Improved Causal Discovery from Longitudinal Data Using a Mixture of DAGs

Eric V. Strobl   ERICVONSTROBL@GMAIL.COM

Department of Psychiatry
Vanderbilt University Medical Center
Nashville, TN 37232, USA

Abstract

Many causal processes in biomedicine contain cycles and evolve. However, most causal discovery algorithms assume that the underlying causal process follows a single directed acyclic graph (DAG) that does not change over time. The algorithms can therefore infer erroneous causal relations with high confidence when run on real biomedical data. In this paper, I relax the single DAG assumption by modeling causal processes using a mixture of DAGs so that the graph can change over time. I then describe a causal discovery algorithm called Causal Inference over Mixtures (CIM) to infer causal structure from a mixture of DAGs using longitudinal data. CIM improves the accuracy of causal discovery on both real and synthetic clinical datasets even when cycles, non-stationarity, non-linearity, latent variables and selection bias exist simultaneously. Code is available at https://github.com/ericstrobl/CIM.

Keywords: Causal discovery, Longitudinal data, Directed acyclic graph, Mixture of DAGs

1. The Problem

Causal discovery refers to the process of inferring causal relations from data. Biomedical scientists usually discover causal relations using randomized controlled trials (RCTs). However, RCTs can be unethical, non-generalizable, time-consuming or expensive. Consider for example trying to discover the long term effects of an illicit substance. Randomly administering a potentially dangerous substance to human subjects is unethical, so scientists often resort to animal studies knowing that the results may not generalize to humans. Scientists may also have trouble implementing RCTs that examine the causal effects of trauma, homelessness or other complex social situations in at-risk populations. We would however like to discover causal relations even in those cases because they play an important role in the practice of medicine.

In this paper, we will describe a method for inferring causation directly from human observational data, or data collected from human subjects without the need for randomized assignment. Denote the variables in the observational dataset with the bolded letter $\boldsymbol{X}$. We can summarize the causal relationships between the variables in $\boldsymbol{X}$ using a *directed graph*. A directed graph contains directed edges between the variables in $\boldsymbol{X}$. For any two variables $X_i, X_j \in \boldsymbol{X}$, we have the directed edge $X_i \to X_j$ if X_i is a *direct cause* of X_j. A sequence of directed edges is called a *directed path*. We have a directed path from X_i to X_j if X_i is a *cause* of X_j (not necessarily direct). A directed graph contains a *cycle* or *feedback loop* when there exists a directed path from X_i to X_j as well as a directed path from X_j to X_i. A directed graph is called a *directed acyclic graph* (DAG) if it does not contain cycles. We provide an
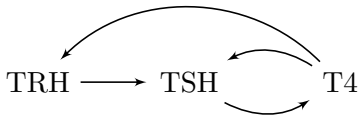
Figure 1: Part of the thyroid system depicted as a directed graph. The thyroid system is an example of a causal process in biomedicine involving feedback loops.

example of a directed graph representing a clinically relevant causal process in Figure 1. The figure depicts part of the thyroid system, where TRH release from the hypothalamus is a direct cause of TSH release from the pituitary which in turn is a direct cause of T4 release from the thyroid gland. The T4 hormone in turn invokes a negative feedback loop by directly causing an inhibition of TRH and TSH release. Note then that TRH release is a cause of T4 release but not a direct cause. Moreover, the directed graph in Figure 1 contains two cycles.

Many methods currently exist for recovering the underlying directed graph from observational data. Most of these methods nevertheless impose assumptions which do not apply to causal processes in biomedicine. For example, some of the most widely used algorithms assume an underlying DAG even though many causal processes in biomedicine are known to contain cycles (Spirtes et al., 2000); we have already provided an example of a process with feedback loops in the previous paragraph. The parathyroid system, cortisol system and multiple neural pathways within the brain and spinal cord contain feedback as well (Habener, 1981; Tasker and Herman, 2011; Saalmann et al., 2007; Cervero et al., 1984). At the single cell level, we may cite the cell cycle targeted by chemotherapeutics as well as glycogenesis/glycogenolysis targeted by metformin and glucagon as additional examples (Baynes and Dominiczak, 2018; Evan and Vousden, 2001). Unfortunately, only a handful of causal discovery algorithms can handle feedback loops.

Among those algorithms that can handle feedback loops, even fewer can also handle *non-stationary distributions*, or distributions which change over time. Most causal discovery algorithms assume that the joint distribution over $\boldsymbol{X}$ is *stationary*. Stationarity however is also violated frequently in biomedicine, a fact which we can appreciate using almost any sizable longitudinal dataset. We take a public longitudinal dataset from the Framingham Heart Study as an example (Mahmood et al., 2014). The dataset contains variables that are routinely measured by primary care physicians over three waves. We plot the empirical cumulative distribution functions (CDFs) of total cholesterol, systolic blood pressure, BMI and glucose across the three waves in Figure 2. If the joint distribution is stationary, we expect the same marginal CDFs across the waves for all 4 variables. However, we reject the null of equivalent CDFs in eight of the 12 pairwise comparisons (4 variables each with 3 comparisons between the waves) as assessed by Kolmogorov-Smirnov tests with a Bonferonni corrected threshold of 0.05/12. Only the marginal distribution of BMI remains stationary throughout all 3 waves. We conclude that the marginal distributions of even some of the most routinely measured variables in medicine are non-stationary; hence, the joint distribution over $\boldsymbol{X}$ is also non-stationary.[1]

Clinical Significance   The purpose of this paper is to introduce an algorithm called Causal Inference over Mixtures (CIM) for performing causal discovery with longitudinal data even

---

1. Recall that the same joint distribution implies the same marginal distributions because the joint distribution uniquely determines the marginals. By contrapositive then, different marginal distributions imply different joint distributions.
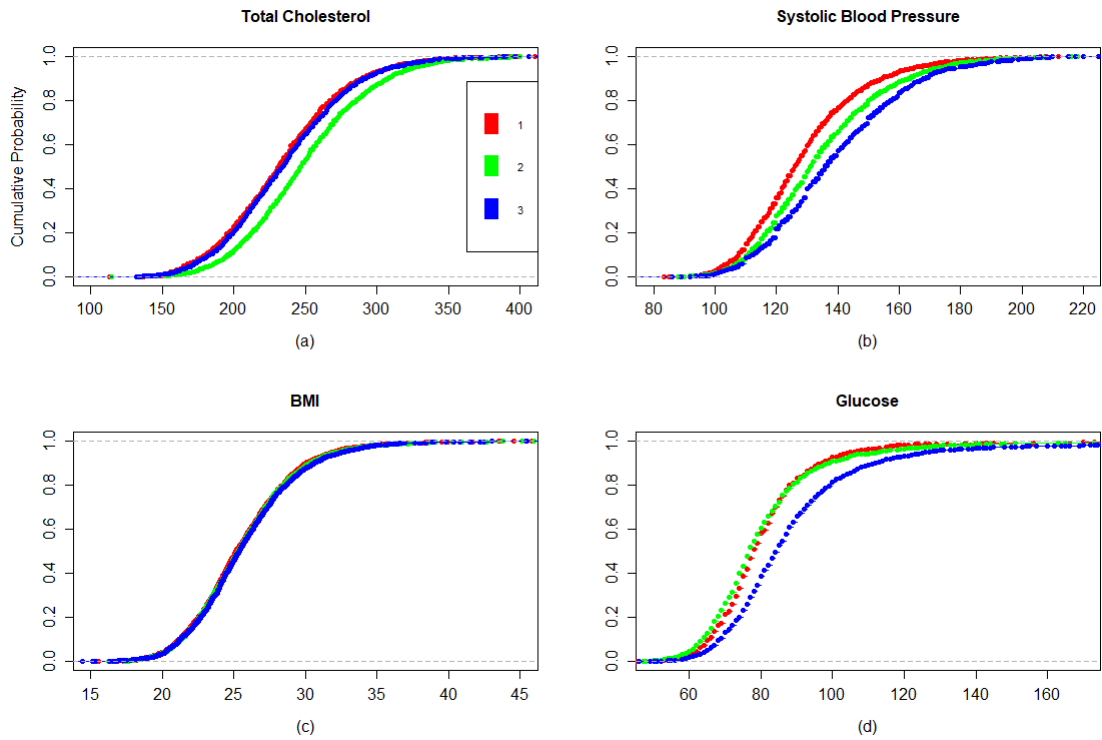
Figure 2: Empirical CDFs of four variables routinely measured in clinical practice. The measurements were obtained from a longitudinal dataset as part of the Framingham Heart Study across 3 waves separated by roughly 6 years each. Only the marginal distribution of BMI remains stationary across all of the waves.

when cycles and non-stationary distributions exist. CIM can also handle other issues which arise with real data including latent variables, selection bias and/or non-parametric distributions. CIM therefore enables more principled causal discovery from messy biomedical data by dropping the restrictive assumptions required by past methods.

Technical Significance CIM achieves the aforementioned feat while maintaining a solid theoretical foundation that guarantees soundness with a conditional independence (CI) oracle. In particular, we propose to represent cycles and non-stationary distributions using a mixture of probabilistic DAGs. We then derive the corresponding *global Markov property* that allows us to read off the CI relations in the joint distribution directly from a graph. This in turn enables the design of the CIM algorithm for the cyclic and non-stationary setting.

We organize the paper as follows. We start by surveying related past works in Section 2. We then provide background material in Section 3. In Section 4, we represent cycles as a mixture of acyclic causal processes. Next, we derive the global Markov property of the proposed representation in Section 5. We then design the CIM algorithm in Section 6 in order to recover causal structure using the global Markov property as well as longitudinal data. Experimental results in Section 7 on both synthetic and real data confirm the utility of the approach even when cycles, non-stationarity, non-linearity, latent variables and selection bias exist simultaneously. We finally conclude the paper in Section 8. The longer proofs are located in the Appendix.

2. Related Past Works

Multiple algorithms currently exist for performing causal discovery with cycles. The Cyclic Causal Discovery (CCD) algorithm recovers causal structure even when cycles exist, but the method assumes linearity, no latent variables, no selection bias and stationarity (Richardson and Spirtes, 1999; Richardson, 1996). Cyclic Causal Inference (CCI) extends CCD to handle latent variables as well as selection bias, but CCI still requires linearity as well as stationarity (Strobl, 2018). Other algorithms utilize satisfiability solvers or answer set programming (ASP), but these methods are generally limited to datasets with less than 10 variables due to scalability issues (Hyttinen et al., 2013, 2914). Both of these methods also assume linearity and stationarity.

The Fast Causal Inference (FCI) algorithm developed in (Spirtes et al., 2000; Zhang, 2008) can recover causal structure with latent variables, selection bias, *non-linearity and cycles* under an acyclic transformation of the directed graph called the *collapsed graph*. This is a previously unrecognized result, so we provide the proof in Proposition 3 in Appendix 9.1. FCI can still work with non-linear cycles because the collapsed graph eliminates all information within the cycles, so the problem of causal discovery with cycles simplifies to that without cycles. A similar approach based on ASP can perform causal discovery with cycles under non-linearity, but the method takes the same approach as FCI by utilizing the collapsed graph (Forré and Mooij, 2017, 2018); hence, it too cannot recover any causal relations within the cycles. Moreover, neither FCI nor the ASP method can recover causal structure with non-stationary distributions.

Other approaches exist for handling non-stationarity and cycles simultaneously. Most of these methods nonetheless require longitudinal or time series data and assume that the probability distribution is faithful to a single directed graph across the time steps (Dagum et al., 1995; Blondel et al., 2017). Even if we simplify the problem by assuming no cycles, existing

methods still fall short under non-stationarity alone because they also assume a single DAG model (Zhang et al., 2017; Magliacane et al., 2016; Triantafillou and Tsamardinos, 2015). This is a serious issue because random simulations frequently violate faithfulness when non-stationarity holds (although violating faithfulness is difficult in the stationary setting at the population level (Uhler et al., 2013)). For example, consider sampling from the two DAGs $X_1 \to X_2 \quad X_3$ and $X_1 \quad X_2 \to X_3$. Let the variable $T = 1$ index the first DAG and $T = 2$ the second. The single composite DAG $X_1 \to X_2 \to X_3$ used in the aforementioned works implies that we have $X_1 \not\perp\!\!\!\perp X_3$. However, we actually have $X_1 \perp\!\!\!\perp X_3$ because $f(X_1, X_3) = f(X_1) \sum_T f(T)f(X_3|T) = f(X_1)f(X_3)$. Notice that this is not a fancy counterexample - it is just a directed path with two edges. We conclude that previous works can fail to recover causal structure when the joint distribution cannot be modeled by a single directed graph.

We are only aware of three methods which can recover causal structure with cycles and a potentially changing graph. Thiesson et al. (1998) proposed the first algorithm which utilizes Bayesian modeling in conjunction with the EM algorithm, but the method cannot handle latent variables. The mixture modeling also inhibits a straightforward extension of the method to the non-parametric setting even in the linear case. Two other algorithms can handle latent variables by replacing Bayesian modeling with CI testing, but these methods depend on mixture modeling as well and therefore do not generalize to the non-parametric setting as is (Strobl, 2017; Zhang and Glymour, 2018). The method proposed in (Strobl, 2017) also often fails to recover any causal structure when we cannot model the joint distribution with a few DAGs. No existing method can therefore recover causal structure with cycles and non-stationarity without imposing strong distributional assumptions.

In summary then, no algorithm currently exists for recovering causal structure when cycles, latent variables, selection bias, and non-linearity exist simultaneously with potentially shifting graphical structure. We will therefore propose such an algorithm in this paper in order to deal with the complexities present within real biomedical data.

3. Background Material

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

3.1 Graphical Terminology

In addition to directed edges, we also consider other edge types including: $\leftrightarrow$ (bidirected), — (undirected), $\circ\!\!\to$ (partially directed), $\circ\!-$ (partially undirected) and $\circ\!\!-\!\!\circ$ (nondirected). Notice that the edges contain three endpoint types: arrowheads, tails and circles. 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 $X_i *\!\!\to X_j \leftarrow\!\!* X_k$ as a *collider* or *v-structure* when each asterisk corresponds to an arbitrary endpoint type. A collider or v-structure is said to be *unshielded* when X_i and X_k are non-adjacent. The triple $X_i *\!\!-\!\!* X_j *\!\!-\!\!* X_k$ is conversely a *triangle* if X_i and X_k are adjacent. The vertex X_j lies within the set $\text{PDS}(X_i)$ if X_i and X_j are adjacent or there exists a path between X_i and X_j such that every triple on the path is a v-structure or a triangle. 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 $X_i \in \text{Anc}_\mathbb{G}(X_j)$ 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 $\boldsymbol{Y} \subseteq \boldsymbol{X}$ or a set of sets $\boldsymbol{Y}' = \{\boldsymbol{Y}^1, \ldots, \boldsymbol{Y}^q\}$ as follows:

$$\mathrm{Anc}_{\mathbb{G}}(\boldsymbol{Y}) = \{X_i | X_i \in \mathrm{Anc}_{\mathbb{G}}(X_j) \text{ for some } X_j \in \boldsymbol{Y}\},$$
$$\mathrm{Anc}_{\mathbb{G}}(\boldsymbol{Y}') = \{X_i | X_i \in \mathrm{Anc}_{\mathbb{G}}(X_j) \text{ for some } X_j \in \cup_{i=1}^q \boldsymbol{Y}^i\}.$$

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 will only consider 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 consider the partition $\boldsymbol{X} = \boldsymbol{O} \cup \boldsymbol{L} \cup \boldsymbol{S}$ denoting observed, latent and selection variables, respectively. We then consider a mixed graph over $\boldsymbol{O}$ which summarizes the ancestral relations in $\mathbb{G}$ using the following endpoint interpretations: we have $O_i * \!\!\to O_j$ in $\mathbb{G}^*$ if $O_j \notin \mathrm{Anc}_{\mathbb{G}}(O_i \cup \boldsymbol{S})$, and we have $O_i * \!\!-\!\! O_j$ in $\mathbb{G}^*$ if $O_j \in \mathrm{Anc}_{\mathbb{G}}(O_i \cup \boldsymbol{S})$. Any mixed graph which satisfies the above two criteria is also known as an *almost ancestral graph* (AAG) (Strobl, 2018).

