Causal Learning for Partially Observed Stochastic Dynamical Systems

Søren Wengel Mogensen
Department of Mathematical Sciences
University of Copenhagen
Copenhagen, Denmark

Daniel Malinsky
Department of Computer Science
Johns Hopkins University
Baltimore, MD, USA

Niels Richard Hansen
Department of Mathematical Sciences
University of Copenhagen
Copenhagen, Denmark

Abstract

Many models of dynamical systems have causal interpretations that support reasoning about the consequences of interventions, suitably defined. Furthermore, local independence has been suggested as a useful independence concept for stochastic dynamical systems. There is, however, no well-developed theoretical framework for causal learning based on this notion of independence. We study independence models induced by directed graphs (DGs) and provide abstract graphoid properties that guarantee that an independence model has the global Markov property w.r.t. a DG. We apply these results to Itô diffusions and event processes. For a partially observed system, directed mixed graphs (DMGs) represent the marginalized local independence model, and we develop, under a faithfulness assumption, a sound and complete learning algorithm of the directed mixed equivalence graph (DMEG) as a summary of all Markov equivalent DMGs.

1 INTRODUCTION

Causal learning has been developed extensively using structural causal models and graphical representations of the conditional independence relations that they induce. The Fast Causal Inference (FCI) algorithm and its variations (RFCI, FCI+, ...) can learn a representation of the independence relations induced by a causal model even when the causal system is only partially observed, i.e., the data is "causally insufficient" in the terminology of Spirtes et al. (2000). FCI is, however, not directly applicable for learning causal relations among entire processes in a continuous-time dynamical system. The dynamic evolution of such a system cannot be modeled using a finite number of variables related via a structural causal model, and standard probabilistic independence cannot adequately capture infinitesimal conditional independence relationships between processes since such relationships can be asymmetric. The asymmetry can intuitively be explained by the fact that the present of one process may be independent of the past of another process, or the reverse, or both.

Local independence was introduced by Schweder (1970) and is a formalization of how the present of one stochastic process depends on the past of others in a dynamical system. This concept directly lends itself to a causal interpretation as dynamical systems develop as functions of their pasts, see e.g. Aalen (1987). Didelez (2000, 2006a, 2008) considered graphical representations of local independence models using directed graphs (DGs) and δ-separation and proved the equivalence of the pairwise and global Markov properties in the case of multivariate counting processes. Nodelman et al. (2002, 2003) and Gunawardana et al. (2011) also considered learning problems in continuous-time models. In this paper, we extend the theory to a broader class of semimartingales, showing the equivalence of pairwise and global Markov properties in DGs. To represent marginalized local independence models, Mogensen and Hansen (2018) introduced directed mixed graphs (DMGs) with μ-separation. Bidirected edges in DMGs (roughly) correspond to dependencies induced by latent processes, and in this sense DMGs can represent partially observed dynamical systems. In contrast to the "causally sufficient" setting as represented by a DG, multiple DMGs may represent the same set of (marginal) local independence relations; thus we use the characterization of Markov equivalent DMGs by Mogensen and Hansen (2018) to propose a sound and complete algorithm for selecting a set of DMGs consistent with a given collection of independence relations.

Proofs omitted from the main text can be found in the supplementary material.

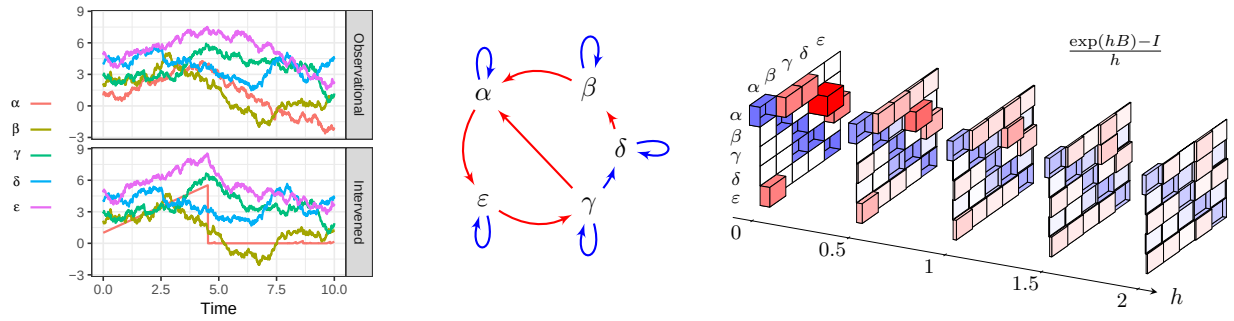

Figure 1: Simulated sample paths (left) for the linear SDE determined by B in (1). The sample paths are from the observational distribution started in the stationary mean as well as under an intervention regime on α. For the local independence graph (middle) the color of the edge $j \to i$ indicates if the nonzero entry B_{ij} is positive (red) or negative (blue). The step size h difference quotient at 0 for the semigroup $t \mapsto \exp(tB)$ (right) determines the discrete time conditional means for time step h transitions. It does not directly reflect the local independences except in the limit $h \to 0$, where it converges to the infinitesimal generator B. Danks and Plis (2013) make a similar point in the case of subsampled time series.

2 CAUSAL DYNAMICAL MODELS

The notion of interventions in a continuous-time model of a dynamical system is not new, and has been investigated thoroughly in the context of control theory. Causal models and interventions for event processes and their relation to graphical independence models have been treated in detail (Didelez, 2008, 2015). Relations to structural causal models have been established for ordinary differential equations (ODEs) (Mooij et al., 2013; Rubenstein et al., 2016). Notions of causality and interventions have also been treated for general stochastic processes such as stochastic differential equations (SDEs) (Aalen et al., 2012; Commenges and Gégout-Petit, 2009; Sokol and Hansen, 2014).

To motivate and explain the general results of this paper, we introduce the toy linear SDE model in $\mathbb{R}^5$ given by $\mathrm{d}X_t = B(X_t - A)\mathrm{d}t + \mathrm{d}W_t$ with $A = (1, 2, 3, 4, 5)^T$,

$$B = \begin{pmatrix} -1.1 & 1 & 1 & \cdot & \cdot \\ \cdot & -1.1 & \cdot & 2.0 & \cdot \\ \cdot & \cdot & -1.1 & \cdot & 1 \\ \cdot & \cdot & -1 & -1.1 & \cdot \\ 1 & \cdot & \cdot & \cdot & -1.1 \end{pmatrix}, \quad (1)$$

and (W_t) a five-dimensional standard Brownian motion. The coordinates of this process will be denoted α, β, γ, δ, and ϵ. If we assume that this SDE has a causal interpretation, we can obtain predictions under interventions via manipulations of the SDE itself, see e.g. Sokol and Hansen (2014). In Figure 1, for instance, we replace the α coordinate of the SDE by

$$\mathrm{d}X^\alpha_t = 1(X^\beta_t > 1)\mathrm{d}t, \quad X^\alpha_t - X^\alpha_{t-} = -X^\alpha_{t-} 1(X^\beta_t \leq 1).$$

The nonzero pattern of the B matrix defines a directed graph which we identify as the *local independence graph* below, which in turn is related to the local independence model of the SDE. It is a main result of this paper that the local independence model satisfies the global Markov property w.r.t. this graph. Under a faithfulness assumption we can identify (aspects of) the causal system from observational data even when some processes are unobserved.

It is well known that

$$X_{t+h} - X_t \mid X_t \sim \mathcal{N}((e^{hB} - I)(X_t - A), \Sigma(h))$$

with $\Sigma(h)$ given in terms of B. Thus a sample of the process at equidistant time points is a vector autoregressive process with correlated errors. We note that $e^{hB} - I$ is a dense matrix that will not reveal the local independence graph unless h is sufficiently small, see Figure 1. The matrix B is, furthermore, a stable matrix, hence there is a stationary solution to the SDE and for $h \to \infty$ we have $\Sigma(h) \to \Sigma$, the invariant covariance matrix. We note that Σ^{-1} is also a dense matrix, thus the invariant distribution does not satisfy the global Markov property w.r.t. to any undirected graph but the complete graph.

In conclusion, the local independence model of the SDE is not encoded directly neither by Markov properties of discrete time samples, nor by Markov properties of the invariant distribution. This is the motivation for our abstract development of local independence models, their relation to continuous-time stochastic processes, and a dedicated learning algorithm.

