtained 2019 samples after removing patients with missing values.

The dataset contains the following known direct causal relations: (1) number of cigarettes per day causes heart rate via cardiac nicotonic acetylcholine receptors (Aronow et al. 1971; Levy 1971; Haass and Kübler 1997; 2) age causes systolic blood pressure due to increased large artery stiffness (Pinto 2007; Safar 2005; 3) age causes cholesterol levels due to changes in cholesterol and lipoprotein metabolism (Parini et al. 1999; 4) BMI causes number of cigarettes per day because smoking cigarettes is a common weight loss strategy (Jo et al. 2002; Chiolero et al. 2008; 5) systolic blood pressure causes diastolic blood pressure and vice versa by definition, because both quantities refer to pressure in the same arteries at different points in time. We can compute sensitivity using this information.

We summarize the results over 50 bootstrapped datasets in Fig. 6a, b, c. We first evaluated sensitivity by running the algorithms using the full time step information. RFCI, FCI and CCI oriented few tails overall, so they obtained lower sensitivity scores (Fig. 6a). PC and CIM had similar sensitivities (t=-0.80, p=0.43). We next combined time steps 2 and 3, so that the algorithms could incorrectly orient tails backwards in time. CIM made fewer errors than PC as indicated by a lower fallout (Fig. 6b, t=-11.85, p=5.37E-16). FCI, RFCI and CCI also achieved low fallout scores, but they again did not orient many tails to begin with. CIM therefore obtained the best overall score when we combined sensitivity and fallout (Fig. 6c, t=-5.60, p=9.70E-7). Timing results in Fig. 7a finally indicate that CIM takes about the same amount of time to complete as the fastest algorithms, PC and RFCI, as predicted by the complexity analysis in Sect. 5.4. We conclude that both CIM and PC orient many tails, but CIM makes fewer errors as evidenced by its high sensitivity and low fallout. We therefore prefer CIM in this dataset.

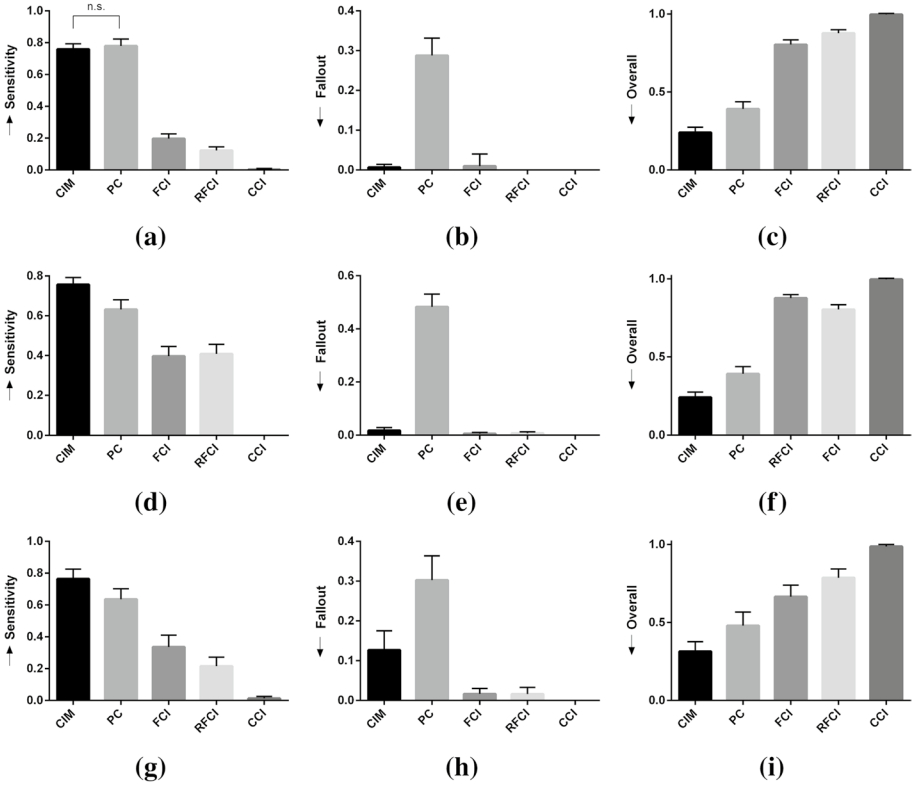


Fig. 6 Results for FHS in (**a**, **b**, **c**), STAR*D in (**d**, **e**, **f**) and the synthetic data in (**g**, **h**, **i**). Bar heights represent empirical means and error bars their 95% confidence intervals. An up-pointing arrow means higher is better and a down-pointing arrow means lower is better. CIM achieves higher sensitivity in (**a**, **d**, **g**) while maintaining a low fallout in (**b**, **e**, **h**). CIM performs the best overall in all cases as shown in (**c**, **f**, **i**)

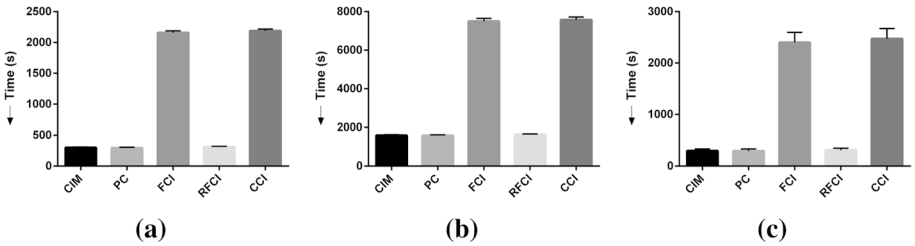


Fig. 7 Timing results for FHS in (**a**), STAR*D in (**b**) and the synthetic data in (**c**). CIM completes in about the same amount of time as PC and RFCI

6.2.2 Sequenced treatment alternatives to relieve depression trial

We next analyzed Level 1 of the Sequenced Treatment Alternatives to Relieve Depression (STAR D) trial (Sinyor et al. 2010). Investigators gave patients an antidepressant called citalopram and then tracked their depression symptoms using a standardized questionnaire

called QIDS-SR-16. We analyzed the 9 QIDS-SR-16 sub-scores measuring components of depression at weeks 0, 2 and 4. We also included age and gender in the first time step. The dataset contains 2043 subjects after removing subjects with missing values.

The 9 QIDS-SR-16 subscores include sleep, mood, appetite, concentration, self-esteem, thoughts of death, interest, energy and psychomotor changes. We asked a psychiatrist to identify direct ground truth causal relations among the subscores before we ran the experiments. The ground truth includes: (1) sleep causes mood (Motomura et al. 2017; 2) energy causes psychomotor changes; (3) appetite causes energy; (4, 5) mood causes appetite and self-esteem (Hepworth et al. 2010; 6) psychomotor changes cause concentration; (7, 8) mood and self-esteem cause thoughts of death (Bhar et al. 2008).