3.2 Probabilistic Interpretation of a Graph

We will assume that a distribution $\mathbb{P}$ obeys a *structural equation model with independent errors* (SEM-IE) with respect to (w.r.t.) a directed graph $\mathbb{G}$. Here we have $X_i = g_i(\mathrm{Pa}_{\mathbb{G}}(X_i), \varepsilon_i)$ for all $X_i \in \boldsymbol{X}$ such that X_i is $\sigma(\mathrm{Pa}_{\mathbb{G}}(X_i), \varepsilon_i)$ measurable and $\varepsilon_i \in \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ denotes a set of mutually independent random variables (Evans, 2016). We can simulate data from an SEM-IE using the fixed point method, where we sample the independent error terms and then iteratively apply the structural equations until each random variable converges almost surely. Note that the values of each random variable may not converge to a unique fixed point for all SEM-IEs, but we only consider those SEM-IEs which do satisfy this property. We refer to the distribution reached at the fixed point as the *equilibrium distribution* $\mathbb{P}$. Notice that $\mathbb{P}$ must be stationary.

If $\mathbb{G}$ is acyclic and $\mathbb{P}$ admits a density, then $\mathbb{P}$ also satisfies the *Markov property* such that its density factorizes into the product of the conditional densities of each variable given its parents:

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

Any distribution which satisfies Markov property 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 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 the algorithm utilizes CI testing to recover some aspects of $\mathbb{G}^*$ as a consequence of the global Markov property and d-separation faithfulness.

Now if $\mathbb{G}$ contains cycles, then $\mathbb{P}$ does not necessarily follow the global Markov property w.r.t. $\mathbb{G}$. The distribution does however follow the global Markov property w.r.t. the *collapsed*

graph of $\mathbb{G}$ constructed by (1) removing all edges within each cycle group, or vertices involved in intersecting cycles, (2) arbitrary numbering the vertices in each cycle group, and then (3) drawing directed edges from the lower to higher numbered vertices in each cycle group (Spirtes, 1995). Moreover, if the structural equations are all linear, then Spirtes (1995) showed that $\mathbb{P}$ more specifically follows the global Markov property w.r.t. the original directed graph $\mathbb{G}$.

4. Cycles as Mixtures of Acyclic Causal Processes

We often cannot model causal processes in biomedicine using equilibriated SEM-IEs. In fact, the data collected in the Framingham Heart Study contradicts the SEM-IE setup because some of the marginal CDFs change over time. The change implies that the values of the variables also change over time, so almost sure equality likely does not hold. We also cannot prescribe the change to measurement error because the investigators implemented a standardized measurement protocol over the waves (Cheng et al., 2012). As a result, any measurement error should have the same distribution over time, so we should at least have equality in the CDFs (i.e., equality in distribution) over the waves across all four variables if the SEM-IE representation holds. Third, we know that systolic blood pressure increases with age primarily due to large artery stiffness (O'Rourke et al., 2002). Any claim for a stationary distribution thus automatically conflicts with known biology.

Other representations of cycles exist, but they have trouble generalizing to biomedical causal processes as well. Dynamic Bayesian networks for example usually require acyclicity to hold within each wave (Malinsky and Spirtes, 2018; Robinson and Hartemink, 2009). Dynamic Bayesian networks which allow intra-wave cycles also impose the equilibrium distribution assumption, which we already argued against in the previous paragraph. Chain graphs require an equilibrium distribution as well (Lauritzen and Richardson, 2002). Furthermore, chain graphs do not represent cycles with directed edges, so we lose directionality information within the cycles which we would ideally like to preserve.

We need a different representation of cycles that can handle non-stationary distributions while preserving information within the cycles. In this report, we will represent a cyclic causal process over time as a set of acyclic ones. Intuitively, a feedback loop occurs when we iteratively "cycle through" or "unravel" the variables in the feedback loop. We can conceptualize this intuition using an example. Consider the cyclic graph involving two variables represented in Figure 3 (a). We will decompose this cycle into two DAGs: $X_1 \rightarrow X_2$ and $X_2 \rightarrow X_1$. In particular, say X_1 first causes X_2 at time point 1. We therefore have the DAG $X_1 \rightarrow X_2$ and obtain samples from $f(X_1, X_2|T = 1) = f(X_2|X_1, T = 1)f(X_1|T = 1)$. We include two samples from $f(X_1, X_2|T = 1)$ in the dataset depicted in Table 3 (b). After X_1 causes X_2, X_2 causes X_1 at time point 2. We thus have the DAG $X_2 \rightarrow X_1$ and obtain samples from $f(X_1, X_2|T = 2) = f(X_1|X_2, T = 2)f(X_2|T = 2)$. We include three samples from $f(X_1, X_2|T = 2)$ as the last three samples in Table 3 (b). We ultimately have samples from the distribution $f(X_1, X_2, T)$ in Table 3 (b), where $T = 1$ or $T = 2$. We now have a clear understanding of the sampling process from the cycle in Figure 3 (a); we have a mixture of samples from two probabilistic DAGs because sometimes we obtain samples when X_1 causes X_2, and other times we obtain samples when X_2 causes X_1.

The above example unfortunately fails to model the complexities of real data in several respects. For example, the causal process may cycle through X_1 and X_2 many more times than once. Second, we may not observe the time variable T but instead sample from $f(X_1, X_2) =$

$$X_1 \rightleftarrows X_2$$

(a)
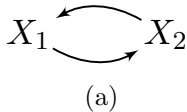
X_1	X_2	T
0.21	-0.20	1
0.68	-0.47	1
1.05	-0.19	2
0.72	-1.40	2
0.13	-0.56	2

(b)

X_1	X_3	X_4	X_7	$\cdots$	T
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)

Figure 3: We decompose the cycle in (a) into two DAGs: $X_1 \rightarrow X_2$ and $X_2 \rightarrow X_1$ at time points 1 and 2, respectively. The first two samples in the table in (b) are generated from the joint density $f(X_1, X_2 | T = 1)$ which factorizes according to $X_1 \rightarrow X_2$. Similarly, the next three are generated from $f(X_1, X_2 | T = 2)$ which factorizes according to $X_2 \rightarrow X_1$. The table in (c) depicts a more realistic dataset containing many more variables and samples.

$\sum_T f(X_1, X_2, T)$ by treating T as a latent variable. Moreover, the ordering of the samples may be scrambled, so we cannot say that any two adjacent samples are taken from similar time points. Fourth, we may have samples from many different time points potentially taken from a continuous rather than a discrete distribution. Fifth, causal processes may involve many more variables in $\boldsymbol{X}$, and we may only observe a subset of them. Some of the variables may also be discrete and others may be continuous. A real dataset therefore looks more like Table 3 (c) than Table 3 (b), where T may or may not be observed. We can fortunately handle all of the aforementioned challenges under the proposed framework, provided that we generalize it from the example given in the previous paragraph.

4.1 The Mixture of DAGs Framework

We develop the new framework as follows. We consider the set of vertices $\boldsymbol{Z} = \boldsymbol{X} \cup T$, where we may divide up $\boldsymbol{Z}$ 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 then consider the joint density $f(\boldsymbol{X}, T = t)$ and assume that it factorizes according to a DAG $\mathbb{G}^t$ over $\boldsymbol{Z}$:

$$f(\boldsymbol{Z}) = f(T)f(\boldsymbol{X}|T)$$
$$= f(T)\prod_{i=1}^{p} f(X_i | \text{Pa}_{\mathbb{G}^T}(X_i), T). \tag{2}$$

Now the set $\text{Pa}_{\mathbb{G}^T}(X_i)$ may not change if we vary the value of T over its codomain. Moreover, we may have $f(X_i | \text{Pa}_{\mathbb{G}^T}(X_i) \setminus T, T) = f(X_i | \text{Pa}_{\mathbb{G}^T}(X_i) \setminus T)$. Let $\boldsymbol{X}^{\emptyset}$ denote the set of variables satisfying the above two critiera. We may write the following for those $X_i \in \boldsymbol{X}^{\emptyset}$:

$$f(X_i | \text{Pa}_{\mathbb{G}^T}(X_i), T) = f(X_i | \text{Pa}_{\mathbb{G}^T}(X_i))$$
$$= f(X_i | \text{Pa}_{\mathbb{G}^{\emptyset}}(X_i)),$$

where $\mathbb{G}^{\emptyset}$ refers to a single graph in $\mathbb{G}'_T$. We use the notation $\mathbb{G}'_T$ to denote the set of unique DAGs over $\boldsymbol{Z}$ indexed by T; $\mathbb{G}'_T$ must have finite size because the cardinality of $\boldsymbol{Z}$ is finite. Hence, we may rewrite Equation (2) as follows:

$$f(T)\prod_{i=1}^{p} f(X_i|\mathrm{Pa}_{\mathbb{G}^T}(X_i), T)$$

$$=f(T)\prod_{i=1}^{r} f(X_i|\mathrm{Pa}_{\mathbb{G}^T}(X_i))\prod_{i=1}^{u} f(X_i|\mathrm{Pa}_{\mathbb{G}^{\emptyset}}(X_i)). \tag{3}$$

where we assume that $T \in \mathrm{Pa}_{\mathbb{G}^T}(X_i)$ for all $X_i \in [\boldsymbol{X} \setminus \boldsymbol{X}^{\emptyset}]$, and $T \notin \mathrm{Pa}_{\mathbb{G}^{\emptyset}}(X_i)$ for all $X_i \in \boldsymbol{X}^{\emptyset}$.

Notice how the above equation differs from Equation (1) for a single DAG; the parent set $\mathrm{Pa}_{\mathbb{G}}(X_i)$ remains constant over time in Equation (1) but the parent set $\mathrm{Pa}_{\mathbb{G}^T}(X_i)$ may vary over time in Equation (3). We thus may have $\mathbb{G}^{t_1} \neq \mathbb{G}^{t_2}$ for any two time points $t_1 \neq t_2$. We assume then that we can at least sample from $f(\boldsymbol{O}|\boldsymbol{S})$, the mixture density defined as follows:

$$f(\boldsymbol{O}|\boldsymbol{S}) = \sum_{T} f(\boldsymbol{O}|T, \boldsymbol{S})f(T|\boldsymbol{S}),$$

where mixing occurs over time T in the integration. We will refer to the above equation as the *mixture of DAGs* framework.

Now observe that Equation (3) can handle both stationary and non-stationary densities. We more technically say that a density is *stationary* when we have $f(\boldsymbol{X}|T) = f(\boldsymbol{X})$ or equality in density over time. Equation (3) simplifies to the usual factorization of a DAG when the density over $\boldsymbol{X}$ is stationary because in that case we have:

$$f(\boldsymbol{Z}) = f(T)f(\boldsymbol{X})$$

$$= f(T)\prod_{i=1}^{p} f(X_i|\mathrm{Pa}_{\mathbb{G}}(X_i))$$

$$= \prod_{i=1}^{p+1} f(Z_i|\mathrm{Pa}_{\mathbb{G}}(Z_i)).$$

Equation (3) can also handle non-stationary densities even when a cycle is *not* present. This can be important for example when causal processes take a long time to complete. We may have $X_i \rightarrow X_j$ in one DAG and no directed path from X_i to X_j in another because X_i takes several years to cause X_j. Thus, X_i lies in the parent set of X_j in the first DAG but not in the second one. A concrete example involves high glucose levels and peripheral sensation in type 2 diabetics, where high glucose levels decrease peripheral sensation in older individuals but not in the younger ones (Said, 2007).

4.2 An Improved Factorization

We can derive a global Markov property using Equation (3). However, the property will not imply many CI relations if $\boldsymbol{X} \setminus \boldsymbol{X}^{\emptyset}$ is large. This result would thus be incongruent with empirical results on real data, where we frequently observe non-stationarity but fail to reject many CI relations. We hypothesize therefore that T may directly index multiple independent random variables in nature. The above formulation unfortunately fails to capture those independent

variables indexed by time because Equation (3) lumps all of them into T. We therefore more specifically consider the set $\boldsymbol{M} = \{M_1, \ldots, M_s\}$ of mutually independent random variables and assume that nature has an instantiation of $\boldsymbol{M}$ at every time point; in other words, we have $f(\boldsymbol{M}|T) = f(\boldsymbol{M}) = \prod_{i=1}^{s} f(M_i)$. The set $\boldsymbol{M}$ may for example correspond to s feedback loops cycling independently. Working with $\boldsymbol{M}$ instead of T directly will allow us to derive a more fine-grained factorization of the joint density.