3 INDEPENDENCE MODELS

Consider some finite set V. An *independence model* over V is a set of triples $\langle A, B \mid C\rangle$ such that $A, B, C \subseteq V$. We let $\mathcal{I}$ denote a generic independence model. Following Didelez (2000, 2008) we will consider independence models that are not assumed to be symmetric in A and B. The independence models we consider do however satisfy other properties which allow us to deduce some independences from others. We define the following properties, some of which have previously been described as *asymmetric (semi)graphoid properties* (Didelez, 2006b, 2008). Many of them are analogous to properties in the literature on conditional independence models (Lauritzen, 1996), though due to the lack of symmetry, one may define both left and right versions.

- Left redundancy: $\langle A, B \mid A\rangle \in \mathcal{I}$
- Left decomposition: $\langle A, B \mid C\rangle \in \mathcal{I}, D \subseteq A \Rightarrow \langle D, B \mid C\rangle \in \mathcal{I}$
- Right decomposition: $\langle A, B \mid C\rangle \in \mathcal{I}, D \subseteq B \Rightarrow \langle A, D \mid C\rangle \in \mathcal{I}$
- Left weak union: $\langle A, B \mid C\rangle \in \mathcal{I}, D \subseteq A \Rightarrow \langle A, B \mid C \cup D\rangle \in \mathcal{I}$
- Right weak union: $\langle A, B \mid C\rangle \in \mathcal{I}, D \subseteq B \Rightarrow \langle A, B \mid C \cup D\rangle \in \mathcal{I}$
- Left intersection: $\langle A, B \mid C\rangle \in \mathcal{I}, \langle C, B \mid A\rangle \in \mathcal{I} \Rightarrow \langle A \cup C, B \mid A \cap C\rangle \in \mathcal{I}$
- Left composition: $\langle A, B \mid C\rangle \in \mathcal{I}, \langle D, B \mid C\rangle \in \mathcal{I} \Rightarrow \langle A \cup D, B \mid C\rangle \in \mathcal{I}$
- Right composition: $\langle A, B \mid C\rangle \in \mathcal{I}, \langle A, D \mid C\rangle \in \mathcal{I} \Rightarrow \langle A, B \cup D \mid C\rangle \in \mathcal{I}$
- Left weak composition: $\langle A, B \mid C\rangle \in \mathcal{I}, D \subseteq C \Rightarrow \langle A \cup D, B \mid C\rangle \in \mathcal{I}$

For disjoint sets $A, C, D \subseteq V$, we say that A and D *factorize* w.r.t. C if there exists a partition $C = C_1 \dot\cup C_2$ such that (i) and (ii) hold:

(i) $\langle A, C_1 \cup D \mid C \cup D\rangle \in \mathcal{I}$
(ii) $\langle D, C_2 \cup A \mid C \cup A\rangle \in \mathcal{I}$.

Definition 1. The independence model $\mathcal{I}$ satisfies *cancellation* if $\langle A, B \mid C \cup \{\delta\}\rangle \in \mathcal{I}$ implies $\langle A, B \mid C\rangle \in \mathcal{I}$ whenever A and $\{\delta\}$ factorize w.r.t. C. Such an independence model is called *cancellative*.

Cancellation is related to ordered downward-stability as defined by Sadeghi (2017) for symmetric independence models over a set with a preorder and studied in relation to separation in acyclic graphs.

3.1 DIRECTED MIXED GRAPHS

We wish to relate a local independence model, as defined in Section 4, to a graph and therefore we need a notion of graphical separation which allows for asymmetry. *Directed mixed graphs* along with μ-separation will provide the means for such graphical modeling of local independence. The subsequent definitions follow Mogensen and Hansen (2018), which we refer to for further details.

Definition 2 (Directed mixed graph)**.** A *directed mixed graph* (DMG) is an ordered pair (V, E) where V is a finite set of vertices (also called nodes) and E is a finite set of edges of the types $\to$ and $\leftrightarrow$. A pair of vertices $\alpha, \beta \in V$ may be joined by any subset of $\{\alpha \to \beta, \alpha \leftarrow \beta, \alpha \leftrightarrow \beta\}$. Note that we allow for loops, i.e., edges $\alpha \to \alpha$ and/or $\alpha \leftrightarrow \alpha$.

Let $\mathcal{G}_1 = (V, E_1)$ and $\mathcal{G}_2 = (V, E_2)$ be DMGs. If $E_1 \subseteq E_2$, then we write $\mathcal{G}_1 \subseteq \mathcal{G}_2$ and say that $\mathcal{G}_2$ is a *supergraph* of $\mathcal{G}_1$. The *complete* DMG on V is the DMG which is a supergraph of all other DMGs with vertices V. Throughout this paper, $\mathcal{G}$ will denote a DMG with node set V and edge set E. We will also consider *directed graphs* (DGs) which are DMGs with no bidirected edges. Let $\alpha, \beta \in V$. We will say that the edge $\alpha \to \beta$ has a *head* at β and a *tail* at α, and that the edge $\alpha \leftrightarrow \beta$ has heads at both α and β. When we write e.g. $\alpha \to \beta$ this does not preclude other edges between these nodes. We use $\alpha \ast\!\!\to \beta$ to denote any edge between α and β that has a head at β. A letter over an edge, e.g. $\alpha \overset{e}{\to} \beta$, denotes simply that e refers to that specific edge. If the edge $\alpha \to \beta$ is in the graph then we say that α is a *parent* of β and if $\alpha \leftrightarrow \beta$ then we say that α and β are *siblings*. Let $\mathrm{pa}(\alpha)$ (or $\mathrm{pa}_{\mathcal{G}}(\alpha)$ to make the graph explicit) denote the set of parents of α in $\mathcal{G}$. Note that due to loops, α can be both a parent and a sibling of itself.

A *walk* is an alternating, ordered sequence of nodes and edges along with an orientation of the edge such that each edge is between its two adjacent nodes, $\langle \nu_1, e_1, \nu_2, \ldots, e_n, \nu_{n+1}\rangle$, where $\nu_i \in V$ and $e_j \in E$. We say that the walk is between ν_1 and ν_{n+1} or from ν_1 to ν_{n+1}. The ν_1 and ν_{n+1} are called the *endpoint nodes* of the walk. A non-endpoint node $\nu_i, i \neq 1, n+1$, is called a collider if the two adjacent edges on the walk both have heads at the node, and otherwise a non-collider. Note that the endpoint nodes are neither colliders nor non-colliders. A walk is called *trivial* if it consists of a single node and no edges. A *path* is a walk where no node is repeated. A path from α to β is *directed* if every edge on the path is directed and points towards β. We say that α is an ancestor of a set $C \subseteq V$ if there exists a (possibly trivial) directed path from α to $\gamma \in C$. We let $\mathrm{an}(C)$ denote the set of nodes that are ancestors to C. Note that $C \subseteq \mathrm{an}(C)$.

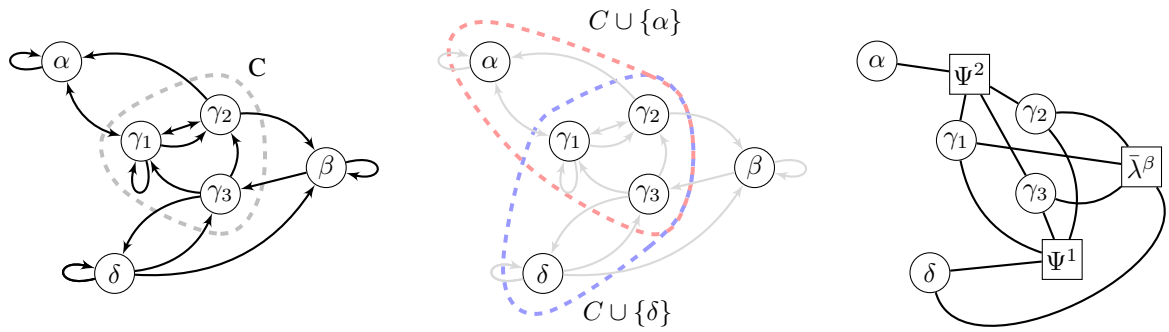

Figure 2: A DMG $\mathcal{G}$ (left) with sets $\{\alpha\}$ and $\{\delta\}$ that factorize w.r.t. $C = \{\gamma_1, \gamma_2, \gamma_3\}$ such that $\alpha \perp_\mu \beta \mid C \cup \{\delta\}$. Any node is μ-separated from either α by $C \cup \{\delta\}$ or δ by $C \cup \{\alpha\}$ (middle), and as $\mathcal{I}(\mathcal{G})$ is cancellative, $\alpha \perp_\mu \beta \mid C$. A corresponding factor graph (right) with the three factor nodes Ψ^1, Ψ^2 and $\bar{\lambda}^\beta$, cf. Theorem 14.