We summarize the sensitivity, fallout and overall performance over 50 bootstrapped datasets in Fig. 6d, e, f. CIM achieved higher sensitivity than all other algorithms (Fig. 6d, t=5.66, p=7.86E-7). CIM also had a smaller fallout score compared to PC (Fig. 6(e), t=-19.19, p<2.20E-16). CIM therefore obtained the highest overall score compared to the other algorithms (Fig. 6(f), t=-14.95, p<2.20E-16). CIM finally completed within a short time frame like PC and RFCI (Fig. 7b). These results corroborate the superiority of CIM in a second real dataset.

6.3 Synthetic data

We next sampled from a mixture of DAGs to see if we could replicate the real data results. We specifically instantiated a linear DAG with an expected neighborhood size of 2, $p = 24$ vertices and linear coefficients drawn from Uniform($[-1, -0.25] \cup [0.25, 1]$). We then uniformly instantiated $q = 5$ to 15 binary variables for $\boldsymbol{T}$ and block randomized the edges in the DAG to each element of $\boldsymbol{T}$. We assigned the first 8 variables to time step 1, the second 8 to time step 2, and the third 8 to time step 3. We added a directed edge from the n^{th} variable in time step 1 to the n^{th} variable in time step 2, and similarly added the directed edges from time step 2 to time step 3 in order to model self-loops. We randomly selected a set of 0-2 latent common causes without replacement from $\boldsymbol{X}$, which we placed in $\boldsymbol{L}$ in addition to the variables in $\boldsymbol{T}$. We then selected a set of 0-2 selection variables $\boldsymbol{S}$ without replacement from the set $\boldsymbol{X} \setminus \boldsymbol{L}$.

We uniformly instantiating the mixing probabilities $p(T_i = 0)$ and $p(T_i = 1)$ for each $T_i \in \boldsymbol{T}$. We then generated 2000 samples as follows. For each sample, we drew an instantiation $\boldsymbol{T} = \boldsymbol{t}$ according to $\prod_{i=1}^{s} p(T_i)$ and created a graph containing the union of the edges associated with those elements in $\boldsymbol{t}$ equal to one. We then sampled the resultant DAG using a multivariate Gaussian distribution. We finally removed the latent variables and introduced selection bias by removing the bottom k^{th} percentile for each selection variable, with k chosen uniformly between 10 and 50.

We report the results in Fig. 6(g, h, i after repeating the above process 50 times. We computed the sensitivity and fallout using the ground truth in time steps 2 and 3. CIM achieved the highest sensitivity (Fig. 6g, t=3.71, p=5.35E-4). PC obtained the second highest sensitivity, but CIM had a lower fallout than PC (Fig. 6(h), t=-4.63,p=2.72E-5). CIM ultimately achieved the best overall score (Fig. 6(i), t=-3.78,p=4.37E-4). We finally provide timing results in Fig. 7c, showing that CIM takes about the same amount of time as PC and RFCI. We conclude that the synthetic data results mimic those seen with the real data.

7 Conclusion

We proposed to model causal processes using a mixture of DAGs to accommodate non-equilibriated distributions, sub-populations and cycles. We then introduced a constraint-based algorithm called CIM to infer causal relations from data even with latent variables and selection bias. The CIM algorithm infers ancestral relations with greater accuracy as compared to several constraint-based algorithms across simulated data and two real datasets with partially known ground truths. CIM thus broadens the scope of causal discovery to processes that do not necessarily follow a single graphical structure. Future work may consider further improving the accuracy of CIM by exhaustive search or continuous optimization, similar to the works: (Hyttinen et al. 2013), (Hyttinen et al. 2014), (Lu et al. 2021) and (Zheng et al. 2018).

Appendix

Equilibrium distribution

Our interpretation of cycles differs from the interpretation used with equilibrium distributions. An equilibrium distribution $\mathbb{P}$ refers to a distribution that obeys a structural equation model with independent errors respecting a potentially cyclic graph $\mathbb{G}$. In other words, we can describe the variables $\boldsymbol{X}$ as $X_i = g_i(\text{Pa}_{\mathbb{G}}(X_i), \varepsilon_i)$ for all $X_i \in \boldsymbol{X}$ such that X_i is measurable according to the sigma algebra $\sigma(\text{Pa}_{\mathbb{G}}(X_i), \varepsilon_i)$; we have $\varepsilon_i \in \boldsymbol{\varepsilon}$, where the set $\boldsymbol{\varepsilon}$ contains jointly independent errors (Evans 2016).

We can simulate data from the equilibrium distribution in practice using the fixed point method (Fisher 1970). The fixed point method involves two steps. We first sample the error terms according to their independent distributions and then initialize $\boldsymbol{X}$ to some values. We then apply the structural equations iteratively until the values of $\boldsymbol{X}$ converge almost surely to a fixed point. The values of $\boldsymbol{X}$ are not guaranteed to converge to a fixed point all of the time for every set of structural equations, but we only consider those structural equations and error distributions which do. Notice therefore that if $X_i \to X_j$ and $X_j \to X_i$ in $\mathbb{G}$, then applying the second step of the fixed point method means X_i is used to instantiate X_j in one iteration, and then that value of X_j is used to instantiate X_i in the next iteration. In other words, X_i causes X_j and then X_j causes X_i; the process of arriving at a fixed point therefore coincides with our mixture of DAGs interpretation of cycles. We however do not consider the causal interpretation *at the fixed point* where X_i and X_j cause each other simultaneously.

Proofs

Let $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ denote disjoint subsets of $\boldsymbol{Z}$. Let $\bar{\mathbb{M}}$ denote the moral graph of $\text{Anc}_{\mathbb{M}}(\boldsymbol{A}' \cup \boldsymbol{B}' \cup \boldsymbol{C}')$. We prove the global Markov property by first finding two sets $\ddot{\boldsymbol{A}} \supseteq \boldsymbol{A}'$ and $\ddot{\boldsymbol{B}} \supseteq \boldsymbol{B}'$ separated by $\boldsymbol{C}'$ in $\bar{\mathbb{M}}$. The vertices $\ddot{\boldsymbol{A}}$ and $\ddot{\boldsymbol{B}}$ represent the random variables $\widetilde{\boldsymbol{A}}^j \supseteq \boldsymbol{A}$ and $\widetilde{\boldsymbol{B}}^j \supseteq \boldsymbol{B}$, respectively, in $\mathbb{G}^j$. We use the cliques in $\bar{\mathbb{M}}$ to decompose the density $\prod_{Z_i \in \widetilde{\boldsymbol{A}}^j \cup \widetilde{\boldsymbol{B}}^j \cup \boldsymbol{C}} p(Z_i | \text{Pa}_{\mathbb{G}^j}(Z_i))$ into a non-negative function involving $\widetilde{\boldsymbol{A}}^j \cup \boldsymbol{C}$ and another non-negative function involving $\widetilde{\boldsymbol{B}}^j \cup \boldsymbol{C}$. Integrating out all of the variables *not* in $\boldsymbol{A} \cup \boldsymbol{B} \cup \boldsymbol{C}$ and then combining the densities across the DAGs in $\mathcal{G}$ finally allows us to represent

$p(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ as a product of a non-negative function involving $\boldsymbol{A} \cup \boldsymbol{C}$ and another non-negative function involving $\boldsymbol{B} \cup \boldsymbol{C}$ – thus arriving at conditional independence.