We now consider the set of vertices $\boldsymbol{Z} = \boldsymbol{X} \cup \boldsymbol{M}$ instead of the original $\boldsymbol{X} \cup T$. We may divide up $\boldsymbol{Z}$ into three non-overlapping sets $\boldsymbol{O}$, $\boldsymbol{L}$ and $\boldsymbol{S}$ as before so that $\boldsymbol{Z} = \boldsymbol{O} \cup \boldsymbol{L} \cup \boldsymbol{S}$. At each time point t, we assume an instantion of $\boldsymbol{M}$ and a joint density $f(\boldsymbol{X}, \boldsymbol{M})$ that factorizes according to a DAG $\mathbb{G}^{\boldsymbol{M}}$ over $\boldsymbol{Z}$:

$$
\begin{aligned}
f(\boldsymbol{Z}) &= f(\boldsymbol{M}) f(\boldsymbol{X}|\boldsymbol{M}) \\
&= \prod_{i=1}^{s} f(M_i) \prod_{i=1}^{p} f(X_i|\mathrm{Pa}_{\mathbb{G}^{\boldsymbol{M}}}(X_i), \boldsymbol{M}).
\end{aligned} \tag{4}
$$

Notice that the above equation mirrors Equation (2). Now the set $\mathrm{Pa}_{\mathbb{G}^{\boldsymbol{M}}}(X_i)$ may not change if we vary the values of $\boldsymbol{M} \setminus \boldsymbol{N}_i$ given any value of $\boldsymbol{N}_i \subseteq \boldsymbol{M}$. Moreover, we may have $f(X_i|\mathrm{Pa}_{\mathbb{G}^{\boldsymbol{M}}}(X_i) \setminus \boldsymbol{N}_i, \boldsymbol{N}_i) = f(X_i|\mathrm{Pa}_{\mathbb{G}^{\boldsymbol{M}}}(X_i) \setminus \boldsymbol{N}_i)$. We can therefore write the following when those two criteria hold:

$$
\begin{aligned}
f(X_i|\mathrm{Pa}_{\mathbb{G}^{\boldsymbol{M}}}(X_i), \boldsymbol{N}_i) &= f(X_i|\mathrm{Pa}_{\mathbb{G}^{\boldsymbol{M}}}(X_i)) \\
&= f(X_i|\mathrm{Pa}_{\mathbb{G}^{\boldsymbol{N}_i}}(X_i)),
\end{aligned}
$$

where $\mathbb{G}^{\boldsymbol{N}_i}$ refers to a single graph within $\mathbb{G}'_{\boldsymbol{N}_i}$, a subset of $\mathbb{G}'_{\boldsymbol{M}}$ containing all those DAGs whose parents of X_i do not change when we vary the values of $\boldsymbol{M} \setminus \boldsymbol{N}_i$ given any value of $\boldsymbol{N}_i$. The set $\mathbb{G}'_{\boldsymbol{M}} = \mathbb{G}'_T$ in turn refers to the set of unique DAGs over $\boldsymbol{Z}$ indexed by $\boldsymbol{M}$. Hence, we may rewrite Equation (4) as follows:

$$
\begin{aligned}
&\prod_{i=1}^{s} f(M_i) \prod_{i=1}^{p} f(X_i|\mathrm{Pa}_{\mathbb{G}^{\boldsymbol{M}}}(X_i), \boldsymbol{M}) \\
&= \prod_{i=1}^{s} f(M_i) \prod_{i=1}^{p} f(X_i|\mathrm{Pa}_{\mathbb{G}^{\boldsymbol{N}_i}}(X_i)),
\end{aligned} \tag{5}
$$

where we assume that all members of $\boldsymbol{N}_i$ and no members of $\boldsymbol{M} \setminus \boldsymbol{N}_i$ are contained within $\mathrm{Pa}_{\mathbb{G}^{\boldsymbol{N}_i}}(X_i)$ for all $X_i \in \boldsymbol{X}$ similar to Equation (3). We will write Equation (5) as follows for shorthand:

$$
\prod_{i=1}^{p+s} f(Z_i|\mathrm{Pa}_{\mathbb{G}^{\boldsymbol{N}_i}}(Z_i)), \tag{6}
$$

where $\mathrm{Pa}_{\mathbb{G}^{\boldsymbol{N}_i}}(Z_i) = \mathrm{Pa}_{\mathbb{G}^{\emptyset}}(Z_i) = \emptyset$ for all $Z_i \in \boldsymbol{M}$.

We assume then that we can at least sample from $f(\boldsymbol{O}|\boldsymbol{S})$, the mixture density defined as follows:

$$
f(\boldsymbol{O}|\boldsymbol{S}) = \sum_{\boldsymbol{M}} f(\boldsymbol{O}|\boldsymbol{M}, \boldsymbol{S}) f(\boldsymbol{M}|\boldsymbol{S}),
$$

Notice that the mixing now occurs over $\boldsymbol{M}$ instead of T because we treat $\boldsymbol{M}$ as a set of variables indexed by T. We therefore also refer to $\boldsymbol{M}$ as the set of *mixture variables*.
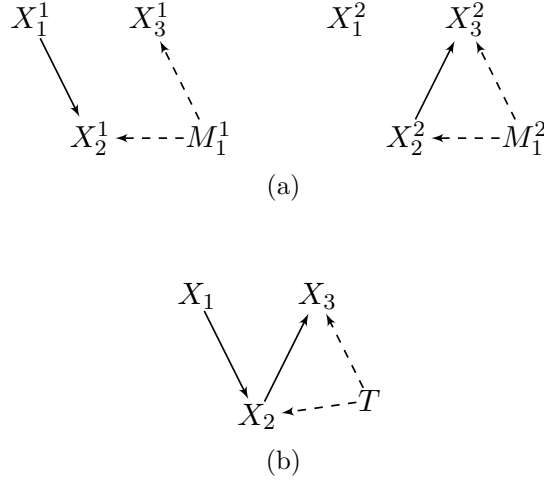
(a)

(b)

Figure 4: The mother graph versus the father graph. We plot the two DAGs over $\boldsymbol{Z}$ next to each to create the mother DAG in (a). In (b), we draw the father graph which misses the independence relation $X_1 \perp\!\!\!\perp X_3$.

5. The Global Markov Property for Mixtures of DAGs

The factorization in Equation (6) 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. The basic idea is to plot the DAGs in $\mathbb{G}'_{\boldsymbol{M}}$ next to each other to form a *mother DAG $\mathcal{M}$*. We then read off the implied CI relations from $\mathcal{M}$ by utilizing d-separation across groups of variables rather than just singletons.

We provide an example in Figure 4 (a) before providing a rigorous characterization. Suppose $\mathbb{G}'_{\boldsymbol{M}}$ contains two DAGs. Let the superscripts of the vertices index the number of each DAG. We can plot each of the unique DAGs over $\boldsymbol{Z} = \{X_1, X_2, X_3, M_1\}$ next to each other as in Figure 4 (a) and call the resultant graph the mother DAG $\mathcal{M}$. Notice that we have $X_1^1 \to X_2^1$ in the first DAG while we have $X_2^2 \to X_3^2$ in the second; however, we do not have the directed path $X_1^j \to X_2^j \to X_3^j$ in either DAG. Now suppose we want to determine if $X_1 \perp\!\!\!\perp X_3$. We can read off the d-separation relation between $\{X_1^1, X_1^2\}$ and $\{X_3^1, X_3^2\}$ in $\mathcal{M}$. The vertices are d-separated in $\mathcal{M}$, so the relation $X_1 \perp\!\!\!\perp X_3$ is implied by the global Markov property of $\mathcal{M}$.

We now make the idea rigorous. Denote the number of unique DAGs indexed by $\boldsymbol{M}$ with $q \in \mathbb{N}^+$. We will place q copies of $\boldsymbol{Z}$ into the set (of sets) $\boldsymbol{Z}'$ so that we have $\boldsymbol{Z}' = \{\boldsymbol{Z}^1, \ldots, \boldsymbol{Z}^q\}$. We also write $\boldsymbol{A}'$ to mean the set $\{\boldsymbol{A}^1, \ldots, \boldsymbol{A}^q\}$ for any subset $\boldsymbol{A} \subseteq \boldsymbol{Z}$. We will now construct the mother graph from which a directed Markov property will follow. Let $\mathbb{G}^j$ correspond to the DAG over $\boldsymbol{Z}^j$. Recall that we use $\mathbb{G}'_{\boldsymbol{M}}$ to denote the set of all q DAGs indexed by $\boldsymbol{M}$. We can associate a single DAG $\mathbb{G}^j \in \mathbb{G}'_{\boldsymbol{M}}$ for each instantiation of $\boldsymbol{M}$ by Equation (4). Note that we may have $\mathbb{G}^{\boldsymbol{m}_1} = \mathbb{G}^{\boldsymbol{m}_2}$ for two instantiations $\boldsymbol{m}_1 \neq \boldsymbol{m}_2$. Now plot each of the q DAGs in $\mathbb{G}'_{\boldsymbol{M}}$ adjacent to each other and denote the resultant graph as $\mathcal{M}$. We have the following property:

Theorem 1. *If $f(\boldsymbol{Z})$ factorizes according to Equation (6), then $f(\boldsymbol{Z})$ obeys the global Markov property with respect to (w.r.t.) $\mathcal{M}$; that is, if $\boldsymbol{A}' \perp\!\!\!\perp_d \boldsymbol{B}' | \boldsymbol{C}'$ in $\mathcal{M}$ where $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ are disjoint subsets of $\boldsymbol{Z}$, then we have $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} | \boldsymbol{C}$.*

We provide the proof in Appendix 9.2. We refer to the converse as *d-separation faithfulness* w.r.t. $\mathcal{M}$: if $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B}|\boldsymbol{C}$, then $\boldsymbol{A}' \perp\!\!\!\perp_d \boldsymbol{B}'|\boldsymbol{C}'$ in $\mathcal{M}$.

Note that Spirtes (1994) also characterized the global Markov property across a mixture of DAGs. There, the author constructed a different graph $\mathcal{F}$ (not necessarily acyclic) by first including a directed edge X_i to X_j if and only if the directed edge exists in any one of the DAGs in $\mathbb{G}'_T$. The author then introduced the vertex T and included a directed edge from T to the vertex $X_i \in \boldsymbol{X}$ whenever $f(X_i|\mathrm{Pa}_{\mathbb{G}^{t_1}}(X_i), T = t_1) \neq f(X_i|\mathrm{Pa}_{\mathbb{G}^{t_2}}(X_i), T = t_2)$ for some $t_1 \neq t_2$. We will call $\mathcal{F}$ the *father graph*. While the global Markov property does hold over $\boldsymbol{X}$ with the father graph, $\mathcal{F}$ implies less CI relations than $\mathcal{M}$. The example provided in Figure 4 also illustrates this fact. We have drawn out $\mathcal{F}$ in Figure 4 (b). Notice that X_1 and X_3 are d-connected in $\mathcal{F}$ even though $\{X_1^1, X_1^2\}$ and $\{X_3^1, X_3^2\}$ are d-separated in $\mathcal{M}$ in Figure 4 (a). We have established an instance where the mother graph implies strictly more independence relations than the father graph.

The mother graph in fact always implies at least the same number of CI relations as the father graph across all possible mixtures of DAGs:

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

We may now claim that $\mathcal{M}$ is superior to $\mathcal{F}$ because (1) $\mathcal{M}$ implies at least as many CI relations as $\mathcal{F}$ across all possible mixtures of DAGs, and (2) there exist instances where $\mathcal{M}$ implies strictly more conditional independencies. The mother graph $\mathcal{M}$ also enumerates the CI relations involving $\boldsymbol{M}$ which Spirtes (1994) did not touch.

6. Designing a Discovery Algorithm

The father graph is nonetheless more intuitive than the mother graph because $\mathcal{F}$ can summarize cycles in one directed graph. We therefore choose to utilize the global Markov property of $\mathcal{M}$ in order to recover an AAG of the father graph ($\mathcal{F}^*$). For example, suppose we have the mother graph drawn in Figure 5 (a). In this case, we assume a cycle involving $\{X_1, X_2, X_4\}$ and consider two slow causal relations: $X_2 \rightarrow X_4$ and $X_4 \rightarrow X_1$. We thus have $X_2 \rightarrow X_4$ in the first DAG in $\mathcal{M}$, but X_4 is overwritten by this causal relation, so we do not observe $X_4 \rightarrow X_1$ here. Likewise, we have $X_4 \rightarrow X_1$ in the second DAG, but X_2 is overwritten, so we do not observe $X_2 \rightarrow X_4$ in this case. We therefore cannot observe X_2 causing X_1 in either DAG due to the two rate limiting steps even though X_2 does cause X_1 in the cycle involving $\{X_1, X_2, X_4\}$. Moreover, if we intervene on the value of X_2, then X_2 cannot be overwritten in the second DAG, so we would observe X_2 causing X_1. Now we have also drawn out $\mathcal{F}$ in Figure 5 (b). Notice that X_2 is an ancestor of X_1 in $\mathcal{F}$ even though $\{X_2^1, X_2^2\}$ is not an ancestor of $\{X_1^1, X_1^2\}$ in $\mathcal{M}$. Discovering $\mathcal{F}^*$ thus allows us to infer cycles that are not present within $\mathcal{M}$ but exist once the DAGs are combined in $\mathcal{F}$.

Recovering $\mathcal{F}^*$ using the global Markov property of $\mathcal{M}$ is however not simple. We say that two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ lie within the same *Markov equivalence class* over $\boldsymbol{O}$ if and only if they share the exact same d-separation and d-connection relations over $\boldsymbol{O}$. We first have the following negative result which states that we cannot infer non-ancestral relations with a CI oracle alone:
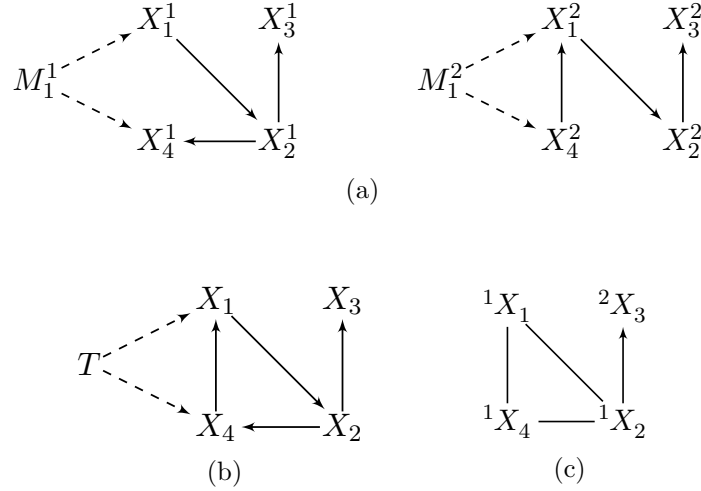
Figure 5: We have the mother graph in (a) and the father graph in (b). Subfigure (c) contains the father AAG as well as wave information.

Proposition 2. *If we have $O_i \notin \mathrm{Anc}_{\mathcal{F}}(O_j \cup \boldsymbol{S})$ in a father graph $\mathcal{F}$ with a corresponding mother graph $\mathcal{M}$, then we have $O_i \in \mathrm{Anc}_{\mathcal{F}_2}(O_j)$ in another father graph $\mathcal{F}_2$ such that its corresponding mother graph $\mathcal{M}_2$ lies within the same Markov equivalence class over $\boldsymbol{O}$ as $\mathcal{M}$.*

We fortunately however can infer ancestral relations. We therefore rely on additional information to orient arrowheads. In this paper, we will utilize *longitudinal data*, or data arising from a *longitudinal density*. Recall that we have $\boldsymbol{Z} = \boldsymbol{O} \cup \boldsymbol{L} \cup \boldsymbol{S}$. We can consider further partitioning the observed variables into w sets or *waves* so that $\boldsymbol{O} = \cup_{k=1}^{w}{}^{k}\boldsymbol{O}$. We then have the following definition:

Definition 1. *(Longitudinal density) A longitudinal density is a density $f(\cup_{k=1}^{w}{}^{k}\boldsymbol{O}, \boldsymbol{L}, \boldsymbol{S})$ that factorizes according to Equation* (6) *such that no variable in wave j is an ancestor of a variable in wave $i < j$ and $w \geq 2$.*

The restriction that no variable in wave j can be an ancestor of a variable in wave $i < j$ will allow us to infer some arrowheads instead of relying on the output of a CI oracle to do so. In contrast, we define a *cross sectional density* as the marginal density $f({}^{1}\boldsymbol{O}, \boldsymbol{L}, \boldsymbol{S})$ where $w = 1$.