3.1.1 μ-separation

Definition 3 (μ-connecting walk)**.** A μ-connecting walk from α to β given C is a non-trivial walk from α to β such that $\alpha \notin C$, every non-collider is not in C and every collider is in $\mathrm{an}(C)$, and such that the final edge has a head at β.

Definition 4. Let $\alpha, \beta \in V, C \subseteq V$. We say that β is μ-separated from α given C in the graph $\mathcal{G}$ if there is no μ-connecting walk from α to β in $\mathcal{G}$ given C. For general sets, $A, B, C \subseteq V$, we say that B is μ-separated from A given C and write $A \perp_\mu B \mid C$ if β is μ-separated from α given C for every $\alpha \in A$ and $\beta \in B$. We write $A \perp_\mu B \mid C\ [\mathcal{G}]$ if we wish to make explicit to which graph the statement applies.

Note that this definition means that B is separated from A given C whenever $A \subseteq C$. We associate an independence model $\mathcal{I}(\mathcal{G})$ with a DMG $\mathcal{G}$ by

$$\langle A, B \mid C \rangle \in \mathcal{I}(\mathcal{G}) \Leftrightarrow A \perp_\mu B \mid C\ [\mathcal{G}].$$

Lemma 5. The independence model $\mathcal{I}(\mathcal{G})$ satisfies left and right {decomposition, weak union, composition} and left {redundancy, intersection, weak composition}. Furthermore, $\langle A, B \mid C \rangle \in \mathcal{I}(\mathcal{G})$ whenever $B = \emptyset$.

Lemma 6. $\mathcal{I}(\mathcal{G})$ satisfies cancellation.

3.1.2 Markov equivalence

We say that DMGs $\mathcal{G}_1 = (V, E_1)$, $\mathcal{G}_2 = (V, E_2)$ are *Markov equivalent* if $\mathcal{I}(\mathcal{G}_1) = \mathcal{I}(\mathcal{G}_2)$ and this defines an equivalence relation. We let $[\mathcal{G}]$ denote the (Markov) equivalence class of $\mathcal{G}$. For DMGs, it does not hold that Markov equivalent graphs have the same adjacencies. Note that the same is true for the directed (cyclic) graphs with no loops considered by Richardson (1996, 1997) in another context. We say that a DMG is *maximal* if it is complete or if no edge can be added without changing the associated Markov equivalence class. Mogensen and Hansen (2018) define for every vertex in a DMG a set of *potential parents* and *potential siblings* (both subsets of V) using the independence model induced by the graph (these definitions are also included in the supplementary material). We let $\mathrm{pp}(\alpha, \mathcal{I})$ denote the set of potential parents of α and $\mathrm{ps}(\alpha, \mathcal{I})$ denote the set of potential siblings of α in the independence model $\mathcal{I}$. If $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent we thus have $\mathrm{pp}(\alpha, \mathcal{I}(\mathcal{G}_1)) = \mathrm{pp}(\alpha, \mathcal{I}(\mathcal{G}_2))$ and $\mathrm{ps}(\alpha, \mathcal{I}(\mathcal{G}_1)) = \mathrm{ps}(\alpha, \mathcal{I}(\mathcal{G}_2))$ for each $\alpha \in V$. Given a DMG $\mathcal{G}$ and independence model $\mathcal{I} = \mathcal{I}(\mathcal{G})$, one can construct another DMG $\mathcal{N}$ in which α is a parent of β if and only if $\alpha \in \mathrm{pp}(\beta, \mathcal{I})$ and α and β are siblings if and only if $\alpha \in \mathrm{ps}(\beta, \mathcal{I})$. Mogensen and Hansen (2018) showed that $\mathcal{N} \in [\mathcal{G}]$, that it is a supergraph of all elements of $[\mathcal{G}]$, and that $\mathcal{N}$ is maximal. This allows one to define a *directed mixed equivalence graph* (DMEG) from the (unique) maximal graph $\mathcal{N}$ in the equivalence class to summarize the entire equivalence class. The DMEG is constructed from $\mathcal{N}$ by partitioning the edge set into two subsets: one consisting of the edges which are common to all graphs in the Markov equivalence class, and one consisting of edges that are present in some members of the equivalence class but absent in others. One may visualize the DMEG by drawing $\mathcal{N}$ and making the edges in the latter set dashed. Note that by collapsing the distinction between dashed and solid edges one may straightforwardly apply μ-separation to a given DMEG.

3.2 MARKOV PROPERTIES

The main result of this section gives conditions on an abstract independence model ensuring equivalence be-

tween the *pairwise* and the *global Markov properties* w.r.t. a directed graph with μ-separation. In the next section we give examples of classes of processes that fulfill these conditions, extending results in Didelez (2008) to a broader class of models. We take an axiomatic approach to proving the equivalence in the sense that we describe some abstract properties and use only these to show the equivalence. This is analogous to what Lauritzen and Sadeghi (2017) did in the case of symmetric independence models.

Definition 7. A DG and an independence model satisfy the pairwise Markov property if for $\alpha, \beta \in V$,

$$\alpha \notin \mathrm{pa}(\beta) \Rightarrow \langle \alpha, \beta \mid V \setminus \{\alpha\} \rangle \in \mathcal{I}$$

A DMG and an independence model satisfy the global Markov property if for $A, B, C \subseteq V$,

$$A \perp_\mu B \mid C \Rightarrow \langle A, B \mid C \rangle \in \mathcal{I}.$$

Theorem 8. Assume that $\mathcal{I}$ is an independence model that satisfies left {redundancy, intersection, decomposition, weak union, weak composition}, right {decomposition, composition}, is cancellative, and furthermore $\langle A, B \mid C \rangle \in \mathcal{I}$ whenever $B = \emptyset$. Let $\mathcal{D}$ be a DG. Then $\mathcal{I}$ satisfies the pairwise Markov property with respect to $\mathcal{D}$ if and only if it satisfies the global Markov property with respect to $\mathcal{D}$.

To keep consistency with earlier literature, we define the pairwise Markov condition above as the absence of an edge, which does not directly generalize to DMGs. Therefore, we prove the equivalence of pairwise and global Markov only in the class of DGs. The main purpose of DMGs is to represent Markov properties from marginalized DGs as defined below, in which case the global Markov property w.r.t. a DMG is inherited from the DG.

Definition 9 (Marginal independence model)**.** Assume that $\mathcal{I}$ is an independence model over V. Then the marginal independence model of $\mathcal{I}$ over $O \subseteq V$, $\mathcal{I}^O$, is the independence model,

$$\mathcal{I}^O = \{\langle A, B \mid C \rangle \mid \langle A, B \mid C \rangle \in \mathcal{I}; A, B, C \subseteq O\}.$$

Mogensen and Hansen (2018) give a marginalization algorithm (a.k.a. a "latent projection"), which outputs a marginal DMG, $\mathcal{G} = (O, F)$, from a DG, $\mathcal{D} = (V, E)$, such that $\mathcal{I}(\mathcal{D})^O = \mathcal{I}(\mathcal{G})$. If $\mathcal{I}$ satisfies the global Markov property w.r.t. $\mathcal{D}$ then

$$\mathcal{I}(\mathcal{G}) = \mathcal{I}(\mathcal{D})^O \subseteq \mathcal{I}^O.$$

This shows that the marginalized independence model $\mathcal{I}^O$ then satisfies the global Markov property w.r.t. the DMG $\mathcal{G}$.

4 LOCAL INDEPENDENCE

This section introduces local independence models and local independence graphs. The main results of the section provide verifiable conditions that ensure that a local independence model satisfies the global Markov property w.r.t. the local independence graph.

Let $X = (X_t^1, \ldots, X_t^n)$ for $t \in [0, T]$ be a càdlàg stochastic process defined on the probability space $(\Omega, \mathcal{F}, P)$. Introduce for $A \subseteq V = \{1, \ldots, n\}$ the filtration $\mathcal{F}_t^A$ as the completed and right continuous version of $\sigma(\{X_s^\alpha, s \leq t, \alpha \in A\})$. Let also $\lambda = (\lambda_t^1, \ldots, \lambda_t^n)$ be an integrable càdlàg stochastic process. This λ-process need not have any specific relation to X *a priori*, but for the main Theorem 14 the relation is through the compatibility processes defined below. Note that some computations below technically require that $E(\cdot \mid \mathcal{F}_t)$ is computed as the optional projection, cf. Theorem VI.7.1 and Lemma VI.7.8 in Rogers and Williams (2000). This is unproblematic, and will not be discussed any further.