Theorem 1 *Let $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ denote disjoint subsets of $\boldsymbol{Z}$. If $\boldsymbol{A}' \perp\!\!\!\perp_d \boldsymbol{B}' | \boldsymbol{C}'$ in $\mathbb{M}$, then $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} | \boldsymbol{C}$.*

Proof We consider a partition of the vertices $\ddot{\boldsymbol{A}} \cup \ddot{\boldsymbol{B}} \cup \boldsymbol{C}' = \text{Anc}_{\mathbb{M}}(\boldsymbol{A}' \cup \boldsymbol{B}' \cup \boldsymbol{C}')$ so that $\boldsymbol{A}' \subseteq \ddot{\boldsymbol{A}}$, $\boldsymbol{B}' \subseteq \ddot{\boldsymbol{B}}$, and $\ddot{\boldsymbol{A}}$, $\ddot{\boldsymbol{B}}$ and $\boldsymbol{C}'$ are disjoint sets of vertices. We require that $\ddot{\boldsymbol{A}}$ and $\ddot{\boldsymbol{B}}$ be separated by $\boldsymbol{C}'$ in $\bar{\mathbb{M}}$; in other words, there does not exist an undirected path between $\ddot{\boldsymbol{A}}$ and $\ddot{\boldsymbol{B}}$ that is active (i.e., unblocked) given $\boldsymbol{C}'$.

We now construct such a partition $(\ddot{\boldsymbol{A}}, \ddot{\boldsymbol{B}})$. First set $\ddot{\boldsymbol{A}}$ to $\boldsymbol{A}'$ and $\ddot{\boldsymbol{B}}$ to $\boldsymbol{B}'$. If $\boldsymbol{A}' \perp\!\!\!\perp_d \boldsymbol{B}' | \boldsymbol{C}'$ in $\mathbb{M}$, then $\boldsymbol{A}'$ and $\boldsymbol{B}'$ are separated by $\boldsymbol{C}'$ in the moral graph of the smallest ancestral set $\text{Anc}_{\mathbb{M}}(\boldsymbol{A}' \cup \boldsymbol{B}' \cup \boldsymbol{C}')$ (Proposition 3 in (Lauritzen et al. 1990)). $\ddot{\boldsymbol{A}}$ and $\ddot{\boldsymbol{B}}$ are therefore separated by $\boldsymbol{C}'$ in $\bar{\mathbb{M}}$ at the moment. Now consider the set of vertices $\boldsymbol{H} = \text{Anc}_{\mathbb{M}}(\boldsymbol{A}' \cup \boldsymbol{B}' \cup \boldsymbol{C}') \setminus (\boldsymbol{A}' \cup \boldsymbol{B}' \cup \boldsymbol{C}')$. We will put members of $\boldsymbol{H}$ into either $\ddot{\boldsymbol{A}}$ or $\ddot{\boldsymbol{B}}$. We have two situations for each vertex $H_i^m \in \boldsymbol{H}$:

1. In $\bar{\mathbb{M}}$, there does not exist an undirected path between H_i^m and $\boldsymbol{A}'$ or an undirected path between H_i^m and $\boldsymbol{B}'$ (or both) that is active given $\boldsymbol{C}'$. More specifically:
   (a) If there does not exist an undirected path between H_i^m and $\boldsymbol{A}'$ that is active given $\boldsymbol{C}'$, but such a path exists between H_i^m and $\boldsymbol{B}'$, then include H_i^m into $\ddot{\boldsymbol{B}}$ so that $\ddot{\boldsymbol{B}} \leftarrow \ddot{\boldsymbol{B}} \cup H_i^m$.
   (b) If there does not exist an undirected path between H_i^m and $\boldsymbol{B}'$ that is active given $\boldsymbol{C}'$, but such a path exists between H_i^m and $\boldsymbol{A}'$, then include H_i^m into $\ddot{\boldsymbol{A}}$ so that $\ddot{\boldsymbol{A}} \leftarrow \ddot{\boldsymbol{A}} \cup H_i^m$.
   (c) If there does not exist an undirected path between H_i^m and $\boldsymbol{A}'$ that is active given $\boldsymbol{C}'$ and there likewise does not exist such a path between H_i^m and $\boldsymbol{B}'$, then include H_i^m into $\ddot{\boldsymbol{A}}$ so that $\ddot{\boldsymbol{A}} \leftarrow \ddot{\boldsymbol{A}} \cup H_i^m$.

2. In $\bar{\mathbb{M}}$, there exists an undirected path between H_i^m and $\boldsymbol{A}'$ and an undirected path between H_i^m and $\boldsymbol{B}'$ that are both active given $\boldsymbol{C}'$. But this implies that $\boldsymbol{A}'$ and $\boldsymbol{B}'$ are connected given $\boldsymbol{C}'$ in $\bar{\mathbb{M}}$ via H_i^m – a contradiction.

We have constructed a disjoint partition of vertices $(\ddot{\boldsymbol{A}}, \ddot{\boldsymbol{B}})$ such that $\ddot{\boldsymbol{A}} \cup \ddot{\boldsymbol{B}} \cup \boldsymbol{C}' = \text{Anc}_{\mathbb{M}}(\boldsymbol{A}' \cup \boldsymbol{B}' \cup \boldsymbol{C}')$. Moreover, $\ddot{\boldsymbol{A}}$ and $\ddot{\boldsymbol{B}}$ are separated given $\boldsymbol{C}'$ in $\bar{\mathbb{M}}$.

We may then consider all of the cliques in $\bar{\mathbb{M}}$ corresponding to each vertex and its married parents. Denote this set of cliques as $\mathcal{E}$. Also let $\mathcal{E}_{\ddot{\boldsymbol{B}}}$ denote the set of cliques in $\mathcal{E}$ that have non-empty intersection with $\ddot{\boldsymbol{B}}$. Because $\ddot{\boldsymbol{A}}$ and $\ddot{\boldsymbol{B}}$ are separated given $\boldsymbol{C}'$, the vertices $\ddot{\boldsymbol{A}}$ and $\ddot{\boldsymbol{B}}$ are also non-adjacent in $\bar{\mathbb{M}}$; this implies that no clique in $\mathcal{E}_{\ddot{\boldsymbol{B}}}$ can contain a member of $\ddot{\boldsymbol{A}}$. We also have $\ddot{\boldsymbol{B}} \cap e = \emptyset$ for all $e \in \mathcal{E} \setminus \mathcal{E}_{\ddot{\boldsymbol{B}}}$.