Now divide each $\boldsymbol{Z}^{j} \in \boldsymbol{Z}'$ into three non-overlapping sets $\boldsymbol{Z}^{j} = \boldsymbol{O}^{j} \cup \boldsymbol{L}^{j} \cup \boldsymbol{S}^{j}$ corresponding to the observable, latent and selection variables, respectively. We require that $\boldsymbol{O}^{j} = \boldsymbol{O}^{k}$ for any $j \neq k$ but we do not require the equality for the latent and selection variables. If $\boldsymbol{Y} \subseteq \boldsymbol{Z}$, then we write $\boldsymbol{Y}'$, ${}^{k}\boldsymbol{Y}$ and ${}^{k}\boldsymbol{Y}'$ to mean $\{\boldsymbol{Y}^{1}, \dots, \boldsymbol{Y}^{q}\}$, $\boldsymbol{Y} \cap {}^{k}\boldsymbol{O}$ and $[{}^{k}\boldsymbol{Y}]'$, respectively. Let $\mathrm{Adj}_{\mathcal{F}^{*}}({}^{k}O_i)$ denote those variables in wave k adjacent to ${}^{k}O_i$ in $\mathcal{F}^{*}$. We will specifically construct $\mathcal{F}^{*}$ with the following adjacencies:

List 1. *(Adjacency Interpretations)*

1. If we have ${}^{a}O_i \mathbin{*\!\!-\!\!*} {}^{b}O_j$ (with possibly $a = b$), then ${}^{a}O'_i \not\perp_d {}^{b}O'_j | \boldsymbol{W}' \cup \boldsymbol{S}'$ in $\mathcal{M}$ for all $\boldsymbol{W} \subseteq \mathrm{Adj}_{\mathcal{F}^{*}}({}^{a}O_i) \setminus {}^{b}O_j$ and all $\boldsymbol{W} \subseteq \mathrm{Adj}_{\mathcal{F}^{*}}({}^{b}O_j) \setminus {}^{a}O_i$.

2. If we do not have ${}^{a}O_i \mathbin{*\!\!-\!\!*} {}^{b}O_j$ (with possibly $a = b$), then ${}^{a}O'_i \perp_d {}^{b}O'_j | \boldsymbol{W}' \cup \boldsymbol{S}'$ in $\mathcal{M}$ for some $\boldsymbol{W} \subseteq {}^{a}\boldsymbol{O} \setminus \{{}^{a}O_i, {}^{b}O_j\}$ or some $\boldsymbol{W} \subseteq {}^{b}\boldsymbol{O} \setminus \{{}^{a}O_i, {}^{b}O_j\}$.

For the first point, notice that we restrict $\boldsymbol{W}$ to include only those variables that are adjacent to ${}^aO_i$ or ${}^bO_j$ and within the same waves. The endpoints of $\mathcal{F}^*$ will also have the following interpretations:

List 2. *(Endpoint Interpretations)*

    1. If we have $O_i *\!\!\rightarrow O_j$, then we have $O_j \notin \operatorname{Anc}_\mathcal{F}(O_i)$.

    2. If we have $O_i *\!\!-O_j$, then we have $O_j \in \operatorname{Anc}_\mathcal{F}(O_i \cup \boldsymbol{S})$.

Note that 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 $\mathcal{F}$ using either wave information or other prior knowledge in practice. Now call $\mathcal{F}^*$ with the above adjacency and endpoint interpretations more specifically the *father AAG*. We have drawn an example of $\mathcal{M}$ in Figure 5 (a) and its corresponding father AAG in Figure 5 (c), where we let $\boldsymbol{O} = \boldsymbol{X}$, $\boldsymbol{L} = \emptyset$, $\boldsymbol{S} = \emptyset$ and $w = 2$.

We unfortunately cannot just apply a constraint-based algorithm like FCI on data arising from a mixture of DAGs and expect to recover a partially oriented father AAG. Generally speaking, FCI assumes a single underlying directed graph, so FCI may make incorrect inferences if $\mathbb{G}'_{\boldsymbol{M}}$ contains more than one DAG. More technically, FCI does not work well because $\mathcal{M}$ often implies more CI relations than $\mathcal{F}$ (Proposition 1). This in turn may cause FCI to infer incorrect arrowheads. Consider for example the mother graph in Figure 6 (a). Notice that ${}^1O_1$ is an ancestor of ${}^1O_3$ in $\mathcal{F}$ drawn in Figure 6 (b), but we have $\{{}^1O_1^1, {}^1O_1^2\} \perp\!\!\!\perp_d \{{}^1O_3^1, {}^1O_3^2\}$ in $\mathcal{M}$, so ${}^1O_1$ and ${}^1O_3$ are independent by Theorem 1. FCI therefore infers the incorrect collider ${}^1O_1 *\!\!\rightarrow {}^1O_2 \leftarrow\!* {}^1O_3$ in $\mathcal{F}^*$ during v-structure discovery. FCI indeed does orient too many colliders with real observational datasets, a phenomenon which we quantify in Section 7.2. The same problem holds with algorithms like CCI that allow cycles because CCI also assumes a single underlying directed graph. We thus require an alternative algorithm for correctly inferring $\mathcal{F}^*$.

6.1 The Algorithm

We now present an algorithm called Causal Inference over Mixtures (CIM) which recovers causal relations assuming a mixture of DAGs framework. We have summarized 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 recovers the adjacencies with interpretations listed in List 1. The algorithm stores the minimal separating sets in the array Sep so that $\operatorname{Sep}(O_i, O_j)$ contains a minimal separating set of O_i and O_j, if such a set exists. We do *not* run FCI's skeleton discovery procedure because FCI can orient erroneous unshielded triples as colliders with enough mixing as we observed in the example illustrated in Figure 6; this causes FCI's conditioning sets to grow large in practice. Notice also that Algorithm 2 only conditions on variables within one wave. We do not condition on all prior waves because we empirically find that this procedure places later waves at a disadvantage by removing too many adjacencies in the finite sample setting.
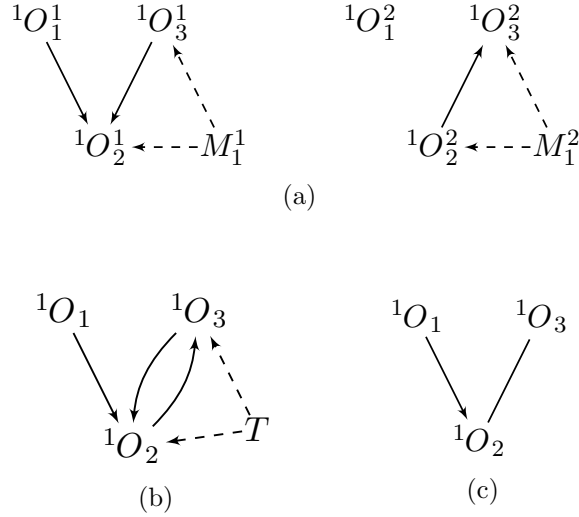
Figure 6: An example where both FCI and CCI fail. We have a mother graph in (a) and its father graph in (b). Subfigure (c) contains the correct father AAG, but FCI and CCI infer the incorrect collider $^1O_1 *\!\!\to {}^1O_2 \leftarrow\!* {}^1O_3$.

Data: CI oracle, waves $\mathcal{W}$, other prior information $\mathcal{P}$
Result: estimated father AAG $\widehat{\mathcal{F}}^*$

1 Run Algorithm 2, a variant of PC-stable's skeleton discovery procedure.

2 If we have $O_i *\!\!-\!\!\circ O_j$ and O_i lies within an earlier wave than O_j according to $\mathcal{W}$ or O_j cannot be an ancestor of O_i according to $\mathcal{P}$, then orient $O_i *\!\!-\!\!\circ O_j$ as $O_i *\!\!\to O_j$ in $\widehat{\mathcal{F}}^*$.

3 If we have $O_i *\!\!\to O_j *\!\!-\!\!* O_k$ with O_i and O_k non-adjacent, $O_j \notin \mathrm{Sep}(O_i, O_j)$ and there exists another minimal separating set $\boldsymbol{W} \subseteq \mathrm{PDS}(O_i) \setminus O_k$ containing O_j, then record $\boldsymbol{W}$ into $\mathrm{Sep2}(O_i, O_j, O_k)$.

4 If we have $O_i *\!\!\to O_j \circ\!\!-\!\!* O_k$ with O_i and O_k non-adjacent, and either $O_j \in \mathrm{Sep}(O_i, O_k)$ or $\mathrm{Sep2}(O_i, O_j, O_k)$ is non-empty, then orient $O_j \circ\!\!-\!\!* O_k$ as $O_j -\!\!* O_k$ in $\widehat{\mathcal{F}}^*$.

5 Execute the following orientation rule until no more edges can be oriented: if we have the sequence of vertices $\langle O_1, \ldots, O_n \rangle$ such that $O_i -\!\!* O_{i+1}$ with $1 \le i \le n-1$, and we have $O_1 \circ\!\!-\!\!* O_n$, then orient $O_1 \circ\!\!-\!\!* O_n$ as $O_1 -\!\!* O_n$ in $\widehat{\mathcal{F}}^*$.

Algorithm 1: Causal Inference over Mixtures (CIM)

15

The CIM algorithm next adds arrowheads in Step 2. Recall that it is impossible to orient arrowheads under a mixture of DAGs framework using a CI oracle alone according to Proposition 2. CIM therefore uses wave information from a longitudinal dataset with the list $\mathcal{W}$ that contains the indices of the variables in each wave. In particular, if we have ${}^aO_i \circ\!\!-\!\!\circ {}^bO_j$ with $b > a$, then CIM orients ${}^aO_i \circ\!\!\rightarrow {}^bO_j$ because we must have ${}^bO_j \notin \mathrm{Anc}_{\mathcal{F}}({}^aO_i)$ (although we may have ${}^bO_j \in \mathrm{Anc}_{\mathcal{F}}(\boldsymbol{S})$). We can similarly orient additional arrowheads using other prior knowledge $\mathcal{P}$. For example, systolic blood pressure cannot be an ancestor of chronological age or gender because intervening on systolic blood pressure can never change age or gender. Step 2 orients many arrowheads in practice, so long as we have at least two waves of data.

For every triple $O_i *\!\!\rightarrow O_j *\!\!-\!\!* O_k$ with O_i and O_k non-adjacent, CIM then attempts to find a minimal separating set that contains O_j in Step 3. These sets are important due to the following lemma which allows us to infer tail endpoints in Step 4:

Lemma 1. *Suppose that we have $O_i' \perp\!\!\!\perp_d O_j' | \boldsymbol{W}' \cup \boldsymbol{S}'$ in $\mathcal{M}$ but we have $O_i' \not\perp\!\!\!\perp_d O_j' | \boldsymbol{V}' \cup \boldsymbol{S}'$ for every $\boldsymbol{V} \subset \boldsymbol{W}$. If $O_k \in \boldsymbol{W}$, then $O_k \in \mathrm{Anc}_{\mathcal{F}}(\{O_i, O_j\} \cup \boldsymbol{S})$.*

Thus, unlike arrowheads, we can fortunately infer tails without additional prior knowledge. CIM finally adds some additional tails in Step 5, a step which we can justify due to the transitivity of the tails.

We now formally claim that Algorithm 1 is sound:

Theorem 2. *Suppose that the longitudinal density $f(\cup_{k=1}^w {}^k\boldsymbol{O}, \boldsymbol{L}, \boldsymbol{S})$ factorizes according to Equation (6). Assume that all arrowheads deduced from $\mathcal{P}$ are correct. Then, under d-separation faithfulness w.r.t. $\mathcal{M}$, the CIM algorithm returns a partially oriented father AAG.*

Proof. Under d-separation faithfulness w.r.t. $\mathcal{M}$, CI and d-separation w.r.t. $\mathcal{M}$ 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 $\mathrm{Adj}_{\mathcal{F}^*}({}^aO_i) \subseteq \mathrm{Adj}_{\widehat{\mathcal{F}}^*}({}^aO_i)$ in Step 9 of Algorithm 2. Step 4 discovers the correct tails by Lemma 1. Step 5 follows directly by transitivity of the tails. □ □

Note that d-separation faithfulness w.r.t. $\mathcal{M}$ is equivalent to normal d-separation faithfulness when we have a single DAG in $\mathbb{G}'_{\boldsymbol{M}}$. The d-separation faithfulness assumption imposed by Theorem 2 is therefore at least as weak as the normal d-separation faithfulness assumption for a single DAG.

CIM finally carries the same polynomial time complexity as the CCD algorithm at $O(u^4)$ due to Step 3, where the maximum number of CI tests is bounded above by $3\binom{u}{3}\sum_{i=0}^q \binom{u-3}{i} \le \frac{(q+1)u^3(u-3)^{q+1}}{q!} = O(u^4)$ with fixed $q \triangleq \max_i |\mathrm{PDS}(O_i)| - 1$ and u denoting the cardinality of $\boldsymbol{O}$. The runtime of the algorithm therefore depends on the underlying sparsity of the mother graph. Practically, we find that CIM completes in about the same amount of time as PC on a variety of longitudinal datasets with less than 10 vertices in each wave using a laptop carrying an Intel (R) Core (TM) i7-6700HQ 2.60 GHz CPU and 16GB of RAM.

6.2 Algorithm Trace

We now run through the CIM algorithm with an example. We consider the mother graph drawn in Figure 7 (a). Step 1 of CIM discovers the skeleton depicted in Figure 7 (b). In Step 2, we assume access to wave information and therefore add the arrowheads at wave 2
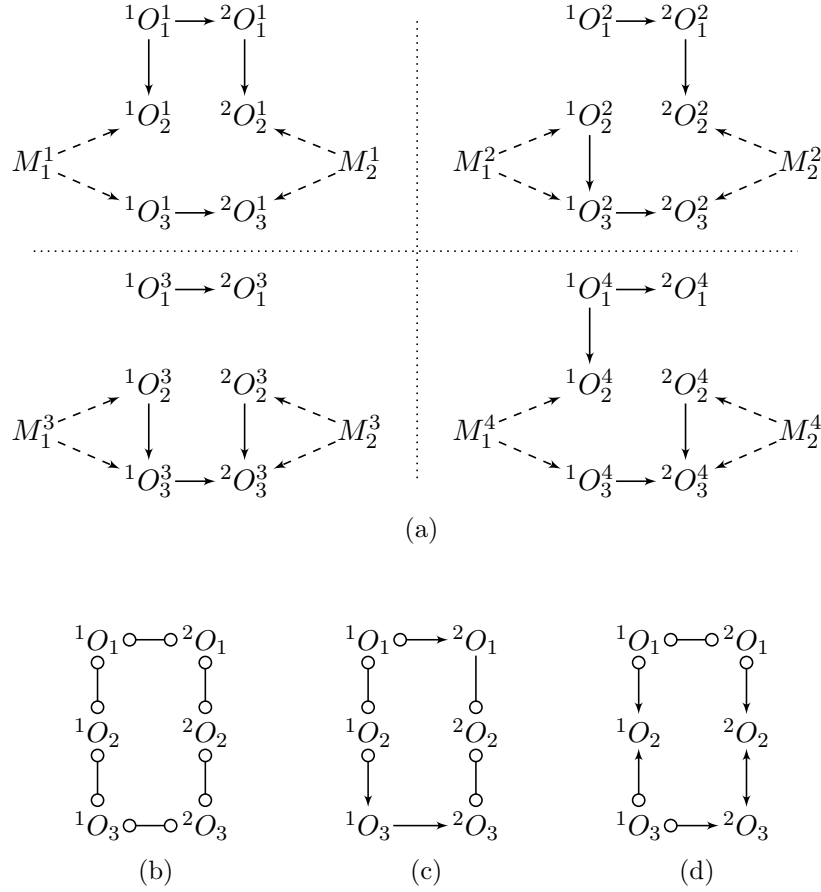
Figure 7: Algorithm trace of CIM. The ground truth mother graph in (a). Step 1 of CIM returns (b). Step 2 orients the arrowheads in (c) and Step 4 the tails. CIM ultimately returns the partially oriented father AAG in (c). In contrast, FCI and CCI return (d) which contains two incorrect arrowheads.

on the edges between wave 1 and wave 2. We may also know that $^1O_3$ is not an ancestor of $^1O_2$ using prior knowledge, so we orient the arrowheads in Figure 7 (c). We then orient the two tails in Figure 7 (c) in Step 4 because we have $\{^1O_1^j\}_{j=1}^4 \perp\!\!\!\perp_d \{^2O_2^j\}_{j=1}^4 | \{^2O_1^j\}_{j=1}^4$ and $\{^1O_2^j\}_{j=1}^4 \perp\!\!\!\perp_d \{^2O_3^j\}_{j=1}^4 | \{^1O_3^j\}_{j=1}^4$ as discovered in Step 1. CIM does not orient any endpoints in Step 5 in this case. In contrast, FCI and CCI discover the partially oriented graph in Figure 7 (d) which contains two incorrect arrowheads: $^2O_3 *\!\!\to\, ^2O_2$ and $^1O_3 *\!\!\to\, ^1O_2$.