Definition 10. We say that B is λ-locally independent of A given C if the process

$$t \mapsto E(\lambda_t^\beta \mid \mathcal{F}_t^{A \cup C})$$

has an $\mathcal{F}_t^C$-adapted version for all $\beta \in B$. In this case we write $A \not\to_\lambda B \mid C$.

This is slightly different from the definition in Didelez (2008) in that β is not necessarily in the conditioning set. This change in the definition makes it possible for a process to be locally independent from itself given some separating set. We define the local independence model, $\mathcal{I}(X, \lambda)$, determined by X and λ via

$$\langle A, B \mid C \rangle \in \mathcal{I}(X, \lambda) \Leftrightarrow A \not\to_\lambda B \mid C.$$

When there is no risk of ambiguity we say that B is locally independent of A given C, and we write $A \not\to B \mid C$ and $\mathcal{I} = \mathcal{I}(X, \lambda)$.

The local independence model satisfies a number of the properties listed in Section 3.

Lemma 11. Let $\mathcal{I}$ be a local independence model. Then it satisfies left {redundancy, decomposition, weak union, weak composition} and right {decomposition, composition} and furthermore $\langle A, B \mid C \rangle \in \mathcal{I}$ whenever $B = \emptyset$. If $\mathcal{F}_t^A \cap \mathcal{F}_t^C = \mathcal{F}_t^{A \cap C}$ holds for all $A, C \subseteq V$ and $t \in [0, T]$, then left intersection holds.

Definition 12. The local independence graph is the directed graph with node set $V = \{1, \ldots, n\}$ such that

$$\alpha \notin \mathrm{pa}(\beta) \Leftrightarrow \alpha \not\to_\lambda \beta \mid V \setminus \{\alpha\}.$$

By Theorem 8 and Lemma 11 a local independence model that satisfies left intersection and is cancellative satisfies the global Markov property w.r.t. the local independence graph. Left intersection holds by Lemma 11 whenever $\mathcal{F}_t^A \cap \mathcal{F}_t^C = \mathcal{F}_t^{A\cap C}$. Theorem 14 below gives a general factorization condition on the distribution of the stochastic processes that ensures a local independence model to be cancellative. This condition is satisfied for example by event and Itô processes.

Introduce for $C \subseteq V$ and $\beta \in V$ the shorthand notation

$$\lambda_t^{C,\beta} = E(\lambda_t^\beta \mid \mathcal{F}_t^C).$$

Furthermore, for $\alpha \in A \subseteq V$ let

$$\Psi_t^{A,\alpha} = \psi_t^\alpha((\lambda_s^{A,\alpha})_{s\le t}, (X_s^\alpha)_{s\le t})$$

denote a càdlàg process that is given in terms of a positive functional ψ_t^α of the history of the $\lambda^{A,\alpha}$- and the X^α-processes up to time t.

Definition 13. We say that P λ-factorizes with compatibility processes $\Psi^{A,\alpha} > 0$ if for all $A \subseteq V$

$$P = \frac{1}{Z_t^A} \prod_{\alpha\in A} \Psi_t^{A,\alpha} \cdot Q_t^A$$

with Q_t^A a probability measure on $(\Omega, \mathcal{F})$ such that $(X_s^\alpha)_{0\le s\le t}$ for $\alpha \in A$ are independent under Q_t^A. Here, Z_t^A is a deterministic normalization constant.

Theorem 14. The local independence model $\mathcal{I}(X, \lambda)$ is cancellative if P λ-factorizes.

Proof. Assume that $A, \{\delta\} \subseteq V$ factorize w.r.t. $C = C_1 \dot\cup C_2$. In this proof, (i) and (ii) refer to the factorization properties, see Definition 1. Let $F = C \cup A \cup \{\delta\}$. Then by (i)

$$\Psi_t^{F,\gamma} = \psi_t^\gamma((\lambda_s^{C\cup\{\delta\},\gamma})_{s\le t}, (X_s^\gamma)_{s\le t}) = \Psi_t^{C\cup\{\delta\},\gamma}$$

for $\gamma \in C_1 \cup \{\delta\}$, and by (ii)

$$\Psi_t^{F,\gamma} = \psi_t^\gamma((\lambda_s^{C\cup A,\gamma})_{s\le t},, (X_s^\gamma)_{s\le t}) = \Psi_t^{C\cup A,\gamma}$$

for $\gamma \in C_2 \cup A$.

It follows that

$$\begin{aligned}\prod_{\gamma\in F} \Psi_t^{F,\gamma} &= \overbrace{\prod_{\gamma\in C_1\cup\{\delta\}} \Psi_t^{C\cup\{\delta\},\gamma}}^{\Psi_t^1} \overbrace{\prod_{\gamma\in C_2\cup A} \Psi_t^{C\cup A,\gamma}}^{\Psi_t^2} \\ &= \Psi_t^1\Psi_t^2,\end{aligned}$$

cf. Figure 2. Note that Ψ_t^2 is $\mathcal{F}_t^{C\cup A}$-adapted. Let $\beta \in B$. We have $\langle A, B \mid C \cup \{\delta\}\rangle \in \mathcal{I}$, hence with $\bar\lambda_t^\beta = \lambda_t^{C\cup\{\delta\},\beta}$

$$\begin{aligned}E(\lambda_t^\beta \mid \mathcal{F}_t^{C\cup A}) &= E(E(\lambda_t^\beta \mid \mathcal{F}_t^{C\cup A\cup\{\delta\}}) \mid \mathcal{F}_t^{C\cup A}) \\ &= E(\bar\lambda_t^\beta \mid \mathcal{F}_t^{C\cup A}) \\ &= \frac{E_{Q_t^F}(\bar\lambda_t^\beta\Psi_t^1\Psi_t^2 \mid \mathcal{F}_t^{C\cup A})}{E_{Q_t^F}(\Psi_t^1\Psi_t^2 \mid \mathcal{F}_t^{C\cup A})} \\ &= \frac{E_{Q_t^F}(\bar\lambda_t^\beta\Psi_t^1 \mid \mathcal{F}_t^{C\cup A})}{E_{Q_t^F}(\Psi_t^1 \mid \mathcal{F}_t^{C\cup A})} \\ &= \frac{E_{Q_t^F}(\bar\lambda_t^\beta\Psi_t^1 \mid \mathcal{F}_t^{C})}{E_{Q_t^F}(\Psi_t^1 \mid \mathcal{F}_t^{C})} \\ &= \lambda_t^{C,\beta}\end{aligned}$$

where the second last identity follows from X^α for $\alpha \in A$ being independent of X^γ for $\gamma \in C \cup \{\delta\}$ under Q_t^F. We conclude that $\langle A, B \mid C\rangle \in \mathcal{I}$, and this shows that $\mathcal{I}$ is cancellative. □

4.1 ITÔ PROCESSES

For X a multivariate Itô process with X^α fulfilling the equation

$$X_t^\alpha = \int_0^t \lambda_s^\alpha \mathrm{d}s + \sigma_t(\alpha) W_t^\alpha$$

with W_t a standard Brownian motion ($\sigma_t(\alpha) > 0$ deterministic) we introduce the compatibility processes

$$\Psi_t^{A,\alpha} = \exp\left(\int_0^t \frac{\lambda_s^{A,\alpha}}{\sigma_s^2(\alpha)}\mathrm{d}X_s^\alpha - \frac{1}{2}\int_0^t \left(\frac{\lambda_s^{A,\alpha}}{\sigma_s(\alpha)}\right)^2 \mathrm{d}s\right).$$

The following result is a consequence of Theorem 7.3 in Liptser and Shiryayev (1977) combined with Theorem VI.8.4 in Rogers and Williams (2000).

Proposition 15. If for all $A \subseteq V$

$$E\left(\prod_{\alpha\in A} (\Psi_t^{A,\alpha})^{-1}\right) = 1 \tag{2}$$

then P λ-factorizes.

It can be shown that the linear SDE introduced earlier satisfies the integrability condition (2).

4.2 EVENT PROCESSES

For X a multivariate counting process with X^α having intensity process λ^α we introduce the compatibility processes

$$\Psi_t^{A,\alpha} = \exp\left(\int_0^t \log(\lambda_{s-}^{A,\alpha})\mathrm{d}X_s^\alpha - \int_0^t \lambda_s^{A,\alpha}\mathrm{d}s\right).$$

Here $\lambda^{A,\alpha}_{s-} = \lim_{r\to s-} \lambda^{A,\alpha}_r$ denotes the left continuous (and thus predictable) version of the intensity process $\lambda^{A,\alpha}_t = E(\lambda^\alpha_t \mid \mathcal{F}^A_t)$. With these compatibility processes, Proposition 15 above holds exactly as formulated for Itô processes, see e.g. Sokol and Hansen (2015) for details and weak conditions ensuring that (2) holds.