Consider an arbitrary graph $\mathbb{G}^j \in \mathcal{G}$. Let $\mathcal{E}^j$ denote the cliques in $\mathcal{E}$ only containing vertices in $\mathbb{G}^j$ – likewise for $\mathcal{E}^j_{\ddot{\boldsymbol{B}}}$. Note that we can associate the *vertices* $\ddot{\boldsymbol{A}}$ with the *random variables* $\widetilde{\boldsymbol{A}}^j = \cup_{H_i^j \in \ddot{\boldsymbol{A}}} H_i$ – and similarly for $\widetilde{\boldsymbol{B}}^j$. We can then write the density factorizing according to $\mathbb{G}^j$ as follows:

$$
\begin{aligned}
&\prod_{Z_i \in \widetilde{\boldsymbol{A}}^j \cup \widetilde{\boldsymbol{B}}^j \cup \boldsymbol{C}} p(Z_i|\mathrm{Pa}_{\mathbb{G}^j}(Z_i)) \\
&= \prod_{\left\{Z_i \cup \mathrm{Pa}_{\mathbb{G}^j}(Z_i)\right\} \in \mathcal{E}^j \setminus \mathcal{E}^j_{\widetilde{\boldsymbol{B}}}} p(Z_i|\mathrm{Pa}_{\mathbb{G}^j}(Z_i)) \prod_{\left\{Z_i \cup \mathrm{Pa}_{\mathbb{G}^j}(Z_i)\right\} \in \mathcal{E}^j_{\widetilde{\boldsymbol{B}}}} p(Z_i|\mathrm{Pa}_{\mathbb{G}^j}(Z_i)) \\
&= f(\widetilde{\boldsymbol{A}}^j, \boldsymbol{C}) f(\widetilde{\boldsymbol{B}}^j, \boldsymbol{C}),
\end{aligned}
$$

where f denotes some non-negative function.

We now proceed by integrating out $[\widetilde{\boldsymbol{A}}^j \cup \widetilde{\boldsymbol{B}}^j] \setminus [\boldsymbol{A} \cup \boldsymbol{B}]$:

$$
\begin{aligned}
p(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) &= \sum_{j=1}^{q} \mathbb{1}_{\boldsymbol{T} \in \mathcal{T}^j} \Big(\sum_{[\widetilde{\boldsymbol{A}}^j \cup \widetilde{\boldsymbol{B}}^j] \setminus [\boldsymbol{A} \cup \boldsymbol{B}]} \prod_{Z_i \in \widetilde{\boldsymbol{A}}^j \cup \widetilde{\boldsymbol{B}}^j \cup \boldsymbol{C}} p(Z_i|\mathrm{Pa}_{\mathbb{G}^j}(Z_i)) \Big) \\
&= \sum_{j=1}^{q} \mathbb{1}_{\boldsymbol{T} \in \mathcal{T}^j} \Big(\sum_{[\widetilde{\boldsymbol{A}}^j \cup \widetilde{\boldsymbol{B}}^j] \setminus [\boldsymbol{A} \cup \boldsymbol{B}]} f(\widetilde{\boldsymbol{A}}^j, \boldsymbol{C}) f(\widetilde{\boldsymbol{B}}^j, \boldsymbol{C}) \Big) \\
&= \sum_{j=1}^{q} \mathbb{1}_{\boldsymbol{T} \in \mathcal{T}^j} \Big(\sum_{[\widetilde{\boldsymbol{A}}^j \setminus \boldsymbol{A}] \cup [\widetilde{\boldsymbol{B}}^j \setminus \boldsymbol{B}]} f(\widetilde{\boldsymbol{A}}^j, \boldsymbol{C}) f(\widetilde{\boldsymbol{B}}^j, \boldsymbol{C}) \Big) \\
&= \sum_{j=1}^{q} \mathbb{1}_{\boldsymbol{T} \in \mathcal{T}^j} \Big(\Big[\sum_{[\widetilde{\boldsymbol{B}}^j \setminus \boldsymbol{B}]} \Big[\sum_{[\widetilde{\boldsymbol{A}}^j \setminus \boldsymbol{A}]} f(\widetilde{\boldsymbol{A}}^j, \boldsymbol{C}) \Big] f(\widetilde{\boldsymbol{B}}^j, \boldsymbol{C}) \Big] \Big) \\
&= \sum_{j=1}^{q} \mathbb{1}_{\boldsymbol{T} \in \mathcal{T}^j} \Big(\sum_{[\widetilde{\boldsymbol{A}}^j \setminus \boldsymbol{A}]} f(\widetilde{\boldsymbol{A}}^j, \boldsymbol{C}) \sum_{[\widetilde{\boldsymbol{B}}^j \setminus \boldsymbol{B}]} f(\widetilde{\boldsymbol{B}}^j, \boldsymbol{C}) \Big) \\
&= \sum_{j=1}^{q} \mathbb{1}_{\boldsymbol{T} \in \mathcal{T}^j} \Big(f(\boldsymbol{A}, \boldsymbol{C}) f(\boldsymbol{B}, \boldsymbol{C}) \Big).
\end{aligned}
$$

The fourth equality follows because $[\widetilde{\boldsymbol{A}}^j \setminus \boldsymbol{A}] \cap [\widetilde{\boldsymbol{B}}^j \setminus \boldsymbol{B}] = \emptyset$ by construction.

Suppose $\boldsymbol{T} \cap (\boldsymbol{A} \cup \boldsymbol{B} \cup \boldsymbol{C}) = \emptyset$. Then $\sum_{j=1}^{q} \mathbb{1}_{\boldsymbol{T} \in \mathcal{T}^j} \Big(f(\boldsymbol{A}, \boldsymbol{C}) f(\boldsymbol{B}, \boldsymbol{C}) \Big) = f(\boldsymbol{A}, \boldsymbol{C}) f(\boldsymbol{B}, \boldsymbol{C})$, so $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} | \boldsymbol{C}$ in this case.