7. Experiments

7.1 Algorithms

We compared the following five algorithms in recovering the ancestral and nonancestral relations in $\mathcal{F}$:

1. CIM

2. PC

3. FCI

4. RFCI

5. CCI

We equipped all algorithms with the nonparametric RCoT test (Strobl et al., 2018) and fixed $\alpha = 0.01$ across all experiments. The F²CI algorithm proposed in (Strobl, 2017) did not orient any endpoints on the real datasets because it assumes mixtures of Gaussians for edge orientation. We therefore do not report the results of this algorithm. PC is the canonical algorithm for causal discovery over a DAG (Spirtes et al., 2000). FCI extends PC to handle latent variables and selection bias (Spirtes et al., 2000; Zhang, 2008). RFCI speeds up FCI by utilizing smaller conditioning sets (Colombo et al., 2012). CCI in turn extends FCI to the cyclic case provided that linearity holds (Strobl, 2018).

We gave all algorithms the same wave information during skeleton discovery in order to avoid giving CIM an unfair advantage. PC, FCI, RFCI and CCI perform much worse without the additional wave information. We additionally provided CIM and PC with the arrowhead information based on the waves and prior knowledge. We could not provide this information to FCI, RFCI and CCI based on their arrowhead interpretation because we often do not know if variables are ancestors of a selection variable.

We had two overarching goals in mind. First, we wanted to evaluate the performance of CIM against the other algorithms on real data with some known ground truth. We also sought to reconstruct the real data results using synthetic data in order to evaluate the mixture of DAGs framework as a reasonable model of nature. We therefore present the real data results first.

7.2 Real Data

We utilized the following three longitudinal datasets and a priori known direct causal relations:

1. Framingham Heart Study (Mahmood et al., 2014)

    (a) number of cigarettes per day causes heart rate (via cardiac nicotonic acetylcholine receptors (Aronow et al., 1971; Levy, 1971; Haass and Kübler, 1997))

    (b) age causes systolic blood pressure (due to increased large artery stiffness (Pinto, 2007; Safar, 2005))

    (c) BMI causes number of cigarettes per day (smoking cigarettes is a common weight loss strategy (Jo et al., 2002; Chiolero et al., 2008))

2. Mayo Clinic Primary Biliary Cirrhosis (Fleming and Harrington, 1991)

    (a) bilirubin & prothrombin cause transplant status (clinical criteria now incorporated into the MELD score (Kamath et al., 2001; Kamath and Kim, 2007))

    (b) histological stage causes hepatomegaly (due to fibrosis and chronic inflammation (Lang, 2009))

    (c) transplant status causes bilirubin & prothrombin (acute decrease post transplantation (Hickman et al., 1997)), hepatomegaly (transplantation immediately reduces liver size)

3. Cognition and Aging USA (McArdle et al., 2015)

   (a) years of education causes numeracy & serial 7s (increased education improves ability to solve math problems)

   (b) age causes episodic memory (decreased episodic memory secondary to neurodegeneration (Rajah and D'esposito, 2005; Head et al., 2008))

We chose the above datasets due to the ability to identify partial ground truth using clinical knowledge. All of the datasets are publicly available. Note also that all of the aforementioned datasets contain three waves of data. We evaluated the algorithms according to known ground truth and time information; that is, (1) the ability to correctly discover the known causal relations listed above, and (2) the ability to avoid discovering incorrect causal relations directed backwards in time. For the second point, we more specifically removed all arrowheads at wave 3 on edges between waves 2 and 3. As a result, the algorithms could orient tails at wave 3 on edges between waves 2 and 3.

We summarize the results from the Framingham Heart Study over 100 bootstrapped datasets in Figure 8. CIM oriented significantly less tails directed backwards in time than PC (t=-12.13, p<2.2E-16; Figure 8 (a)). RFCI, FCI and CCI oriented few backwards tails as well, but the output of CIM contained many more tails on average than any of those algorithms (min t = 39.10, p<2.2E-16; Figure 8 (b)). Note that CCI almost never oriented tails; the algorithm oriented about 0.57 tails on average (95% CI: [0.40, 0.74]) as opposed to 34.42 for CIM (95% CI: [33.11, 35.73]). Recall that these findings are congruent with the mixture of DAGs framework, where we expect traditional algorithms to orient too many arrowheads. Third, CIM discovered more known causal relations than FCI, RFCI and CCI (min t=38.64, p<2.2E-16; Figure 8 (c)). CIM outperformed PC in this case too (t=20.05, p<2.2E-16), although this relation appeared to be dataset dependent because it did not hold in the other two datasets (Appendix Figures 12 (c) and 13 (c)). Next, CIM took the least amount of time to complete (max t=-9.02, p=1.53E-14; Figure 8 (d)). We finally replicated all of the aforementioned results in the Mayo Clinic Primary Biliary Cirrhosis and Cognition and Aging USA studies as presented in Figures 12 and 13 in Appendix 9.4, respectively. We conclude that, with real data, CIM (1) makes few errors based on time and (2) detects many known causal relations while (3) orienting many tails and (4) completing within a short time frame. In contrast, PC orients many incorrect tails, and FCI, RFCI and CCI orient too many arrowheads.

7.3 Synthetic Data

We seek to approximate the real data results using synthetic data generated from a mixture of DAGs. Empirical results on the real data suggest a sparse underlying mother graph with a high degree of mixing causing algorithms like CCI to orient relatively few tails. To account for the sparsity and the few tails, we hypothesized the existence of many variables in $\boldsymbol{M}$ and only a few children for each variable in $\boldsymbol{M}$. Thus, even if X_i causes X_j in the father graph, we may not have a directed path from X_i' to X_j' in the mother graph like in Figure 4.

We therefore sampled according to a mixture of DAGs as follows. We generated 100 probabilistic models by drawing $q = 5$ to 15 binary variables in $\boldsymbol{M}$. For each $M_i \in \boldsymbol{M}$, we then instantiated two Gaussian DAGs each with an expected neighborhood size $\mathbb{E}(N) = 1/q$ and $p = 24$ vertices. For each DAG, we permuted the ordering of the 24 vertices and then generated an upper triangular adjacency matrix $\mathcal{A}$ using independent realizations of Bernoulli($\mathbb{E}(N)/(p-1)$)
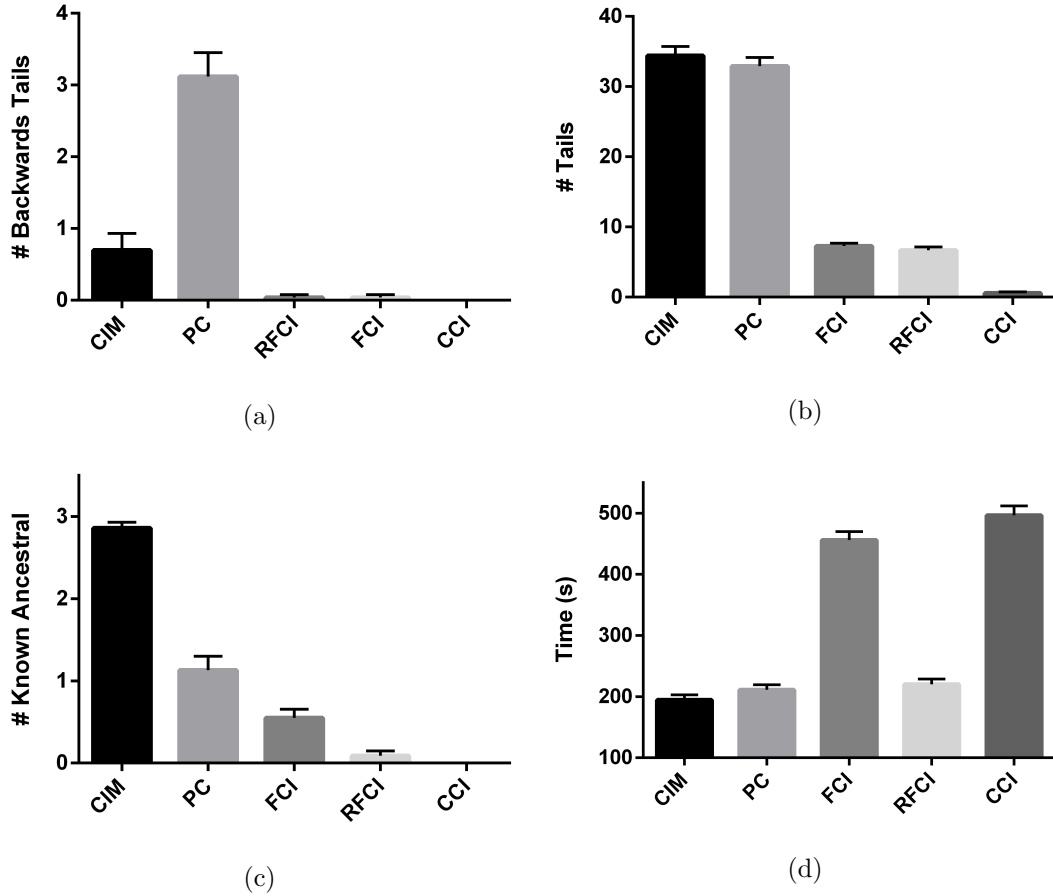
(a)

(b)

(c)

(d)

Figure 8: Framingham Heart Study results. Bar heights represent empirical means and error bars their 95% confidence intervals (CIs). (a) CIM orients less tails from wave 3 to wave 2 than PC. (b) CIM orients more tails than FCI, RFCI and CCI combined. (c) CIM detects the most known ancestral relations. (d) CIM takes the least amount of time to complete.

random variables. Restoring the variable order to before the permutation allowed us to obtain a DAG over $\boldsymbol{X}$ with an adjacency matrix not necessarily restricted to the upper triangle. We then replaced the non-zero entries in $\mathcal{A}$ with independent realizations of Uniform($[-1, -0.25] \cup [0.25, 1]$) random variables. We thus generated 10 to 30 DAGs indexed by 5 to 15 binary variables in $\boldsymbol{M}$.

We next modified each DAG to represent a longitudinal density. We assigned the first 8 variables to wave 1, the second 8 to wave 2, and the third 8 to wave 3. We then added a directed edge from the nth variable in wave 1 to the nth variable in wave 2 for the DAG associated with $M_i = 0$, and similarly added the directed edges from wave 2 to wave 3 for the DAG associated with $M_i = 1$ in order to model self-loops. Next, we deleted all edges from a higher to a lower numbered wave to remove ancestral relations directed backwards in time. We finally introduced latent and selection variables as follows. We first 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 $\boldsymbol{M}$. We then selected a set of 0-2 selection variables $\boldsymbol{S}$ without replacement from the set $\boldsymbol{X} \setminus \boldsymbol{L}$.

We generated 1000 samples from each probabilistic model as follows. We uniformly instantiated the mixing probabilities $f(M_i = 0)$ and $f(M_i = 1)$ for each $M_i \in \boldsymbol{M}$ so that we have one probability value for each of its two DAGs. For each sample, we then drew an instantiation $\boldsymbol{M} = \boldsymbol{m}$ according to $\prod_{i=1}^{q} f(M_i)$ and created the directed graph containing the union of the edges present in each of the 5 to 15 DAGs associated with $\boldsymbol{m}$. If this directed graph was cyclic, we removed a directed edge in each cycle to ensure acyclicity. We can thus associate each instantiation of $\boldsymbol{M}$ with a DAG but cycles may exist in the father graph. Repeating the above process 1000 times thus generates 1000 samples over $\boldsymbol{Z}$ according to Equation (6). 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 Figure 9. Like with the real data, CIM discovered significantly less backwards tails than PC (max t=-16.85, p<2.2E-16). CIM also oriented a larger number of tails than FCI, RFCI and CCI (min t=10.94, p<2.2E-16). We next analyzed the algorithm outputs within waves 2 and 3 as well as from waves 2 to 3, where CIM can orient tails with the synthetic data. Here, the proposed algorithm identified the most known ancestral relations based on the ground truth father graph (min t = 8.03, p=2.09E-12). PC discovered the second most known ancestral relations on average, but it also oriented many backwards tails. We therefore conclude that FCI, RFCI and CCI output too few tails just like with the real data. Moreover, PC can often output incorrect tails.