5 LEARNING ALGORITHMS

In this section, we assume that we have access to a local independence oracle that can answer whether or not some independence statement is in $\mathcal{I}$. In applications, the oracle would of course be substituted with statistical tests of local independence. The local independence model, $\mathcal{I}$, is assumed to be faithful to some DMG $\mathcal{G}_0$, i.e. $\mathcal{I} = \mathcal{I}(\mathcal{G}_0)$.

Meek (2014) described a related algorithm for learning local independence graphs which is, however, not complete when the system of stochastic processes is only partially observed. In the FCI algorithm, which learns an equivalence class of MAGs (Maximal Ancestral Graphs), one can exploit the fact that Markov equivalent graphs have the same adjacencies, so the learning algorithm can first find this so-called *skeleton* of the graph and then orient the edges by applying a finite set of rules (Zhang, 2008; Ali et al., 2009). Since Markov equivalent DMGs may have different adjacencies, we cannot straightforwardly copy the FCI strategy here, and our procedure is more complicated.

5.1 A THREE-STEP PROCEDURE

As described in Section 3.1.2, we know that there exists a unique graph which is Markov equivalent to $\mathcal{G}_0$ and a supergraph of all DMGs in $[\mathcal{G}_0]$ and we denote this graph by $\mathcal{N}$. In this section we give a learning algorithm exploiting this fact. Having learned the maximal DMG $\mathcal{N}$ we can subsequently construct a DMEG to summarize the Markov equivalence class.

The characterization of Markov equivalence of DMGs in Mogensen and Hansen (2018) implies a learning algorithm to construct $\mathcal{N}$ which is Markov equivalent to $\mathcal{G}_0$. For each pair of nodes α, β there exists a well-defined list of independence tests such that $\alpha \to \beta$ is in $\mathcal{N}$ if and only if all requirements in the list is met by $\mathcal{I}(\mathcal{G}_0)$, analogously for the edge $\alpha \leftrightarrow \beta$ (see conditions (p1)-(p4) and (s1)-(s3) in the supplementary material). This means that we can use these lists of tests to construct a maximal graph $\mathcal{N}$ such that $\mathcal{I}(\mathcal{N}) = \mathcal{I}(\mathcal{G}_0)$. However such an algorithm would perform many more independence tests than needed and one can reduce the number of independence tests conducted by a kind of preprocessing. Our proposed algorithm starts from the complete DMG and removes edges that are not in $\mathcal{G}_0$ by an FCI-like approach, exploiting properties of DMGs and μ-separation, and then in the end applies the potential parents and potential siblings definitions (see the supplementary material), but only if and when needed.

input : a local independence oracle for $\mathcal{I}$
output: a DMG, $\mathcal{G} = (V, E)$
initialize $\mathcal{G}$ as the complete DMG, set $n = 0$, initialize $\mathcal{L}_s = \emptyset, \mathcal{L}_n = \emptyset$;
while $n \leq \max_{\beta\in V} |\mathrm{pa}_{\mathcal{G}}(\beta)|$ **do**
  foreach $\alpha \to \beta \in E$ **do**
    foreach $C \subseteq \mathrm{pa}_{\mathcal{G}}(\beta)\backslash\{\alpha\}$, $|C| = n$ **do**
      if $\alpha \not\to_\lambda \beta \mid C$ **then**
        delete $\alpha \to \beta$ and $\alpha \leftrightarrow \beta$ from $\mathcal{G}$;
        update $\mathcal{L}_s = \mathcal{L}_s \cup \{\langle \alpha, \beta \mid C\rangle\}$;
      else
        update $\mathcal{L}_n = \mathcal{L}_n \cup \{\langle \alpha, \beta \mid C\rangle\}$;
      end
    end
  end
  update $n = n + 1$;
end
set $n = 1$;
while $n \leq \max_{\alpha,\beta\in V} |D_{\mathcal{G}}(\alpha, \beta)|$ **do**
  foreach $\alpha \to \beta \in E$ **do**
    foreach $C \subseteq D_{\mathcal{G}}(\alpha, \beta)$, $|C| = n$ **do**
      if $\alpha \not\to_\lambda \beta \mid C$ **then**
        delete $\alpha \to \beta$ and $\alpha \leftrightarrow \beta$ from $\mathcal{G}$;
        update $\mathcal{L}_s = \mathcal{L}_s \cup \{\langle \alpha, \beta \mid C\rangle\}$;
      else
        update $\mathcal{L}_n = \mathcal{L}_n \cup \{\langle \alpha, \beta \mid C\rangle\}$;
      end
    end
    update $n = n + 1$;
  end
end
return $\mathcal{G}$, $\mathcal{L}_s$, $\mathcal{L}_n$

Subalgorithm 1: Separation step

In this section we describe three steps (and three subalgorithms): a *separation*, a *pruning*, and a *potential* step, and then we argue that we can construct a sound and complete algorithm by using these steps. For all three steps, we sequentially remove edges starting from the complete DMG on nodes V. We will also along the way update a set of triples $\mathcal{L}_s$ corresponding to independence statements that we know to be in $\mathcal{I}$ and a set of triples $\mathcal{L}_n$ corresponding to independence statements that we know to not be in $\mathcal{I}$. We keep track of this information as we will reuse some of it to reduce the number of independence tests that we conduct. Figure 3 illustrates what

input : a separability graph, $\mathcal{S}$, a set of known independencies $\mathcal{L}_s$
output: a DMG
initialize $\mathcal{G} = \mathcal{S}$;
foreach *unshielded* W*-structure in* $\mathcal{S}$, ${}_w(\alpha, \beta, \gamma)$ **do**
  if $\beta \in S_{\alpha,\gamma}$ *such that* $\langle \alpha, \gamma \mid S_{\alpha,\gamma} \rangle \in \mathcal{L}_s$ **then**
    if $\beta \leftrightarrow \gamma$ *is in* $\mathcal{G}$ **then**
      delete $\beta \leftrightarrow \gamma$ from $\mathcal{G}$;
    end
  else
    if $\beta \rightarrow \gamma$ *is in* $\mathcal{G}$ **then**
      delete $\beta \rightarrow \gamma$ from $\mathcal{G}$;
    end
  end
end
return $\mathcal{G}$

Subalgorithm 2: Pruning step

input : a local independence oracle for $\mathcal{I}$, a DMG $\mathcal{G} = (V, E)$, a set of known dependencies $\mathcal{L}_n$
output: a DMG
foreach $\alpha \overset{e}{\rightarrow} \beta \in E$ **do**
  if $\mathcal{I}(\mathcal{G} - e) \cap \mathcal{L}_n = \emptyset$ **then**
    if $\alpha \notin \mathrm{pp}(\beta, \mathcal{I})$ **then**
      delete $\alpha \rightarrow \beta$ in $\mathcal{G}$;
    end
  end
end
foreach $\alpha \overset{e}{\leftrightarrow} \beta \in E$ **do**
  if $\mathcal{I}(\mathcal{G} - e) \cap \mathcal{L}_n = \emptyset$ **then**
    if $\alpha \notin \mathrm{ps}(\beta, \mathcal{I})$ **then**
      delete $\alpha \leftrightarrow \beta$ in $\mathcal{G}$;
    end
  end
end
return $\mathcal{G}$

Subalgorithm 3: Potential step

each subalgorithm outputs for an example $\mathcal{G}_0$.

5.1.1 The separation step

When we have an independence model $\mathcal{I}$ over V, we will for $\alpha, \beta \in V$ say that β is *inseparable* from α if there exists no $C \subseteq V \setminus \{\alpha\}$ such that $\langle \alpha, \beta \mid C \rangle \in \mathcal{I}$. Let

$$u(\beta, \mathcal{I}) = \{\gamma \in V \mid \beta \text{ is inseparable from } \gamma \text{ in } \mathcal{I}\}.$$

The purpose of the first step is to output a *separability graph*. The separability graph of an independence model $\mathcal{I}$ is the DMG such that the edge $\alpha \rightarrow \beta$ is in the DMG if and only if $\alpha \in u(\beta, \mathcal{I})$ and the edge $\alpha \leftrightarrow \beta$ is in the DMG if and only if $\alpha \in u(\beta, \mathcal{I})$ and $\beta \in u(\alpha, \mathcal{I})$.