Now assume $\boldsymbol{T} \cap (\boldsymbol{A} \cup \boldsymbol{B} \cup \boldsymbol{C}) \neq \emptyset$. Let $\boldsymbol{U} = \boldsymbol{T} \cap (\boldsymbol{A} \cup \boldsymbol{C})$ and $\boldsymbol{V} = \boldsymbol{T} \cap (\boldsymbol{B} \cup \boldsymbol{C})$. Also let $\mathcal{U}$ denote the set of all values of $\boldsymbol{U}$. The values in $\mathcal{U}$ index the r unique functions $f_{\boldsymbol{U}}((\boldsymbol{A} \cup \boldsymbol{C}) \setminus \boldsymbol{U}) = f(\boldsymbol{A}, \boldsymbol{C})$. Let $\mathcal{U}^k$ more specifically denote those values of $\boldsymbol{U}$ associated with the k^{th} unique function over $(\boldsymbol{A} \cup \boldsymbol{C}) \setminus \boldsymbol{U}$, denoted by $f^k_{\boldsymbol{U}}((\boldsymbol{A} \cup \boldsymbol{C}) \setminus \boldsymbol{U})$. Similarly, let $\mathcal{V}$ denote the set of all values of $\boldsymbol{V}$ indexing s unique functions $f_{\boldsymbol{V}}((\boldsymbol{B} \cup \boldsymbol{C}) \setminus \boldsymbol{V}) = f(\boldsymbol{B}, \boldsymbol{C})$. Also let $\mathcal{V}^l$ refer to the values of $\boldsymbol{V}$ associated with the l^{th} unique function over $(\boldsymbol{B} \cup \boldsymbol{C}) \setminus \boldsymbol{V}$, denoted by $f^l_{\boldsymbol{V}}((\boldsymbol{B} \cup \boldsymbol{C}) \setminus \boldsymbol{V})$. We must have:

$$
\begin{aligned}
&\sum_{j=1}^{q} \mathbb{1}_{\boldsymbol{T} \in \mathcal{T}^j} f_{\boldsymbol{U}}((\boldsymbol{A} \cup \boldsymbol{C}) \setminus \boldsymbol{U}) f_{\boldsymbol{V}}((\boldsymbol{B} \cup \boldsymbol{C}) \setminus \boldsymbol{V}) \\
&\quad = \sum_{k=1}^{r} \mathbb{1}_{\boldsymbol{U} \in \mathcal{U}^k} f^k_{\boldsymbol{U}}((\boldsymbol{A} \cup \boldsymbol{C}) \setminus \boldsymbol{U}) \sum_{l=1}^{s} \mathbb{1}_{\boldsymbol{V} \in \mathcal{V}^l} f^l_{\boldsymbol{V}}((\boldsymbol{B} \cup \boldsymbol{C}) \setminus \boldsymbol{V}),
\end{aligned} \tag{8}
$$

because $f_{\boldsymbol{U}=\boldsymbol{u}}((\boldsymbol{A} \cup \boldsymbol{C}) \setminus \boldsymbol{U}) f_{\boldsymbol{V}=\boldsymbol{v}}((\boldsymbol{B} \cup \boldsymbol{C}) \setminus \boldsymbol{V})$ is the product of a unique function over $(\boldsymbol{A} \cup \boldsymbol{C}) \setminus \boldsymbol{U}$ *and* a unique function over $(\boldsymbol{B} \cup \boldsymbol{C}) \setminus \boldsymbol{V}$. We can therefore write:

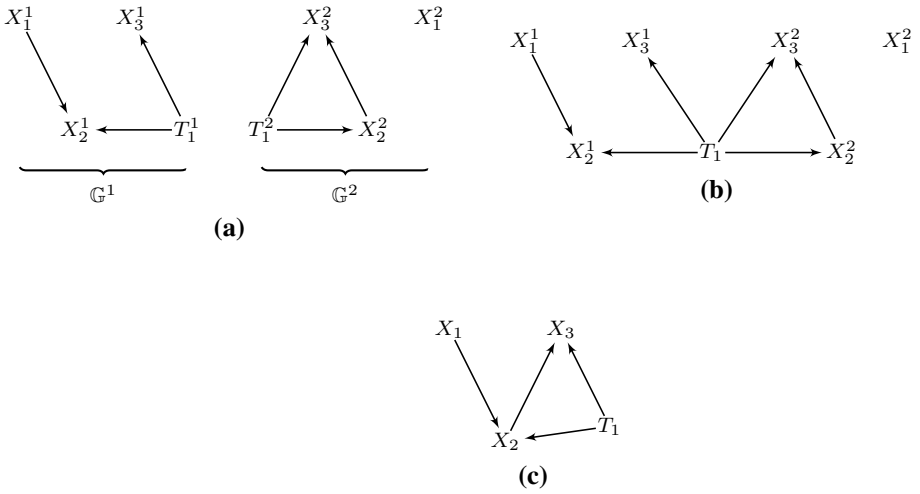


Fig. 8 An example where $\mathbb{M}$ implies more CI relations than $\mathbb{F}$. The two DAGs in $\mathcal{G}$ are plotted in (**a**). In (**b**), $X_1' \perp\!\!\!\perp_d X_3'$ in $\mathbb{M}$ implies $X_1 \perp\!\!\!\perp X_3$. The fused graph $\mathbb{F}$ in (**c**) however does not imply the independence relation

$$\sum_{j=1}^{q} \mathbb{1}_{\boldsymbol{T}\in\mathcal{T}^j}\Big(f(\boldsymbol{A},\boldsymbol{C})f(\boldsymbol{B},\boldsymbol{C})\Big)$$
$$= \Big(\sum_{k=1}^{r} \mathbb{1}_{\boldsymbol{U}\in\mathcal{U}^k} f_{\boldsymbol{U}}^k((\boldsymbol{A}\cup\boldsymbol{C})\setminus\boldsymbol{U})\Big)\Big(\sum_{l=1}^{s} \mathbb{1}_{\boldsymbol{V}\in\mathcal{V}^l} f_{\boldsymbol{V}}^l((\boldsymbol{B}\cup\boldsymbol{C})\setminus\boldsymbol{V})\Big).$$

The conclusion follows by this last equality. □

Lemma 1 *Suppose* ${}^aO_i' \perp\!\!\!\perp_d {}^bO_j' | \boldsymbol{W}' \cup \boldsymbol{S}'$ *in* $\mathbb{M}$ *but* ${}^aO_i' \not\perp\!\!\!\perp_d {}^bO_j' | \boldsymbol{V}' \cup \boldsymbol{S}'$ *for every* $\boldsymbol{V} \subset \boldsymbol{W}$. *If* ${}^cO_k \in \boldsymbol{W}$, *then* ${}^cO_k \in \mathrm{Anc}_{\mathbb{F}}({}^aO_i \cup {}^bO_j \cup \boldsymbol{S})$.