We can also compute other more common performance measures with the synthetic data because we know the ground truth father graphs. We plot the precision results in Figure 10 (a) by analyzing the endpoints within waves 2 and 3 as well as from waves 2 to 3. Notice that CIM under-performed all algorithms with this metric (max t=-5.05, p=2.02E-6). Precision however gives algorithms which orient few tails an unfair advantage because the algorithms do not need to orient any tails to achieve a perfect score. We therefore also summarize the recall results in Figure 10 (b). Notice that CIM obtained the highest recall (min t = 3.31, p=1.29E-3). CIM outperformed FCI, RFCI and CCI as well with the F1 score which combines precision and recall (min t=14.60, p< 2.2E-16; Figure 10 (c)). CIM did not achieve a higher F1 score than PC, but CIM is theoretically sound under the proposed setup with feedback loops, latent variables, selection bias and/or non-stationary distributions, whereas PC is not. We can appreciate this difference experimentally by removing the time information between
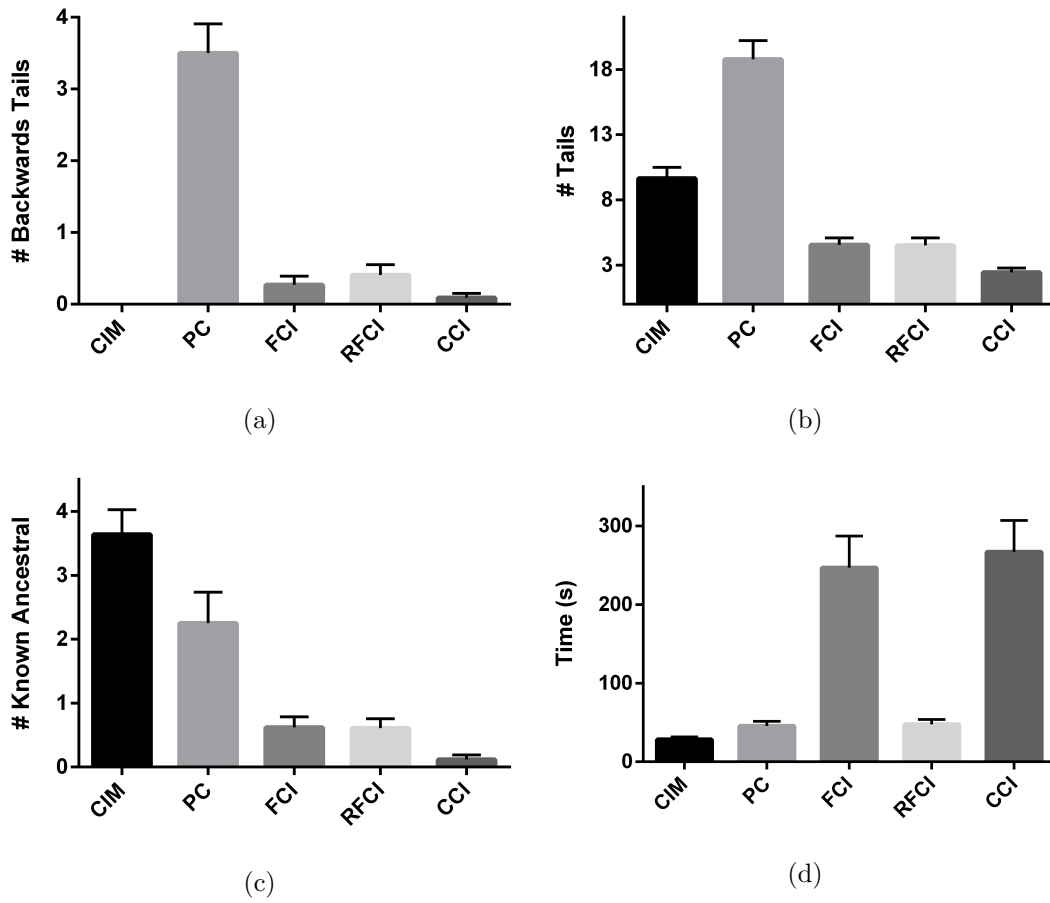
(a)

(b)

(c)

(d)

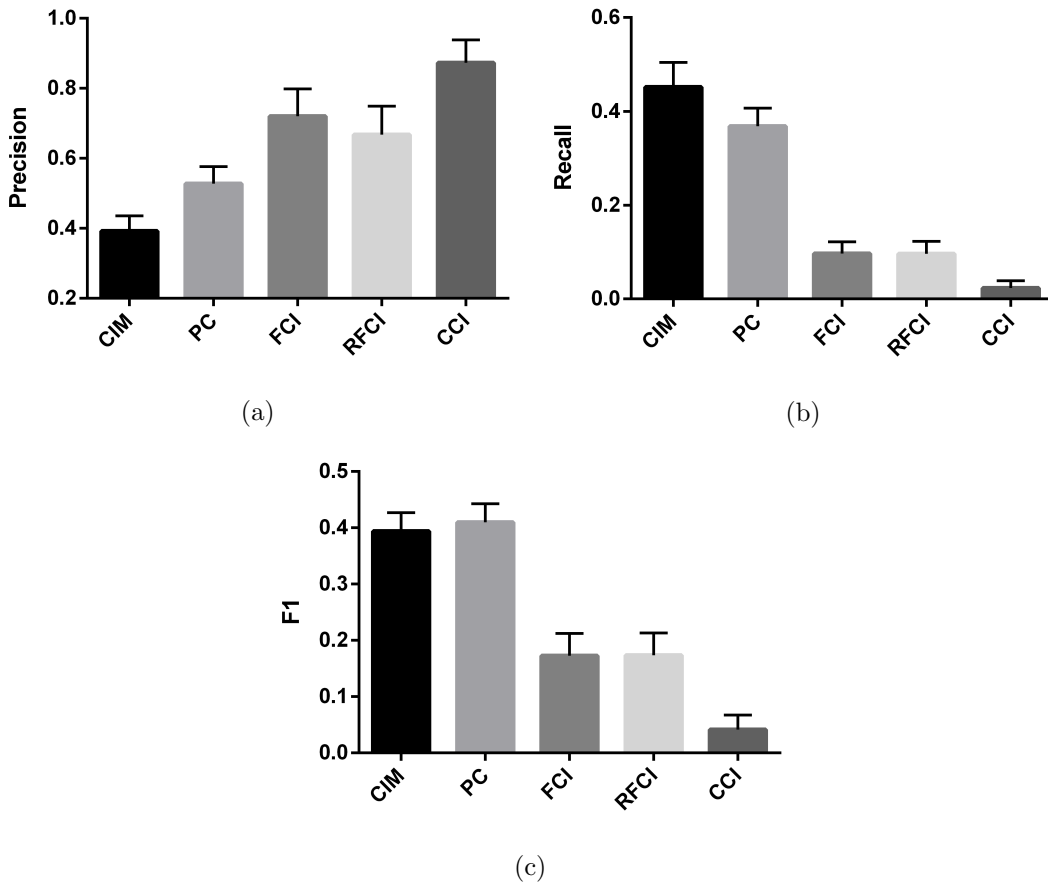Figure 9: Synthetic data results presented in the same format as Figure 8.

(a)

(b)

(c)

Figure 10: (a) Precision, (b) recall and (c) the F1 score for the synthetic data. CIM achieves the lowest precision but highest recall.

waves 2 and 3. In this case, CIM maintains an average precision of 0.507 (95% CI: [0.457, 0.557]), whereas PC drops around 0.15 points to an average precision of 0.373 (95% CI: [0.327, 0.418]; CIM vs. PC t=5.26, p=8.19E-7); PC's lower precision occurs because the algorithm discovers erroneous causal relations without the additional time information. We conclude that CIM helps users detect tails the most accurately under a mixture of DAGs framework while maintaining theoretical soundness.

Now simply sampling from a directed graph with latent and selection variables does not reproduce the real data results as well as the aforementioned setup. To demonstrate this, we generated 1000 samples each from 100 Gaussian directed graphs with $\mathbb{E}(N) = 2$ and $p = 24$ vertices again over 3 waves with 8 vertices in each wave. We connected the nth variable in wave m to the nth variable in wave $m+1$ to model the self-loops as before. We also included 0-2 latent and 0-2 selection variables by sampling the variables without replacement. We summarize the key difference in Figure 11. Notice that CIM no longer orients as many more tails than FCI, RFCI and CCI (PC oriented 21.07 tails on average, so we do not plot its results). Recall that this occurs because we in general have more CI relations with a mixture of DAGs than with a single directed graph. We conclude that sampling according to Equation (6) reproduces the
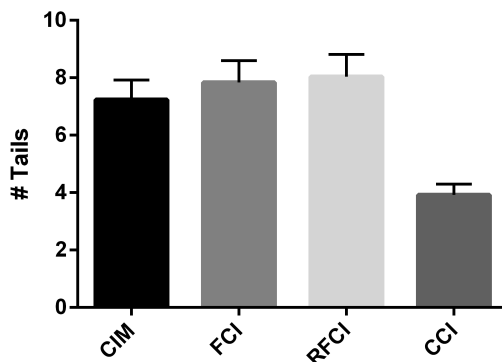
Figure 11: Number of tails recovered by the algorithms after simulating data from single directed graphs. Notice that CIM no longer orients as many more tails than the other algorithms.

real data results more accurately than sampling from a probabilistic model following a single directed graph.

8. Conclusion

We studied the problem of performing causal discovery with mixtures of DAGs. We showed that the mixture of DAGs framework helps model causal processes even when cycles, non-stationarity, non-linearity, latent variables and selection bias exist simultaneously. We then proposed an algorithm called CIM for recovering ancestral causal relations with longitudinal data under the framework. Experimental results showed that CIM more accurately discovers causal relations under a mixture of DAGs than previously proposed strategies using both simulated and real data. We conclude that the proposed ideas improve the accuracy of causal discovery under more realistic situations.

References

Wilbert S Aronow, James Dendinger, and Stanley N Rokaw. 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, 1971.

J.W. Baynes and M.H. Dominiczak. *Medical Biochemistry*. Elsevier Health Sciences, 2018. ISBN 9780702073007. URL https://books.google.com/books?id=Bs9FDwAAQBAJ.

Gilles Blondel, Marta Arias, and Ricard Gavaldà. Identifiability and transportability in dynamic causal networks. *International Journal of Data Science and Analytics*, 3(2): 131–147, Mar 2017. ISSN 2364-4168. doi: 10.1007/s41060-016-0028-8. URL https://doi.org/10.1007/s41060-016-0028-8.

Fernande Cervero et al. A positive feedback loop between spinal cord nociceptive pathways and antinociceptive areas of the cat's brain stem. *Pain*, 20(2):125–138, 1984.

S. Cheng, V. Xanthakis, L. M. Sullivan, and R. S. Vasan. Blood pressure tracking over the adult life course: patterns and correlates in the Framingham heart study. *Hypertension*, 60 (6):1393–1399, Dec 2012.

Arnaud Chiolero, David Faeh, Fred Paccaud, and Jacques Cornuz. Consequences of smoking for body weight, body fat distribution, and insulin resistance. *The American journal of clinical nutrition*, 87(4):801–809, 2008.

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

Diego Colombo, Marloes Maathius, Marcus Kalisch, and Thomas Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *Annals of Statistics*, 40(1):294–321, April 2012. doi: 10.1214/11-AOS940. URL http://projecteuclid.org/euclid.aos/1333567191.

Robert G. Cowell, Steffen L. Lauritzen, A. Philip David, and David J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer-Verlag, Berlin, Heidelberg, 1st edition, 1999. ISBN 0387987673.

Paul Dagum, Adam Galper, Eric Horvitz, and Adam Seiver. Uncertain reasoning and forecasting. *International Journal of Forecasting*, 11:73–87, 1995.

Gerard I Evan and Karen H Vousden. Proliferation, cell cycle and apoptosis in cancer. *Nature*, 411(6835):342, 2001.

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

Thomas R. Fleming and David P. Harrington. *Counting Processes and Survival Analysis*. John Wiley & Sons, 1991.

Patrick Forré and Joris M. Mooij. Markov properties for graphical models with cycles and latent variables. *arXiv.org preprint*, arXiv:1710.08775 [math.ST], October 2017. URL https://arxiv.org/abs/1710.08775.

Patrick Forré and Joris M. Mooij. 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)*, 2018.

Markus Haass and Wolfgang Kübler. Nicotine and sympathetic neurotransmission. *Cardiovascular Drugs and Therapy*, 10(6):657–665, 1997.

Joel F Habener. Regulation of parathyroid hormone secretion and biosynthesis. *Annual review of physiology*, 43(1):211–223, 1981.

Denise Head, Karen M Rodrigue, Kristen M Kennedy, and Naftali Raz. Neuroanatomical and cognitive mediators of age-related differences in episodic memory. *Neuropsychology*, 22(4): 491, 2008.

Peter E Hickman, Julia M Potter, and Amadeo J Pesce. Clinical chemistry and post-liver-transplant monitoring. *Clinical chemistry*, 43(8):1546–1554, 1997.

Antti Hyttinen, Patrik O. Hoyer, Frederick Eberhardt, and Matti Järvisalo. 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*, 2013. URL https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=2391&proceeding_id=29.

Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. 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, 2914. AUAI Press. ISBN 978-0-9749039-1-0. URL http://dl.acm.org/citation.cfm?id=3020751.3020787.

Young-Hwan Jo, David A Talmage, and Lorna W Role. Nicotinic receptor-mediated effects on appetite and food intake. *Journal of neurobiology*, 53(4):618–632, 2002.

Patrick S Kamath and W Ray Kim. The model for end-stage liver disease (meld). *Hepatology*, 45(3):797–805, 2007.

Patrick S Kamath, Russell H Wiesner, Michael Malinchoc, Walter Kremers, Terry M Therneau, Catherine L Kosberg, Gennaro D'Amico, E Rolland Dickson, and W Ray Kim. A model to predict survival in patients with end-stage liver disease. *Hepatology*, 33(2):464–470, 2001.

Florian Lang. *Encyclopedia of molecular mechanisms of disease*. Springer Science & Business Media, 2009.

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

Steffen L. Lauritzen and Thomas S. Richardson. Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):321–348, 2002. doi: 10.1111/1467-9868.00340. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00340.

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

Sara Magliacane, Tom Claassen, and Joris M. Mooij. Joint causal inference on observational and experimental datasets. *CoRR*, abs/1611.10351, 2016. URL http://arxiv.org/abs/1611.10351.

Syed S Mahmood, Daniel Levy, Ramachandran S Vasan, and Thomas J Wang. The framingham heart study and the epidemiology of cardiovascular disease: a historical perspective. *The Lancet*, 383(9921):999 – 1008, 2014. ISSN 0140-6736. doi: http://doi.org/10.1016/S0140-6736(13)61752-3. URL http://www.sciencedirect.com/science/article/pii/S0140673613617523.

Daniel Malinsky and Peter Spirtes. 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`.

J. McArdle, W. Rodgers, and R. Willis. Cognition and aging in the usa (cogusa), 2007-2009, 2015. URL `http://www.icpsr.umich.edu/icpsrweb/NACDA/studies/36053`.

M. F. O'Rourke, J. A. Staessen, C. Vlachopoulos, D. Duprez, and G. E. Plante. Clinical applications of arterial stiffness; definitions and reference values. *Am. J. Hypertens.*, 15(5): 426–444, May 2002.

Elisabete Pinto. Blood pressure and ageing. *Postgraduate medical journal*, 83(976):109–114, 2007.

M Natasha Rajah and Mark D'esposito. Region-specific changes in prefrontal function with age: a review of pet and fmri studies on working and episodic memory. *Brain*, 128(9): 1964–1983, 2005.

T. Richardson. A discovery algorithm for directed cyclic graphs. In *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence*, UAI'96, pages 454–461, San Francisco, CA, USA, 1996. Morgan Kaufmann Publishers Inc. ISBN 1-55860-412-X. URL `http://dl.acm.org/citation.cfm?id=2074284.2074338`.

T. Richardson and P. Spirtes. Automated causal discovery under linear feedback. In *Computation, Causation, and Discovery*, pages 253–302. AAAI Press, Menlo Park, CA, 1999.

Joshua W. Robinson and Alexander J. Hartemink. Non-stationary dynamic bayesian networks. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1369–1376. Curran Associates, Inc., 2009. URL `http://papers.nips.cc/paper/3571-non-stationary-dynamic-bayesian-networks.pdf`.

Yuri B Saalmann, Ivan N Pigarev, and Trichur R Vidyasagar. Neural mechanisms of visual attention: how top-down feedback highlights relevant locations. *Science*, 316(5831):1612–1615, 2007.

Michel E Safar. Systolic hypertension in the elderly: arterial wall mechanical properties and the renin–angiotensin–aldosterone system. *Journal of hypertension*, 23(4):673–681, 2005.

Gérard Said. Diabetic neuropathya review. *Nature Reviews Neurology*, 3(6):331, 2007.

P. Spirtes. Conditional independence properties in directed cyclic graphical models for feedback. Technical report, Carnegie Mellon University, 1994.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

Peter Spirtes. Directed cyclic graphical representations of feedback models. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, pages 491–498, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1-55860-385-9. URL `http://dl.acm.org/citation.cfm?id=2074158.2074214`.

Eric V Strobl. *Causal Discovery Under Non-Stationary Feedback.* PhD thesis, University of Pittsburgh, 2017.