We say that γ is *directedly collider connected to* β if there exists a non-trivial walk from γ to β such that every non-endpoint node on the walk is a collider and such that the final edge has a head at β. As shorthand, we write $\gamma \twoheadrightarrow \beta$. We define the separator set of β from α,

$$D_{\mathcal{G}}(\alpha, \beta) = \{\gamma \in \mathrm{an}(\alpha, \beta) \mid \gamma \twoheadrightarrow \beta\} \setminus \{\alpha\}.$$

If there exists a subset of $V \setminus \{\alpha\}$ that separates β from α, then this set does (Mogensen and Hansen, 2018). This set will play a role analogous to that of the set **Possible-D-Sep** in the FCI algorithm (Spirtes et al., 2000).

In the first part of Subalgorithm 1, we consider pairs of nodes, α, β, and test if they can be separated by larger and larger conditioning sets, though only subsets of $\mathrm{pa}_{\mathcal{G}}(\beta) \setminus \{\alpha\}$ in the current $\mathcal{G}$. In the second part, we use all subsets of the current separator set $D_{\mathcal{G}}(\alpha, \beta)$ to determine separability of each pair of nodes. Note that separability is not symmetric, hence, one needs to determine separability of β from α and of α from β. The candidate separator sets may be chosen in more-or-less efficient ways, but we will not discuss this aspect of the algorithm (Colombo et al., 2012; Claassen et al., 2013).

Lemma 16. Subalgorithm 1 outputs the separability graph of $\mathcal{I}$, $\mathcal{S}$, and furthermore $\mathcal{N} \subseteq \mathcal{S}$.

5.1.2 The pruning step

Let $\mathcal{S}$ denote the graph in the output of Subalgorithm 1. One can use some of the information encoded by the graph along with the set $\mathcal{L}_s$ to further prune the graph. For this purpose, we consider W*-structures* which are triples of nodes α, β, γ such that $\alpha \neq \beta \neq \gamma$, and $\alpha \rightarrow \beta \ast\!\!\rightarrow \gamma$. We denote such a triple by ${}_w(\alpha, \beta, \gamma)$. We will say that a W-structure is *unshielded* if the edge $\alpha \rightarrow \gamma$ is not in the graph. For every unshielded W-structure ${}_w(\alpha, \beta, \gamma)$, there exists exactly one triple $\langle \alpha, \gamma \mid C \rangle$ in $\mathcal{L}_s$ (output from Subalgorithm 1) and we let $S_{\alpha,\gamma}$ denote the separating set C.

Lemma 17. Subalgorithm 2 outputs a supergraph of $\mathcal{N}$.

5.1.3 Potential step

In the final step, we sequentially consider each edge which is still in the graph. If $\mathcal{G} = (V, E)$ and $e \in E$ we let $\mathcal{G} - e$ denote the DMG $(V, E \setminus \{e\})$. We then check if $\mathcal{I}(\mathcal{G} - e) \cap \mathcal{L}_n = \emptyset$. If not, we leave this edge in the graph. On the other hand, if the intersection is the empty set, we check if the edge is between a pair of potential parents/siblings using the definition of these sets. That is, in the case of a directed edge we check each of

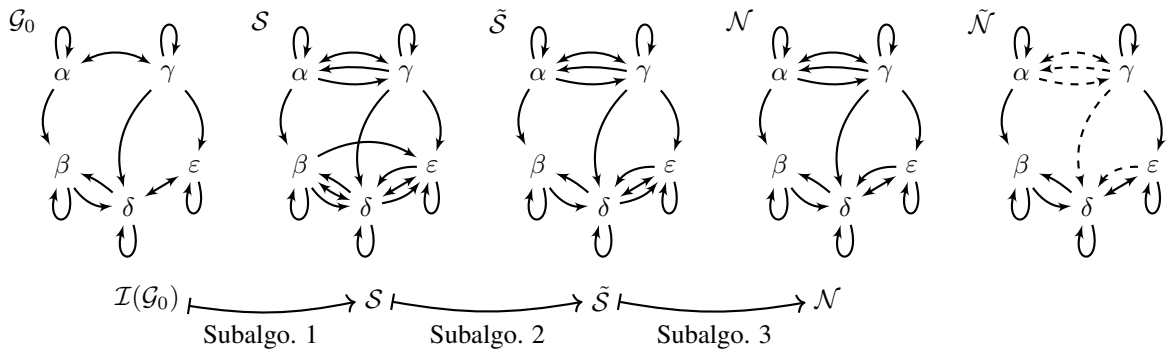

Figure 3: Illustration of the learning algorithm. The DMG $\mathcal{G}_0$ is the underlying graph and we have access to $\mathcal{I} = \mathcal{I}(\mathcal{G}_0)$. Subalgorithm 1 outputs $\mathcal{S}$, the separability graph of $\mathcal{I}(\mathcal{G}_0)$. Subalgorithm 2 prunes $\mathcal{S}$ and outputs $\tilde{\mathcal{S}}$. Note e.g. the unshielded W-structure $\alpha \to \beta \to \varepsilon$ in $\mathcal{S}$. The DMG $\mathcal{N}$ is the maximal element in $[G_0]$. Note that $\delta \to \varepsilon$ has been removed by Subalgorithm 3 using the potential parent criteria. The final graph $\tilde{\mathcal{N}}$ is the DMEG constructed from $\mathcal{N}$.

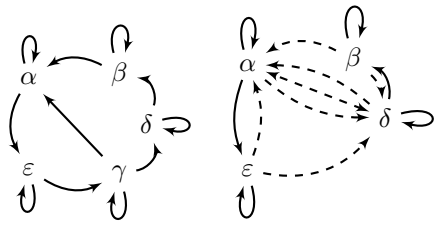

Figure 4: Left: linear SDE example (see Figure 1). Right: the DMEG after marginalization over γ. It is not possible to decide if a loop is directed or bidirected from the independence model only and we choose only to draw the directed loop and to not present it as dashed.

the conditions (p1)-(p4) and in the case of a bidirected edge each of the conditions (s1)-(s3); both sets of conditions are in the supplementary material. Note that if $\alpha \in \mathrm{ps}(\beta, \mathcal{I})$, then also $\beta \in \mathrm{ps}(\alpha, \mathcal{I})$.

Theorem 18. The algorithm defined by first doing the separation step, then the pruning, and finally the potential step outputs $\mathcal{N}$, the maximal element of $[\mathcal{G}_0]$.

Using properties of maximal DMGs, Mogensen and Hansen (2018) showed how one can construct the DMEG efficiently. The learning algorithm that is defined by first constructing $\mathcal{N}$ and then constructing the DMEG is sound and complete in the sense that if an edge is absent in the DMEG, then it is also absent in any element of $[\mathcal{G}_0]$ and therefore also in $\mathcal{G}_0$. If it is present and not dashed in the DMEG, then it is present in all elements of $[\mathcal{G}_0]$ and therefore also in $\mathcal{G}_0$. Finally, if it is present and dashed in the DMEG, then there exist $\mathcal{G}_1, \mathcal{G}_2 \in [\mathcal{G}_0]$ such that the edge is present in $\mathcal{G}_1$ and absent in $\mathcal{G}_2$ and therefore it is impossible to determine if the edge is in $\mathcal{G}_0$ using knowledge of $\mathcal{I}(\mathcal{G}_0)$ only.

One could also skip the potential step to reduce the computational requirements. The resulting DMG is then a supergraph of the true graph. A small simulation study (supplementary material) indicates that one could save quite a number of tests and still get close to the true $\mathcal{N}$.

6 CONCLUSION AND DISCUSSION

We have shown that for a given directed graph with μ-separation it is possible to specify abstract properties that ensure equivalence of the pairwise and global Markov properties in asymmetric independence models. We have shown that under certain conditions these properties hold in local independence models of Itô diffusions and event processes, extending known results.

Assuming faithfulness, we have given a sound and complete learning algorithm for the Markov equivalence class of directed mixed graphs representing a marginalized local independence model. Faithfulness is not an innocuous assumption and it remains an open research question how common this property is in different classes of stochastic processes.

Acknowledgements

SWM and NRH were supported by research grant 13358 from VILLUM FONDEN. DM was supported by research grant R01 AI127271-01A1 from the National Institutes of Health.

References

Odd O. Aalen. Dynamic modelling and causality. *Scandinavian Actuarial Journal*, pages 177–190, 1987.

Odd O. Aalen, Kjetil Røysland, Jon Michael Gran, and Bruno Ledergerber. Causality, mediation and time: a dynamic viewpoint. *Journal of the Royal Statistical Society, Series A*, 175(4):831–861, 2012.

Ayesha R. Ali, Thomas S. Richardson, and Peter Spirtes. Markov equivalence for ancestral graphs. *The Annals of Statistics*, 37(5B):2808–2837, 2009.

Tom Claassen, Joris Mooij, and Tom Heskes. Learning sparse causal models is not NP-hard. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, pages 172–181, 2013.