Proof We invoke Lemma 15 in (Strobl 2018) by setting $\boldsymbol{R} = \emptyset$, $O_i = {}^aO_i'$, $O_j = {}^bO_j'$, $\boldsymbol{W} = \boldsymbol{W}'$ and $\boldsymbol{S} = \boldsymbol{S}'$ in that paper. We can then conclude that ${}^cO_k' \in \mathrm{Anc}_{\mathbb{M}}({}^aO_i' \cup {}^bO_j' \cup \boldsymbol{S}')$. Moreover, if ${}^cO_k' \in \mathrm{Anc}_{\mathbb{M}}({}^aO_i' \cup {}^bO_j' \cup \boldsymbol{S}')$, then ${}^cO_k \in \mathrm{Anc}_{\mathbb{F}}({}^aO_i \cup {}^bO_j \cup \boldsymbol{S})$ by construction of $\mathbb{F}$. □

Theorem 2 *Suppose the longitudinal density* $p(\cup_{k=1}^{w} {}^k\boldsymbol{O}, \boldsymbol{L}, \boldsymbol{S})$ *factorizes according to Eq.* (5). *Assume that all prior information* $\mathcal{P}$ *is correct. Then, under d-separation faithfulness w.r.t.* $\mathbb{M}$, *the CIM algorithm returns* $\widehat{\mathbb{F}}^*$ *– the mixed graph* $\mathbb{F}^*$ *partially oriented.*

Proof Under d-separation faithfulness w.r.t. $\mathbb{M}$, CI and d-separation w.r.t. $\mathbb{M}$ are equivalent by Theorem 1, so we can refer to them interchangeably. Algorithm 2 finds the adjacencies in List 1 because we must always have ${}^a_b\mathrm{Adj}_{\mathbb{F}^*}({}^aO_i) \subseteq {}^a_b\mathrm{Adj}_{\widehat{\mathbb{F}}^*}({}^aO_i)$ in Step 8 of Algorithm 2. Step 4 discovers the correct tails by Lemma 1. Step 5 follows directly by transitivity of the tails. □

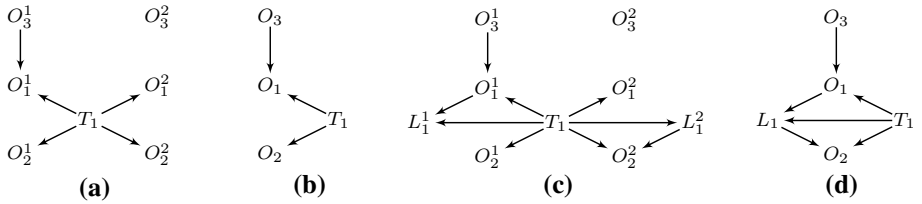


Fig. 9 Example showing that we cannot infer v-structures without additional assumptions. **a** and **b** show the pair $(\mathbb{M}_1, \mathbb{F}_1)$, respectively, where $O_1 \notin \text{Anc}_{\mathbb{F}_1}(O_2)$. **c** and **d** on the other hand show the pair $(\mathbb{M}_2, \mathbb{F}_2)$, where $O_1 \in \text{Anc}_{\mathbb{F}_2}(O_2 \cup \boldsymbol{S})$

Comparison to previous global Markov property

Spirtes (1994) also characterized the global Markov property across a mixture of DAGs using $\mathbb{F}$ under the additional assumption that $\boldsymbol{T}$ is discrete, latent and univariate. The fused graph however implies less CI relations than $\mathbb{M}$ as illustrated in Fig. 8. We have drawn $\mathbb{F}$ in Fig. 8c. X_1 and X_3 are d-connected in $\mathbb{F}$ even though X_1' and X_3' are d-separated in $\mathbb{M}$ in Fig. 8b. We have established an instance where the mixture graph implies strictly more independence relations than the fused graph.

The mixture graph in fact always implies at least the same number of CI relations as the fused graph:

Proposition 1 *Let $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ denote disjoint subsets of $\boldsymbol{Z}$. If $\boldsymbol{A} \perp\!\!\!\perp_d \boldsymbol{B}|\boldsymbol{C}$ in $\mathbb{F}$, then $\boldsymbol{A}' \perp\!\!\!\perp_d \boldsymbol{B}'|\boldsymbol{C}'$ in $\mathbb{M}$.*

Proof We create q copies of $\mathbb{F}$ and plot them adjacent to each other. Denote the resultant graph as $\mathbb{F}'$. As a result, we have $\boldsymbol{A} \perp\!\!\!\perp_d \boldsymbol{B}|\boldsymbol{C}$ in $\mathbb{F}$ if and only if $\boldsymbol{A}' \perp\!\!\!\perp_d \boldsymbol{B}'|\boldsymbol{C}'$ in $\mathbb{F}'$. Create a new graph $\mathbb{F}''$ as follows. First set $\mathbb{F}''$ equal to $\mathbb{F}'$. Then merge $T_i' \subseteq \boldsymbol{T}'$ into a single vertex T_i for each $T_i \in \boldsymbol{T}$. Denote the resultant graph as $\mathbb{F}''$.

We will show that $\boldsymbol{A}' \perp\!\!\!\perp_d \boldsymbol{B}'|\boldsymbol{C}'$ in $\mathbb{F}''$ implies $\boldsymbol{A}' \perp\!\!\!\perp_d \boldsymbol{B}'|\boldsymbol{C}'$ in $\mathbb{F}'$. Denote the moral graph of $\text{Anc}_{\mathbb{F}''}(\boldsymbol{A}' \cup \boldsymbol{B}' \cup \boldsymbol{C}')$ by $\bar{\mathbb{F}}''$. If $\boldsymbol{A}' \perp\!\!\!\perp_d \boldsymbol{B}'|\boldsymbol{C}'$ in $\mathbb{F}''$, then there exists an active path $\Pi_{\boldsymbol{A}'\boldsymbol{B}'}$ between $\boldsymbol{A}'$ and $\boldsymbol{B}'$ given $\boldsymbol{C}'$ in $\bar{\mathbb{F}}''$ by Proposition 3 in (Lauritzen et al. 1990). We can replace an arbitrary vertex Z_i^m on $\Pi_{\boldsymbol{A}'\boldsymbol{B}'}$ with Z_i^n on $\mathbb{G}^n \in \mathcal{G}$. Repeating this process for every vertex on $\Pi_{\boldsymbol{A}'\boldsymbol{B}'}$ creates a non-simple path (i.e. with potentially repeated vertices) between $\boldsymbol{A}^n$ and $\boldsymbol{B}^n$ that does not contain any member of $\boldsymbol{C}^n$. There thus exists a simple path without repeated vertices between $\boldsymbol{A}^n$ and $\boldsymbol{B}^n$ that does not contain any member of $\boldsymbol{C}^n$ in $\bar{\mathbb{F}}''$ – so in $\bar{\mathbb{F}}'$ as well. Hence $\boldsymbol{A}' \perp\!\!\!\perp_d \boldsymbol{B}'|\boldsymbol{C}'$ in $\mathbb{F}'$ again by Proposition 3 in (Lauritzen et al. 1990).