Eric V. Strobl. A constraint-based algorithm for causal discovery with cycles, latent variables and selection bias. *International Journal of Data Science and Analytics*, Nov 2018. ISSN 2364-4168. doi: 10.1007/s41060-018-0158-2. URL `https://doi.org/10.1007/s41060-018-0158-2`.

Eric V. Strobl, Kun Zhang, and Shyam Visweswaran. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, 2018. doi: 10.1515/jci-2018-0017.

Jeffrey G Tasker and James P Herman. Mechanisms of rapid glucocorticoid feedback inhibition of the hypothalamic–pituitary–adrenal axis. *Stress*, 14(4):398–406, 2011.

Bo Thiesson, Christopher Meek, David Maxwell Chickering, and David Heckerman. Learning mixtures of dag models. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI'98, pages 504–513, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 1-55860-555-X. URL `http://dl.acm.org/citation.cfm?id=2074094.2074154`.

Sofia Triantafillou and Ioannis Tsamardinos. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *J. Mach. Learn. Res.*, 16(1):2147–2205, January 2015. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=2789272.2886819`.

Caroline Uhler, Garvesh Raskutti, Peter Bhlmann, and Bin Yu. Geometry of the faithfulness assumption in causal inference. *Ann. Statist.*, 41(2):436–463, 04 2013. doi: 10.1214/12-AOS1080. URL `https://doi.org/10.1214/12-AOS1080`.

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artif. Intell.*, 172(16-17):1873–1896, November 2008. ISSN 0004-3702. doi: 10.1016/j.artint.2008.08.001. URL `http://dx.doi.org/10.1016/j.artint.2008.08.001`.

Kun Zhang and Madelyn Glymour. Unmixing for causal inference: Thoughts on mccaffrey and danks. *The British Journal for the Philosophy of Science*, page axy040, 2018. doi: 10.1093/bjps/axy040. URL `http://dx.doi.org/10.1093/bjps/axy040`.

Kun Zhang, Biwei Huang, Jiji Zhang, Clark Glymour, and Bernhard Schölkopf. 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, 2017. ISBN 978-0-9992411-0-3. URL `http://dl.acm.org/citation.cfm?id=3171642.3171833`.

9. Appendix

9.1 FCI is Sound and Complete Under Sigma-Separation

Define σ-separation as in Definition 2.8.1 in (Forré and Mooij, 2017). We also say that σ-*separation faithfulness* holds when conditional independence implies σ-separation.

Proposition 3. *Consider a unique equilibrium distribution $\mathbb{P}$ of an SEM-IE with directed graph $\mathbb{G}$. If σ-separation faithfulness holds w.r.t. $\mathbb{G}$, then FCI is sound and complete for $\mathbb{P}$.*

Proof. Suppose there exists a unique equilibrium distribution $\mathbb{P}$ of an SEM-IE with $\mathbb{G}$. If X_i and X_j are σ-separated given $\boldsymbol{W} \subseteq \boldsymbol{X} \setminus \{X_i, X_j\}$ in $\mathbb{G}$, then we have $X_i \perp\!\!\!\perp X_j | \boldsymbol{W}$ in $\mathbb{P}$ (see Figure 3 in (Forré and Mooij, 2017) where csSEP implies gdGMP).

Now choose an arbitary collapsed graph $\mathbb{G}^1$. Note that X_i and X_j are d-separated given $\boldsymbol{W}$ in $\mathbb{G}^1$ if and only if X_i and X_j are σ-separated given $\boldsymbol{W}$ in $\mathbb{G}$ (Corollary 2.8.4 in (Forré and Mooij, 2017)). Thus, if X_i and X_j are d-separated given $\boldsymbol{W}$ in $\mathbb{G}^1$, then X_i and X_j are σ-separated given $\boldsymbol{W}$ in $\mathbb{G}$. We therefore have $X_i \perp\!\!\!\perp X_j | \boldsymbol{W}$ in $\mathbb{P}$. In other words, $\mathbb{P}$ obeys the global Markov property w.r.t. $\mathbb{G}^1$.

Suppose now that $\mathbb{P}$ obeys σ-separation faithfulness w.r.t. $\mathbb{G}$. This implies that, if we have $X_i \perp\!\!\!\perp X_j | \boldsymbol{W}$ in $\mathbb{P}$, then X_i and X_j are σ-separated given $\boldsymbol{W}$ in $\mathbb{G}$. By the equivalence of d-separation and σ-separation mentioned in the paragraph above, we conclude that X_i and X_j are d-separated given $\boldsymbol{W}$ in $\mathbb{G}^1$. In other words, $\mathbb{P}$ also obeys d-separation faithfulness w.r.t. $\mathbb{G}^1$.

From Theorem 4 in (Zhang, 2008), we know that FCI is sound and complete for any probability distribution obeying the global Markov property w.r.t. a DAG and d-separation faithfulness w.r.t. that DAG. Hence, FCI is sound and complete for any probability distribution obeying the global Markov property w.r.t. $\mathbb{G}^1$ and d-separation faithfulness w.r.t. $\mathbb{G}^1$. It follows that FCI is sound and complete for $\mathbb{P}$. □ □

9.2 Proofs

Theorem 1. *If $f(\boldsymbol{Z})$ factorizes according to Equation (6), then $f(\boldsymbol{Z})$ obeys the global Markov property with respect to $\mathcal{M}$; that is, if $\boldsymbol{A}' \perp\!\!\!\perp_d \boldsymbol{B}' | \boldsymbol{C}'$ in $\mathcal{M}$ where $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ are disjoint subsets of $\boldsymbol{Z}$, then we have $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} | \boldsymbol{C}$.*

Proof. Let $\phi(Z_i^j, \mathrm{Pa}_{\mathcal{M}}(Z_i^j))$ be a non-negative function equal to $f(Z_i | \mathrm{Pa}_{\mathbb{G}^j}(Z_i))$ whenever $f(Z_i | \mathrm{Pa}_{\mathbb{G}^j}(Z_i)) = f(Z_i | \mathrm{Pa}_{\mathbb{G}^{\boldsymbol{N}_i}}(Z_i))$ and 1 otherwise. Then $f(\boldsymbol{Z})$ factorizes according to the DAG $\mathcal{M}$ as follows:

$$f(\boldsymbol{Z}) = \prod_{i=1}^{p+s} f(Z_i | \mathrm{Pa}_{\mathbb{G}^{\boldsymbol{N}_i}}(Z_i))$$

$$= \prod_{i=1}^{p+s} \prod_{j=1}^{q} \phi(Z_i^j, \mathrm{Pa}_{\mathcal{M}}(Z_i^j))$$

$$\doteq \rho(\boldsymbol{Z}').$$

Note that $\rho(\boldsymbol{A}')$ also factorizes according to any subgraph of $\mathcal{M}$ for any set of vertices $\boldsymbol{A} \subseteq \boldsymbol{Z}$ such that $\boldsymbol{A}'$ contains all its ancestors (i.e. $\mathrm{Anc}_{\mathcal{M}}(\boldsymbol{A}') = \cup_{j=1}^{q} \boldsymbol{A}^j$).

Now denote the moral graph of $\mathcal{M}$ as $\mathcal{M}'$. Observe that $\rho(\boldsymbol{Z}')$ factorizes according to $\mathcal{M}'$:

$$\rho(\boldsymbol{Z}') = \prod_{\{Z_i^j \cup \mathrm{Pa}_{\mathcal{M}}(Z_i^j)\} \in \mathcal{D}} \phi(Z_i^j, \mathrm{Pa}_{\mathcal{M}}(Z_i^j)), \tag{7}$$

where $\mathcal{D}$ denotes the set of cliques in $\mathcal{M}'$. If $\boldsymbol{A}' \perp\!\!\!\perp_d \boldsymbol{B}' | \boldsymbol{C}'$ in $\mathcal{M}$, then $\boldsymbol{A}'$ and $\boldsymbol{B}'$ are also separated by $\boldsymbol{C}'$ in $\mathcal{M}'_{\mathrm{Anc}_{\mathcal{M}}(\boldsymbol{A}' \cup \boldsymbol{B}' \cup \boldsymbol{C}')}$, the moral graph of the ancestral set containing $\boldsymbol{A}' \cup \boldsymbol{B}' \cup \boldsymbol{C}'$ (page 72 in (Cowell et al., 1999)). Let $\mathcal{E}$ denote the set of cliques in $\mathcal{M}'_{\mathrm{Anc}_{\mathcal{M}}(\boldsymbol{A}' \cup \boldsymbol{B}' \cup \boldsymbol{C}')}$.

We construct the sets $\ddot{\boldsymbol{A}}'$ and $\ddot{\boldsymbol{B}}'$ such that $\boldsymbol{A} \subseteq \ddot{\boldsymbol{A}}$, $\boldsymbol{B} \subseteq \ddot{\boldsymbol{B}}$, $\ddot{\boldsymbol{A}} \cup \ddot{\boldsymbol{B}} \cup \boldsymbol{C} = \mathrm{Anc}_{\mathcal{M}}(\boldsymbol{A}' \cup \boldsymbol{B}' \cup \boldsymbol{C}')$ and $\ddot{\boldsymbol{A}}$, $\ddot{\boldsymbol{B}}$ and $\boldsymbol{C}$ are disjoint. We also require that $\ddot{\boldsymbol{A}}'$ and $\ddot{\boldsymbol{B}}'$ be separated by $\boldsymbol{C}'$ in $\mathcal{M}'_{\mathrm{Anc}_{\mathcal{M}}(\boldsymbol{A}' \cup \boldsymbol{B}' \cup \boldsymbol{C}')}$; such a partition of $\mathrm{Anc}_{\mathcal{M}}(\boldsymbol{A}' \cup \boldsymbol{B}' \cup \boldsymbol{C}')$ is possible because $\boldsymbol{A}'$ and $\boldsymbol{B}'$ are already separated by $\boldsymbol{C}'$ in $\mathcal{M}'_{\mathrm{Anc}_{\mathcal{M}}(\boldsymbol{A}' \cup \boldsymbol{B}' \cup \boldsymbol{C}')}$.

Let $\mathcal{E}_{\ddot{\boldsymbol{A}}}$ denote the set of cliques in $\mathcal{E}$ that have non-empty intersection with $\ddot{\boldsymbol{A}}'$. Observe then that we must have $\ddot{\boldsymbol{B}}' \cap e = \emptyset$ for all $e \in \mathcal{E}_{\ddot{\boldsymbol{A}}}$. Similarly, $\ddot{\boldsymbol{A}}' \cap e = \emptyset$ for all $e \in \mathcal{E} \setminus \mathcal{E}_{\ddot{\boldsymbol{A}}}$. We thus obtain the following similar to Equation (7):

$$
\begin{aligned}
\rho(\ddot{\boldsymbol{A}}' \cup \ddot{\boldsymbol{B}}' \cup \boldsymbol{C}') &= f(\ddot{\boldsymbol{A}} \cup \ddot{\boldsymbol{B}} \cup \boldsymbol{C}) \\
&= \prod_{\{Z_i^j \cup \mathrm{Pa}_{\mathcal{M}}(Z_i^j)\} \in \mathcal{E}} \phi(Z_i^j, \mathrm{Pa}_{\mathcal{M}}(Z_i^j)) \\
&= \prod_{e \in \mathcal{E}_{\ddot{\boldsymbol{A}}}} \gamma(e) \prod_{e \in \mathcal{E} \setminus \mathcal{E}_{\ddot{\boldsymbol{A}}}} \gamma(e) \\
&= \psi_1(\ddot{\boldsymbol{A}} \cup \boldsymbol{C}) \psi_2(\ddot{\boldsymbol{B}} \cup \boldsymbol{C}),
\end{aligned}
$$

where $\gamma(\cdot)$ denotes a non-negative function. We can obtain $f(\boldsymbol{A} \cup \boldsymbol{B} \cup \boldsymbol{C})$ by marginalizing over $f(\ddot{\boldsymbol{A}} \cup \ddot{\boldsymbol{B}} \cup \boldsymbol{C})$:

$$
\begin{aligned}
f(\boldsymbol{A} \cup \boldsymbol{B} \cup \boldsymbol{C}) &= \sum_{[\ddot{\boldsymbol{A}} \cup \ddot{\boldsymbol{B}}] \setminus [\boldsymbol{A} \cup \boldsymbol{B}]} f(\ddot{\boldsymbol{A}} \cup \ddot{\boldsymbol{B}} \cup \boldsymbol{C}) \\
&= \sum_{[\ddot{\boldsymbol{A}} \setminus \boldsymbol{A}] \cup [\ddot{\boldsymbol{B}} \setminus \boldsymbol{B}]} f(\ddot{\boldsymbol{A}} \cup \ddot{\boldsymbol{B}} \cup \boldsymbol{C}) \\
&= \sum_{[\ddot{\boldsymbol{A}} \setminus \boldsymbol{A}] \cup [\ddot{\boldsymbol{B}} \setminus \boldsymbol{B}]} \psi_1(\ddot{\boldsymbol{A}} \cup \boldsymbol{C}) \psi_2(\ddot{\boldsymbol{B}} \cup \boldsymbol{C}) \\
&= \Big[\sum_{[\ddot{\boldsymbol{B}} \setminus \boldsymbol{B}]} \Big[\sum_{[\ddot{\boldsymbol{A}} \setminus \boldsymbol{A}]} \psi_1(\ddot{\boldsymbol{A}} \cup \boldsymbol{C}) \Big] \psi_2(\ddot{\boldsymbol{B}} \cup \boldsymbol{C}) \Big] \\
&= \sum_{[\ddot{\boldsymbol{A}} \setminus \boldsymbol{A}]} \psi_1(\ddot{\boldsymbol{A}} \cup \boldsymbol{C}) \sum_{[\ddot{\boldsymbol{B}} \setminus \boldsymbol{B}]} \psi_2(\ddot{\boldsymbol{B}} \cup \boldsymbol{C}) \\
&= \psi_3(\boldsymbol{A} \cup \boldsymbol{C}) \psi_4(\boldsymbol{B} \cup \boldsymbol{C}),
\end{aligned}
$$

where the fifth equality follows because $[\ddot{\boldsymbol{A}} \setminus \boldsymbol{A}] \cap [\ddot{\boldsymbol{B}} \setminus \boldsymbol{B}] = \emptyset$ by construction. The conclusion follows by the sixth equality. $\qquad\square \qquad\qquad\qquad \square$

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

Proof. We construct a new graph $\mathcal{R}$ over $\boldsymbol{X}' \cup T'$ as follows. For each DAG in $\mathbb{G}'_{\boldsymbol{M}}$, introduce a vertex T and draw directed edges from T to all of the children of $\boldsymbol{M}$. Then remove all of the variables in $\boldsymbol{M}$ as well as all the directed edges from $\boldsymbol{M}$. Finally plot the modified DAGs in $\mathbb{G}'_{\boldsymbol{M}}$ next to each other to form the graph $\mathcal{R}$. Notice that, if we have $\boldsymbol{A}' \perp\!\!\!\perp_d \boldsymbol{B}' | \boldsymbol{C}'$ in $\mathcal{R}$, then we also have $\boldsymbol{A}' \perp\!\!\!\perp_d \boldsymbol{B}' | \boldsymbol{C}'$ in $\mathcal{M}$ by construction.