Diego Colombo, Marloes H. Maathuis, Markus Kalisch, and Thomas S. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 40(1): 294–321, 2012.

Daniel Commenges and Anne Gégout-Petit. A general dynamical statistical model with causal interpretation. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 71(3):719–736, 2009.

David Danks and Sergey Plis. Learning causal structure from undersampled time series. In *JMLR: Workshop and Conference Proceedings*, volume 10, pages 1–10, 2013.

Vanessa Didelez. *Graphical Models for Event History Analysis based on Local Independence*. PhD thesis, Universität Dortmund, 2000.

Vanessa Didelez. Graphical models for composable finite Markov processes. *Scandinavian Journal of Statistics*, 34(1):169–185, 2006a.

Vanessa Didelez. Asymmetric separation for local independence graphs. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, 2006b.

Vanessa Didelez. Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society, Series B*, 70(1):245–264, 2008.

Vanessa Didelez. Causal reasoning for events in continuous time: A decision-theoretic approach. In *Proceedings of the UAI 2015 Workshop on Advances in Causal Inference*, 2015.

Asela Gunawardana, Christopher Meek, and Puyang Xu. A model for temporal dependencies in event streams. In *Advances in Neural Information Processing Systems 24 (NIPS 2011)*, 2011.

Steffen Lauritzen. *Graphical Models*. Oxford: Clarendon, 1996.

Steffen Lauritzen and Kayvan Sadeghi. Unifying Markov properties for graphical models. 2017. URL `https://arxiv.org/abs/1608.05810`.

R.S. Liptser and A.N. Shiryayev. *Statistics of Random Processes I: General Theory*. Springer-Verlag, 1977.

Christopher Meek. Toward learning graphical and causal process models. In *Proceedings of the UAI 2014 Workshop on Causal Inference: Learning and Prediction*, 2014.

Søren Wengel Mogensen and Niels Richard Hansen. Markov equivalence of marginalized local independence graphs. 2018. URL `https://arxiv.org/abs/1802.10163`.

Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. From ordinary differential equations to structural causal models: the deterministic case. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, 2013.

U. Nodelman, C. R. Shelton, and D. Koller. Continuous time Bayesian networks. In *Proceedings of the 18th Conference on Uncertainty in Artifical Intelligence*, pages 378–87, 2002.

U. Nodelman, C. R. Shelton, and D. Koller. Learning continuous time Bayesian networks. In *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence*, pages 451–8, 2003.

Thomas S. Richardson. A discovery algorithm for directed cyclic graphs. In *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence*, 1996.

Thomas S. Richardson. A characterization of Markov equivalence for directed cyclic graphs. *International Journal of Approximate Reasoning*, 17:107–162, 1997.

L. C. G. Rogers and David Williams. *Diffusions, Markov processes, and martingales. Vol. 2*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 2000.

Paul K. Rubenstein, Stephan Bongers, Joris M. Mooij, and Bernhard Schölkopf. From deterministic ODEs to dynamic structural causal models. *arXiv.org preprint*, arXiv:1608.08028 [cs.AI], 2016. URL `http://arxiv.org/abs/1608.08028`.

Kayvan Sadeghi. Faithfulness of probability distributions and graphs. *Journal of Machine Learning Research*, 18(148):1–29, 2017.

Tore Schweder. Composable Markov processes. *Journal of Applied Probability*, 7(2):400–410, 1970.

Alexander Sokol and Niels Richard Hansen. Causal interpretation of stochastic differential equations. *Electronic Journal of Probability*, 19(100):1–24, 2014.

Alexander Sokol and Niels Richard Hansen. Exponential martingales and changes of measure for counting processes. *Stochastic Analysis and Applications*, 33(5): 823–843, 2015.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2000.

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172:1873–1896, 2008.

SUPPLEMENTARY MATERIAL

This supplementary material contains proofs that were omitted in the paper. It also contains the potential parent and potential sibling criteria and reports the results of a small simulation study illustrating the cost and the impact of the potential step in the learning algorithm.

A PROOFS OF LEMMAS 5 AND 6

Lemma 5. The independence model $\mathcal{I}(\mathcal{G})$ satisfies left and right {decomposition, weak union, composition} and left {redundancy, intersection, weak composition}. Furthermore, $\langle A, B \mid C\rangle \in \mathcal{I}(\mathcal{G})$ whenever $B = \emptyset$.

Proof. Left redundancy, left and right decomposition and left and right composition follow directly from the definition of μ-separation. Left and right weak union are also immediate. Left weak composition follows from left redundancy, left decomposition and left composition. It is also clear that $\langle A, B \mid C\rangle \in \mathcal{I}(\mathcal{G})$ if $B = \emptyset$.

For left intersection, consider a μ-connecting walk, $\omega = \langle \nu_1, e_1, \ldots, e_n, \nu_{n+1}\rangle$ from $\delta = \nu_1 \in A \cup C$ to $\beta = \nu_{n+1} \in B$ given $A \cap C$. This walk is by definition non-trivial. Consider now the shortest possible non-trivial subwalk of ω of the form $\tilde{\omega} = \langle \nu_i, e_i, \ldots, e_n, \nu_{n+1}\rangle$ such that $\nu_i \in (A \cup C) \setminus (A \cap C)$. Such a subwalk always exists and it is μ-connecting either from A to B given C or from C to B given A. □

Lemma 6. $\mathcal{I}(\mathcal{G})$ satisfies cancellation.

Proof. The contrapositive of $A \perp_\mu B \mid C \cup \{\delta\} \Rightarrow A \perp_\mu B \mid C$ is $A \not\perp_\mu B \mid C \Rightarrow A \not\perp_\mu B \mid C \cup \{\delta\}$. So we have that $A \perp_\mu C_1 \cup \{\delta\} \mid C \cup \{\delta\}$, $\delta \perp_\mu C_2 \cup A \mid C \cup A$, and $A \not\perp_\mu B \mid C$ and want to show that $A \not\perp_\mu B \mid C \cup \{\delta\}$. Note that $A \perp_\mu \delta \mid C \cup \{\delta\}$ by right decomposition.

There exists a μ-connecting walk ω from $\alpha \in A$ to some $\beta \in B$ given C, and we argue that this walk is also μ-connecting given $C \cup \{\delta\}$. Suppose not, for contradiction. Note that $\alpha \notin C$ so $\alpha \notin C \cup \{\delta\}$ since by factorization $A, C, \{\delta\}$ are disjoint. Also every collider on ω is in $\mathrm{an}(C)$ so it is in $\mathrm{an}(C \cup \{\delta\})$. Thus if ω is not μ-connecting given $C \cup \{\delta\}$ it must be because there is some non-collider on ω which is not in C but is in $C \cup \{\delta\}$, i.e., the non-collider is δ. Choose now a subwalk of ω between some (possibly different) $\alpha \in A$ and δ such that no non-endpoint node of this subwalk is in $A \cup \{\delta\}$. Again, $\alpha \notin C \cup \{\delta\}$. Such a subwalk always exists.

There are two possibilities: either there is an arrowhead into δ on this subwalk of ω or there is not. In the first case, the subwalk of ω from α into δ is μ-connecting given $C \cup \{\delta\}$, i.e., $A \not\perp_\mu \delta \mid C \cup \{\delta\}$. Contradiction. In the second case, we consider a collider ε on the subwalk between α and δ (if there is no collider on the walk, then the directed walk from δ to α is μ-connecting given $C \cup A$). Either $\varepsilon \in C_1$, $\varepsilon \in C_2$, or there is a (non-trivial) directed walk from ε to some ε' that is either in C_1 or C_2. If $\varepsilon \in C_1$, there is a μ-connecting subwalk of ω from α to $\varepsilon \in C_1$ given C. Since there are no non-colliders on this walk in $\{\delta\}$, it is also μ-connecting given $C \cup \{\delta\}$. If $\varepsilon \in C_2$, likewise there is a μ-connecting walk from δ to C_2 given $C \cup A$ (note that there are no non-colliders in A on this walk by choice of α). Either way, contradiction.

If $\varepsilon \notin C$, we consider concatenating one of the aforementioned walks to ε with the directed path ω' from ε to $\varepsilon' \in C$. Either δ appears on ω' or it does not. In the first case, then there is an arrowhead at δ on ω' and so $A \not\perp_\mu \delta \mid C \cup \{\delta\}$ as before. In the latter case, there are two subcases to consider: either there is some vertex in A on ω' or there is not. If there is, choose $\alpha' \in A$ on ω' such that there are no vertices in A nearer to ε on ω'. Then the the walk from δ to α' is μ-connecting given $C \cup A$. If there is no vertex in A on ω', then by concatenating a subwalk of ω to ω' we get a μ-connecting walk from α or δ to ε' in C_1 or C_2 given $C \cup \{\delta\}$ or $C \cup A$, respectively. In any case, contradiction. □

B PROOF OF THEOREM 8

In this section, we first prove some lemmas and then use these to prove Theorem 8.