Note that all of the edges in $\mathbb{M}$ are contained within $\mathbb{F}''$. We can therefore write $\boldsymbol{A}' \perp\!\!\!\perp_d \boldsymbol{B}'|\boldsymbol{C}'$ in $\mathbb{M}$ implies $\boldsymbol{A}' \perp\!\!\!\perp_d \boldsymbol{B}'|\boldsymbol{C}'$ in $\mathbb{F}''$, which implies $\boldsymbol{A}' \perp\!\!\!\perp_d \boldsymbol{B}'|\boldsymbol{C}'$ in $\mathbb{F}'$, which implies $\boldsymbol{A} \perp\!\!\!\perp_d \boldsymbol{B}|\boldsymbol{C}$ in $\mathbb{F}$. The conclusion follows by contrapositive.

□

$\mathbb{M}$ is thus superior to $\mathbb{F}$ because (1) $\mathbb{M}$ implies at least as many CI relations as $\mathbb{F}$, and (2) $\mathbb{M}$ implies strictly more CI relations in some cases.

Counterexample

We sometimes cannot even detect v-structures with a CI oracle alone. Consider for example the mixture and fused graph pairs shown in Fig. 9 where $\{T_1, L_1\} = \boldsymbol{L}$ and $\boldsymbol{S} = \emptyset$. We have $O_1 \notin \text{Anc}_{\mathbb{F}_1}(O_2)$ and $O_1 \in \text{Anc}_{\mathbb{F}_2}(O_2 \cup \boldsymbol{S})$. However, all of the following relations hold in both $\mathbb{M}_1$ and $\mathbb{M}_2$: $O'_3 \perp\!\!\!\perp_d O'_2$, $O'_3 \not\perp\!\!\!\perp_d O'_2|O'_1$, $O'_3 \not\perp\!\!\!\perp_d O'_1$, $O'_3 \not\perp\!\!\!\perp_d O'_1|O'_2$, $O'_2 \not\perp\!\!\!\perp_d O'_1$, and $O'_2 \not\perp\!\!\!\perp_d O'_1|O'_3$. In other words, $O'_i \perp\!\!\!\perp_d O'_j|\boldsymbol{W}' \cup \boldsymbol{S}'$ in $\mathbb{M}_1$ if and only if $O'_i \perp\!\!\!\perp_d O'_j|\boldsymbol{W}' \cup \boldsymbol{S}'$ in $\mathbb{M}_2$ for any $O_i, O_j \in \boldsymbol{O}$ and $\boldsymbol{W} \subseteq \boldsymbol{O} \setminus \{O_i, O_j\}$. We therefore cannot distinguish the two mixed graphs $O_3 \ast\!\!\rightarrow O_1 \leftarrow\!\!\ast\, O_2$ and $O_3 \ast\!\!\rightarrow O_1 -\!\!\ast\, O_2$ for $\mathbb{F}_1^*$ and $\mathbb{F}_2^*$, respectively, using CI relations alone under d-separation faithfulness w.r.t. $\mathbb{M}$; this holds even though $\mathbb{F}_1$ and $\mathbb{F}_2$ are acyclic.

Author contributions Conceptualization, methodology, formal analysis and investigation, writing completed by EVS

Funding None.

Availability of data and material Synthetic data available at https://github.com/ericstrobl/CIM.

Code availability Available at https://github.com/ericstrobl/CIM.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

Alberts, B., Wilson, J. H., & Hunt, T. (2015). *Molecular biology of the cell* (6th ed.). Garland Science: Taylor and Francis Group.

Aronow, W. S., Dendinger, J., & Rokaw, S. N. (1971). Heart rate and carbon monoxide level after smoking high-, low-, and non-nicotine cigarettes: A study in male patients with angina pectoris. *Annals of Internal Medicine*, *74*(5), 697–702.

Bellot, A., Branson, K., & van der Schaar, M. (2021). Consistency of mechanistic causal discovery in continuous-time using neural odes. *arXiv preprint* arXiv:2105.02522.

Bhar, S., Ghahramanlou-Holloway, M., Brown, G., & Beck, A. T. (2008). Self-esteem and suicide ideation in psychiatric outpatients. *Suicide and life-threatening behavior*, *38*(5), 511–516.

Chiolero, A., Faeh, D., Paccaud, F., & Cornuz, J. (2008). Consequences of smoking for body weight, body fat distribution, and insulin resistance. *The American journal of clinical nutrition, 87*(4), 801–809.

Colombo, D., & Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, *15*(1), 3741–3782. ISSN 1532-4435. URL: http://dl.acm.org/citation.cfm?id=2627435.2750365.

Dagum, P., Galper, A., Horvitz, E., & Seiver, A. (1995). Uncertain reasoning and forecasting. *International Journal of Forecasting, 11*, 73–87.

Entner, D., & Hoyer, P. O. (2010). On causal discovery from time series data using fci. *Probabilistic Graphical Models*, pp. 121–128.

Evans, R. J. (2016). Graphs for margins of Bayesian networks. *Scandinavian Journal of Statistics, 43*(3), 625–648.

Fisher, F. M. (1970). A correspondence principle for simultaneous equation models. *Econometrica*, *38*(1), 73–92. URL: https://EconPapers.repec.org/RePEc:ecm:emetrp:v:38:y:1970:i:1:p:73-92.

Forré, P., & Mooij, J. M. (2017). Markov properties for graphical models with cycles and latent variables. arXiv:1710.08775 [math.ST]

Forré, Patrick, & Mooij, Joris M. (2018). Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders. In: *Proceedings of the 34th Annual Conference on Uncertainty in Artificial Intelligence (UAI-18)*.

Haass, M., & Kübler, W. (1997). Nicotine and sympathetic neurotransmission. *Cardiovascular Drugs and Therapy, 10*(6), 657–665.

Hepworth, R., Mogg, K., Brignell, C., & Bradley, B. P. (2010). Negative mood increases selective attention to food cues and subjective appetite. *Appetite*, *54*(1), 134–142.

Hyttinen, A., Hoyer, P. O., Eberhardt, F., & Järvisalo, M. (2013). Discovering cyclic causal models with latent variables: A general sat-based procedure. In: *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, UAI 2013, Bellevue, WA, USA, August 11-15*. URL: https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=2391&proceeding_id=29.

Hyttinen, A., Eberhardt, F., & Järvisalo, M. (2014). Constraint-based causal discovery: Conflict resolution with answer set programming. In: *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI'14, pages 340–349, Arlington, Virginia, United States. AUAI Press. ISBN 978-0-9749039-1-0. URL: http://dl.acm.org/citation.cfm?id=3020751.3020787.

Jaber, A., Kocaoglu, M., Shanmugam, K., & Bareinboim, E. (2020). Causal discovery from soft interventions with unknown targets: Characterization and learning. *Advances in neural information processing systems*, *33*.