Now create another new graph $\mathcal{S}$ by replacing all of the DAGs in $\mathcal{M}$ with $\mathcal{F}$. Notice that we have $\boldsymbol{A}' \perp\!\!\!\perp_d \boldsymbol{B}' | \boldsymbol{C}'$ in $\mathcal{S}$ if and only if $\boldsymbol{A} \perp\!\!\!\perp_d \boldsymbol{B} | \boldsymbol{C}$ in $\mathcal{F}$. Moreover, $\mathcal{S}$ contains all of the edges in $\mathcal{R}$ so, if we have $\boldsymbol{A}' \perp\!\!\!\perp_d \boldsymbol{B}' | \boldsymbol{C}'$ in $\mathcal{S}$, then $\boldsymbol{A}' \perp\!\!\!\perp_d \boldsymbol{B}' | \boldsymbol{C}'$ in $\mathcal{R}$. Combining this with

the aforementioned if and only if relation, we have $\boldsymbol{A} \perp\!\!\!\perp_d \boldsymbol{B}|\boldsymbol{C}$ in $\mathcal{F}$ implies $\boldsymbol{A}' \perp\!\!\!\perp_d \boldsymbol{B}'|\boldsymbol{C}'$ in $\mathcal{S}$ which in turn implies $\boldsymbol{A}' \perp\!\!\!\perp_d \boldsymbol{B}'|\boldsymbol{C}'$ in $\mathcal{R}$. Hence, we have $\boldsymbol{A}' \perp\!\!\!\perp_d \boldsymbol{B}'|\boldsymbol{C}'$ in $\mathcal{M}$ by the end of the previous paragraph. $\qquad\qquad\square$ $\qquad\qquad\square$

Proposition 2. *If we have $O_i \notin \mathrm{Anc}_{\mathcal{F}}(O_j \cup \boldsymbol{S})$ in a father graph $\mathcal{F}$ with a corresponding mother graph $\mathcal{M}$, then we have $O_i \in \mathrm{Anc}_{\mathcal{F}_2}(O_j)$ in another father graph $\mathcal{F}_2$ such that its corresponding mother graph $\mathcal{M}_2$ lies within the same Markov equivalence class over $\boldsymbol{O}$ as $\mathcal{M}$.*

Proof. Set $\mathbb{G}'_2$ to $\mathbb{G}'_{\boldsymbol{M}}$, where $\mathbb{G}'_2$ refers to the set of DAGs in $\mathcal{M}_2$. If $|\mathbb{G}'_{\boldsymbol{M}}| = 1$, then add two copies of the DAG into $\mathbb{G}'_2$. Now add one new latent variable L_m into each DAG in $\mathbb{G}'_2$ as follows. For all but the last DAG, draw the two directed edge $O_i \to L_m$. For the last DAG, draw $L_m \to O_j$. Next, introduce a new latent common cause M_{s+1} for L_m and O_j into every DAG. Observe that all of the new paths are d-separating among the observed variables in $\mathcal{G}_2$, so no new d-separation or d-connection relations were introduced among the observed variables in $\mathcal{M}_2$. However, we have $O_i \in \mathrm{Anc}_{\mathcal{F}_2}(O_j)$ with the directed path $O_i \to L_m \to O_j$. $\qquad\square$ $\qquad\square$

The above result implies that it is impossible to infer non-ancestral relations with a CI oracle, unless we make further assumptions. Traditional methods skirt the above result by restricting the number of DAGs in $\mathbb{G}'_{\boldsymbol{M}}$ and $\mathbb{G}'_2$ both to one. The result also elucidates a warning for the suspiciously high number of non-ancestral relations inferred by past constraint-based algorithms in general; Proposition 2 implies that the arrowheads should not have been inferred in the first place under a mixture of DAGs framework.

While we cannot infer non-ancestral relations in general, we can infer ancestral ones using a CI oracle:

Lemma 1. *Suppose that we have $O'_i \perp\!\!\!\perp_d O'_j|\boldsymbol{W}' \cup \boldsymbol{S}'$ in $\mathcal{M}$ but we have $O'_i \not\perp\!\!\!\perp_d O'_j|\boldsymbol{V}' \cup \boldsymbol{S}'$ for every $\boldsymbol{V} \subset \boldsymbol{W}$. If $O_k \in \boldsymbol{W}$, then $O_k \in \mathrm{Anc}_{\mathcal{F}}(\{O_i, O_j\} \cup \boldsymbol{S})$.*

Proof. We first invoke Lemma 15 in (Strobl, 2018) by setting $\boldsymbol{R} = \emptyset$, $O_i = O'_i$, $O_j = O'_j$, $\boldsymbol{W} = \boldsymbol{W}'$ and $\boldsymbol{S} = \boldsymbol{S}'$ in that paper. We can then conclude that $O'_k \in \mathrm{Anc}_{\mathcal{M}}(O'_i \cup O'_j \cup \boldsymbol{S}')$. If $O'_k \in \mathrm{Anc}_{\mathcal{M}}(O'_i \cup O'_j \cup \boldsymbol{S}')$, then there exists at least one DAG $\mathbb{G} \in \mathbb{G}'_{\boldsymbol{M}}$ such that $O_k \in \mathrm{Anc}_{\mathbb{G}}(\{O_i, O_j\} \cup \boldsymbol{S})$ because no paths exist between the DAGs in $\mathbb{G}'_{\boldsymbol{M}}$. The conclusion follows because every edge in $\mathbb{G}$ is also in $\mathcal{F}$. $\qquad\square$ $\qquad\square$

9.3 Skeleton Discovery

We have summarized the skeleton discovery procedure of CIM in Algorithm 2. Algorithm 2 learns the skeleton as follows. First, the algorithm initializes a fully connected nondirected graph $\widehat{\mathcal{F}}^*$ in Step 1. Algorithm 2 then determines whether ${}^aO_i$ and ${}^bO_j$ are conditionally dependent given all subsets of $\mathrm{Adj}_{\widehat{\mathcal{F}}^*}({}^aO_i) \setminus {}^bO_j$ and all subsets of $\mathrm{Adj}_{\widehat{\mathcal{F}}^*}({}^bO_j) \setminus {}^aO_i$ in Step 10. If the algorithm finds a CI relation, then Algorithm 2 removes the edge between ${}^aO_i$ and ${}^bO_j$ in $\widehat{\mathcal{F}}^*$ and records the separating set in Sep. The algorithm ultimately outputs both $\widehat{\mathcal{F}}^*$ and Sep for later use by Algorithm 1.

9.4 Extra Experimental Results

We have summarized the results for the Mayo Clinic Primary Biliary Cirrhosis and Cognition and Aging USA datasets in Figures 12 and 13, respectively.

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

1 Form a complete graph $\widehat{\mathcal{F}}^*$ over $\boldsymbol{O}$ with edges $\circ\!\!-\!\!\circ$

2 $l \leftarrow -1$

3 repeat

4 $\quad$ Let $l = l + 1$

5 $\quad$ For each vertex $^aO_i \in \boldsymbol{O}$, form the adjacency set $\mathrm{Adj}_{\widehat{\mathcal{F}}^*}(^aO_i)$

6 $\quad$ **repeat**

7 $\qquad$ Select a new ordered pair of vertices $(^aO_i, {}^bO_j)$ that are adjacent in $\widehat{\mathcal{F}}^*$ and satisfy $|\mathrm{Adj}_{\widehat{\mathcal{F}}^*}(^aO_i) \setminus {}^bO_j| \geq l$

8 $\qquad$ **repeat**

9 $\qquad\quad$ Choose a new set $\boldsymbol{W} \subseteq \mathrm{Adj}_{\widehat{\mathcal{F}}^*}(^aO_i) \setminus {}^bO_j$ with $|\boldsymbol{W}| = l$

10 $\qquad\quad$ **if** $^aO_i \perp\!\!\!\perp {}^bO_j | \boldsymbol{W} \cup \boldsymbol{S}$ **then**

11 $\qquad\qquad$ Delete the edge $^aO_i \circ\!\!-\!\!\circ {}^bO_j$ from $\widehat{\mathcal{F}}^*$

12 $\qquad\qquad$ Let $\mathrm{Sep}(^aO_i, {}^bO_j) = \mathrm{Sep}(^bO_j, {}^aO_i) = \boldsymbol{W}$

13 $\qquad\quad$ **end**

14 $\qquad$ **until** *^aO_i and ^bO_j are no longer adjacent in $\widehat{\mathcal{F}}^*$ or all $\boldsymbol{W} \subseteq \mathrm{Adj}_{\widehat{\mathcal{F}}^*}(^aO_i) \setminus {}^bO_j$ with $|\boldsymbol{W}| = l$ have been considered;*

15 $\quad$ **until** *all ordered pairs of adjacent vertices $(^aO_i, {}^bO_j)$ in $\widehat{\mathcal{F}}^*$ with $|\mathrm{Adj}_{\widehat{\mathcal{F}}^*}(^aO_i) \setminus {}^bO_j| \geq l$ have been considered;*

16 until *all pairs of adjacent vertices $(^aO_i, {}^bO_j)$ in $\widehat{\mathcal{F}}^*$ satisfy $|\mathrm{Adj}_{\widehat{\mathcal{F}}^*}(^aO_i) \setminus {}^bO_j| \leq l$;*

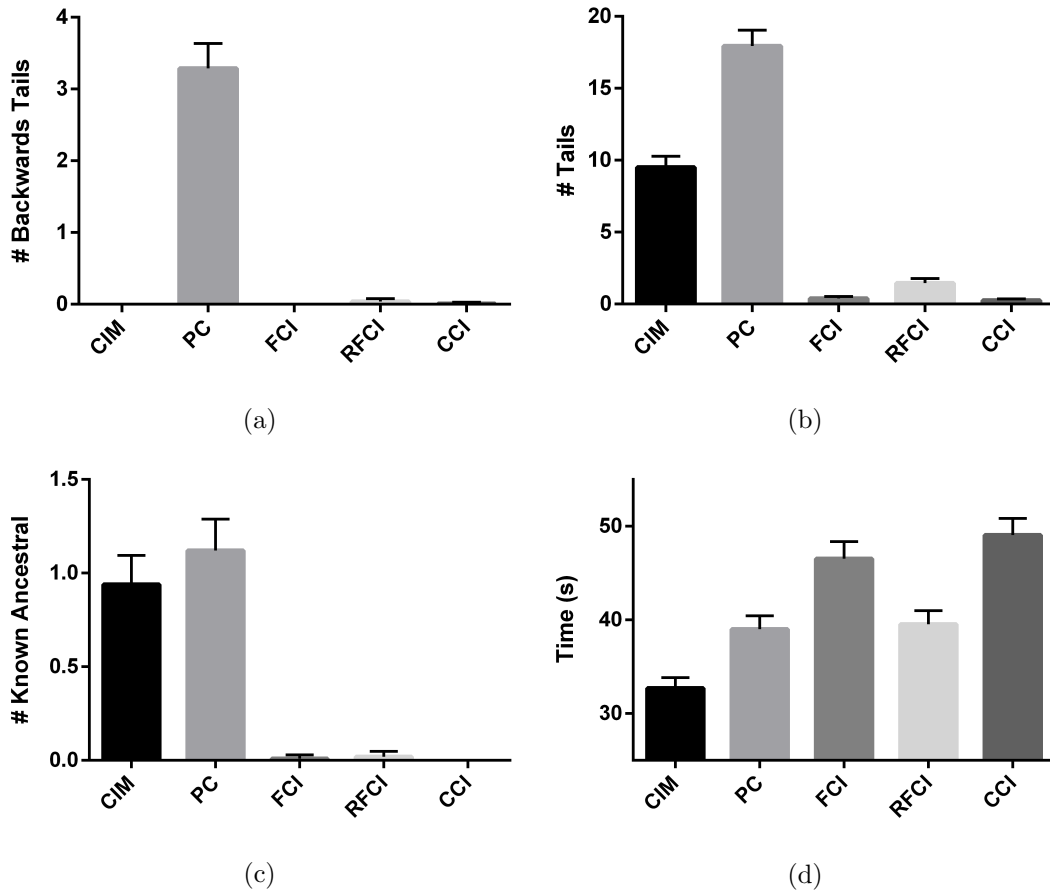**Algorithm 2:** CIM's skeleton discovery procedure

(a)

(b)

(c)

(d)

Figure 12: Mayo Clinic Primary Biliary Cirrhosis study results presented in the same format as Figure 8.
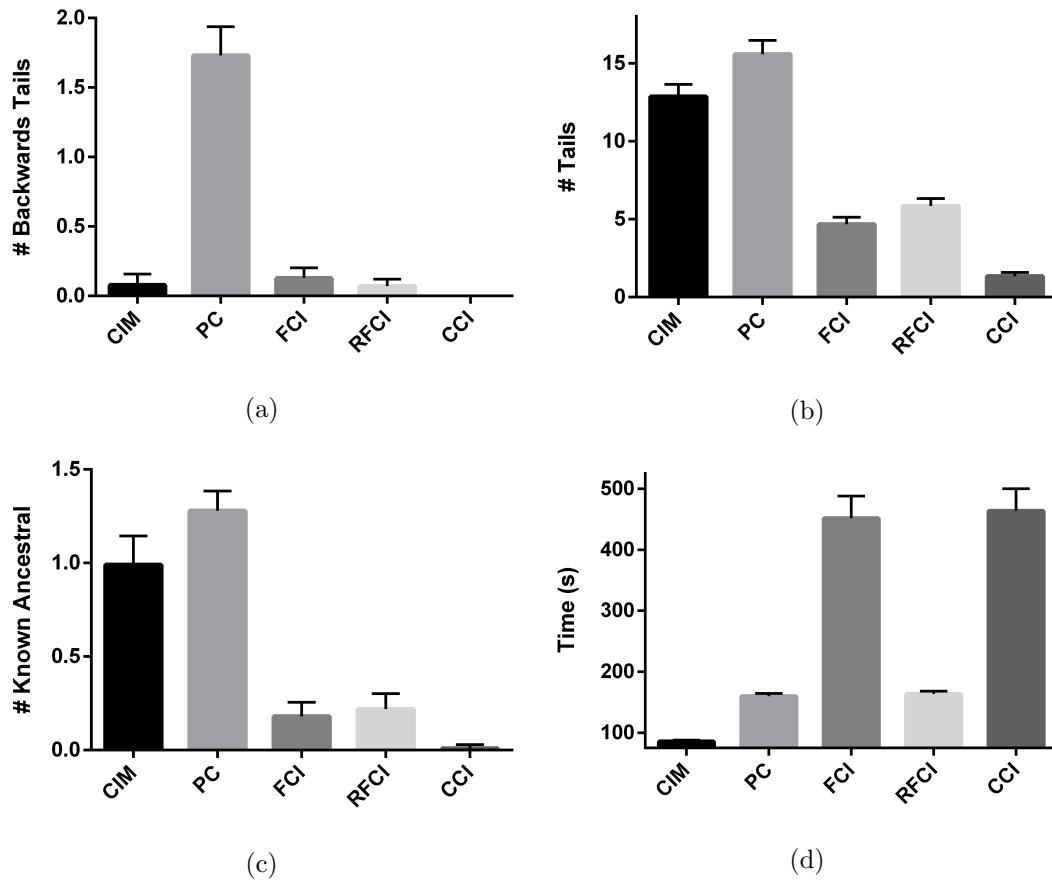
(a)

(b)

(c)

(d)

Figure 13: Cognition and Aging USA study results presented in the same format as Figure 8.