Jo, Y.- H., Talmage, D. A., & Role, L. W. (2002). Nicotinic receptor-mediated effects on appetite and food intake. *Journal of neurobiology*, *53*(4), 618–632.

Ke, N. R., Bilaniuk, O., Goyal, A., Bauer, S., Larochelle, H., Schölkopf, B., Mozer, M. C., Pal, C. & Yoshua B. (2019). Learning neural causal models from unknown interventions. *arXiv preprint* arXiv:1910.01075.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N., & Leimer, H. G. (1990). Independence Properties of Directed Markov Fields. *Networks, 20*(5), 491–505. https://doi.org/10.1002/net.3230200503

Levy, M. N. (1971). Brief reviews: sympathetic-parasympathetic interactions in the heart. *Circulation research*, *29*(5): 437–445.

Ni Y. L., Kun Z., & Changhe Y. (2021). Improving causal discovery by optimal bayesian network learning. In: *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, UAI'98.

Lucke, C., Hehrmann, R., Von Mayersbach, K., & von Zur Mühlen, A. (1977). Studies on circadian variations of plasma tsh, thyroxine and triiodothyronine in man. *European Journal of Endocrinology*, *86*(1), 81–88.

Mahmood, S. S., Levy, D., Vasan, R. S., & Wang. T. J. (2014). The framingham heart study and the epidemiology of cardiovascular disease: a historical perspective. *The Lancet*, *383* (9921): 999 – 1008, 2014. ISSN 0140-6736. https://doi.org/10.1016/S0140-6736(13)61752-3.

Malinsky, D., & Spirtes, P. (2018). Causal structure learning from multivariate time series in settings with unmeasured confounding. In: *Proceedings of 2018 ACM SIGKDD Workshop on Causal Discovery, volume 92 of Proceedings of Machine Learning Research*, pages 23–47, London, UK, 20 Aug 2018. PMLR. URL: http://proceedings.mlr.press/v92/malinsky18a.html.

Mooij, J. M. & Claassen, T. (2020). Constraint-based causal discovery using partial ancestral graphs in the presence of cycles. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1159–1168. PMLR.

Mooij, J. M., Magliacane, S., & Claassen, T. (2020). Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, *21*(99), 1–108.

Motomura, Y., Katsunuma, R., Yoshimura, M., & Mishima, K. (2017). Two days' sleep debt causes mood decline during resting state via diminished amygdala-prefrontal connectivity. *Sleep*, *40*(10).

Murphy, K. P. (2002). *Dynamic bayesian networks: Representation, inference and learning*. Berkeley: University of California.

Parini, P., Angelin, B., & Rudling, M. (1999). Cholesterol and lipoprotein metabolism in aging: Reversal of hypercholesterolemia by growth hormone treatment in old rats. *Arteriosclerosis, Thrombosis, and Vascular Biology, 19*(4), 832–839.

Perkins, N. J., & Schisterman, E. F. (2006). The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *American journal of epidemiology*, *163*(7), 670–675.

Pinto, E. (2007). Blood pressure and ageing. *Postgraduate Medical Journal, 83*(976), 109–114.
Pirahanchi, Y., Toro, F., & Jialal, I. (2021) Physiology, thyroid stimulating hormone. *StatPearls*.
Richardson, T. (1996). A discovery algorithm for directed cyclic graphs. In: *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence*, pp. 454–461.
Rubenstein, P., Bongers, S., Schölkopf, B., & Mooij, J. M. (2018). From deterministic odes to dynamic structural causal models. In: *34th Conference on Uncertainty in Artificial Intelligence (UAI 2018)*, pp. 114–123. Curran Associates, Inc.
Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., & Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, *5*(11), eaau4996.
Saeed, B, Panigrahi, S., & Uhler, C. (2020). Causal structure discovery from distributions arising from mixtures of dags. *arXiv preprint* arXiv:2001.11940.
Safar, M. E. (2005). Systolic hypertension in the elderly: arterial wall mechanical properties and the renin–angiotensin–aldosterone system. *Journal of Hypertension*, *23*(4), 673–681.
Shah, R. D. Peters, J., et al. (2020). The hardness of conditional independence testing and the generalised covariance measure. *Annals of Statistics*, *48*(3), 1514–1538.
Sinyor, M., Schaffer, A., & Levitt, A. (2010). The sequenced treatment alternatives to relieve depression (star* d) trial: a review. *The Canadian Journal of Psychiatry, 55*(3), 126–135.
Spirtes, P. (1994). Conditional independence properties in directed cyclic graphical models for feedback. Technical report, Carnegie Mellon University.
Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, Prediction, and Search*. MIT press, 2nd edition.
Spirtes, P. (1995). Directed cyclic graphical representations of feedback models. In: *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 491–498.
Squires, C., Wang, Y., & Uhler, C. (2020). Permutation-based causal structure learning with unknown intervention targets. In: *Conference on Uncertainty in Artificial Intelligence*, pp. 1039–1048. PMLR.
Strobl, E. V. (2017). *Causal Discovery Under Non-Stationary Feedback*. PhD thesis, University of Pittsburgh.
Strobl, E. V. (2018). A constraint-based algorithm for causal discovery with cycles, latent variables and selection bias. *International Journal of Data Science and Analytics*. ISSN 2364-4168. https://doi.org/10.1007/s41060-018-0158-2.
Strobl, E. V. (2019). Improved causal discovery from longitudinal data using a mixture of dags. In: Le, T. D., Li, J., Zhang, K., Cui, E. K. P., & Hyvärinen, A., (Eds.), *Proceedings of Machine Learning Research, volume 104 of Proceedings of Machine Learning Research*, pp. 100–133, Anchorage, Alaska, USA, 05 Aug 2019. PMLR. URL http://proceedings.mlr.press/v104/strobl19a.html.
Strobl, E. V., Zhang, K., & Visweswaran, S. (2018). Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*. https://doi.org/10.1515/jci-2018-0017
Taris, T. W. (2000). *A Primer in Longitudinal Data Analysis*. Sage.
Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artifical Intelligence*, *172*(16–17), 1873–1896, November 2008. ISSN 0004-3702. https://doi.org/10.1016/j.artint.2008.08.001.
Zhang, K., & Glymour, M. (2018). Unmixing for causal inference: Thoughts on mccaffrey and danks. *The British Journal for the Philosophy of Science*, p. axy040.
Zhang, K., Huang, B., Zhang, J., Glymour, C., & Schölkopf, B. (2017). Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pages 1347–1353. AAAI Press. ISBN 978-0-9992411-0-3. URL http://dl.acm.org/citation.cfm?id=3171642.3171833.
Zheng, X. Aragam, B., Ravikumar, P. K. & Xing, E. P. (2018). Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, *31*.