Lemma 19. If $A \perp_\mu B \mid C$ and $A \perp_\mu D \mid C$, then $A \perp_\mu B \mid C \cup D$.

Proof. This follows from right composition, right weak union, and right decomposition of μ-separation. □

Lemma 20. Assume $\gamma \in \mathrm{an}(A \cup B \cup C)$ and $\alpha, \gamma \notin C$. If there is a walk between $\alpha \in A$ and γ such that no non-collider is in C and every collider is in $\mathrm{an}(C)$, and there

is a μ-connecting walk from γ to $\beta \in B$ given C, then there is a μ-connecting walk from A to B given C.

If $\omega = \langle \nu_1, e_1, \nu_2, \ldots, e_n, \nu_{n+1} \rangle$ is a walk, then the inverse, ω^{-1}, is the walk $\langle \nu_{n+1}, e_n, \nu_n, \ldots, e_1, \nu_1 \rangle$.

Proof. If $\gamma \in \mathrm{an}(C)$, then simply compose the walks. Assume $\gamma \notin \mathrm{an}(C)$. If $\gamma \in \mathrm{an}(A)$ let π denote the directed path from γ to $\bar{\alpha} \in A$. We have that there is no node in C on π and composing π^{-1} with the μ-connecting walk from γ to B gives a μ-connecting walk from $\bar{\alpha} \in A$ to $\beta \in B$ given C. If $\gamma \in \mathrm{an}(B)$ compose the walk from α to γ with the directed path from γ to B (which is μ-connecting given C as $\gamma \notin \mathrm{an}(C)$). □

Lemma 21. Assume that $\mathcal{I}$ satisfies left weak composition, left intersection, and left decomposition. If $A \cap D = \emptyset$ then

$$\langle A, B \mid C \cup D \rangle \in \mathcal{I}, \langle D, B \mid C \cup A \rangle \in \mathcal{I} \Rightarrow \langle A \cup D, B \mid C \rangle \in \mathcal{I}.$$

Proof. By left weak composition $\langle A \cup C, B \mid C \cup D \rangle \in \mathcal{I}, \langle D \cup C, B \mid C \cup A \rangle \in \mathcal{I}$. It follows by left intersection that $\langle A \cup C \cup D, B \mid C \rangle \in \mathcal{I}$ and by left decomposition the result follows. □

Lemma 22. Let $\mathcal{D} = (V, E)$ be a DG, and let $\alpha, \beta \in V$. Then $\alpha \notin \mathrm{pa}_{\mathcal{D}}(\beta)$ if and only if $\alpha \perp_\mu \beta \mid V \setminus \{\alpha\}$.

In the following proofs, we will use $\sim$ to denote an arbitrary edge.

Proof. Assume first that $\alpha \notin \mathrm{pa}_{\mathcal{D}}(\beta)$, and consider a walk between α and β that has a head at β, $\alpha \sim \ldots \sim \gamma \to \beta$. We must have that $\alpha \neq \gamma$ and therefore the walk is not μ-connecting given $V \setminus \{\alpha\}$.

Assume instead that $\alpha \perp_\mu \beta \mid V \setminus \{\alpha\}$. The edge $\alpha \to \beta$ would constitute a μ-connecting walk given $V \setminus \{\alpha\}$ and therefore we must have that $\alpha \notin \mathrm{pa}_{\mathcal{D}}(\beta)$. □

Theorem 8. Assume that $\mathcal{I}$ is an independence model that satisfies left {redundancy, intersection, decomposition, weak union, weak composition}, right {decomposition, composition}, is cancellative, and furthermore $\langle A, B \mid C \rangle \in \mathcal{I}$ whenever $B = \emptyset$. Let $\mathcal{D}$ be a DG. Then $\mathcal{I}$ satisfies the pairwise Markov property with respect to $\mathcal{D}$ if and only if it satisfies the global Markov property with respect to $\mathcal{D}$.

Proof. It follows directly from the definitions and Lemma 22 that the global Markov property implies the pairwise Markov property. Assume that $\mathcal{I}$ satisfies the pairwise Markov property w.r.t. $\mathcal{D}$ and let $A, B, C \subseteq V$. Assume $A \perp_\mu B \mid C$. We wish to show that $\langle A, B \mid C \rangle \in \mathcal{I}$.

Assume $|V| = n > 0$. We will proceed using reverse induction on $|C|$. As the induction base, $C = V$. The result follows by noting that $\langle V, B \mid V \rangle \in \mathcal{I}$ by left redundancy of $\mathcal{I}$. By left decomposition of $\mathcal{I}$, we get $\langle A, B \mid V \rangle \in \mathcal{I}$.

For the induction step, consider a node $\gamma \notin C$. Note first that if $A \subseteq C$, then the result once again follows using left redundancy and then left decomposition, and therefore assume that $A \setminus C \neq \emptyset$, and take $\alpha \in A \setminus C$ (note that $\alpha = \gamma$ is allowed). Assume first that we cannot choose α and γ such that $\alpha \neq \gamma$. This means that $C = V \setminus \{\alpha\}$. By right decomposition of $\mathcal{I}(\mathcal{G})$ we have that $A \perp_\mu \beta \mid C$ for all $\beta \in B$, and by left decomposition of $\mathcal{I}(\mathcal{G})$ we have $\alpha \perp_\mu \beta \mid C$. If $B = \emptyset$, then the result follows by assumption, and else by the pairwise Markov property and Lemma 22 we have $\langle \alpha, \beta \mid C \rangle \in \mathcal{I}$ for all $\beta \in B$ and by right composition of $\mathcal{I}$ we have $\langle \alpha, B \mid C \rangle \in \mathcal{I}$. By left weak composition, we have $\langle A, B \mid C \rangle \in \mathcal{I}$.

Now assume $\gamma \neq \alpha$. We split the proof into two cases, (i) and (ii), depending on whether or not we can choose γ as an ancestor to $A \cup B \cup C$.

Case (i): $\gamma \in \mathrm{an}(A \cup B \cup C)$
We have that $\gamma \perp_\mu B \mid C$ or $A \perp_\mu \gamma \mid C$ by Lemma 20. We split into two subcases, (i-1) and (i-2).

Case (i-1): $\gamma \perp_\mu B \mid C$
By left composition of $\mathcal{I}(\mathcal{G})$, $A \cup \{\gamma\} \perp_\mu B \mid C$ and by left weak union $A \cup \{\gamma\} \perp_\mu B \mid C \cup \{\gamma\}$ as well as $A \cup \{\gamma\} \perp_\mu B \mid C \cup (A \setminus \{\gamma\})$. By the induction hypothesis and noting that $C \cup \{\gamma\} \neq C \neq C \cup (A \setminus \{\gamma\})$, $\langle A \cup \{\gamma\}, B \mid C \cup \{\gamma\} \rangle \in \mathcal{I}$, and $\langle A \cup \{\gamma\}, B \mid C \cup (A \setminus \{\gamma\}) \rangle \in \mathcal{I}$. By left decomposition of $\mathcal{I}$ and Lemma 21, the result follows.

Case (i-2): $A \perp_\mu \gamma \mid C$
In this case, we can assume that $\gamma \notin A$, as otherwise by left decomposition of $\mathcal{I}(\mathcal{G})$ we would also have $\gamma \perp_\mu B \mid C$ which is case (i-1). Moreover, either $\gamma \perp_\mu B \mid C$ or $\gamma \perp_\mu A \setminus C \mid C$, as otherwise $A \perp_\mu B \mid C$ would not hold (Lemma 20). $\gamma \perp_\mu B \mid C$ is the above case, so assume that $\gamma \not\perp_\mu B \mid C$ and $\gamma \perp_\mu A \setminus C \mid C$. Using right weak union of $\mathcal{I}(\mathcal{G})$, we have $A \perp_\mu \gamma \mid C \cup \{\gamma\}$ and $\gamma \perp_\mu A \setminus C \mid C \cup A$. Using the induction assumption, we have that $\langle A, \gamma \mid C \cup \{\gamma\} \rangle \in \mathcal{I}$ and $\langle \gamma, A \setminus C \mid C \cup A \rangle \in \mathcal{I}$. We have $A \perp_\mu B \mid C$ and $A \perp_\mu \gamma \mid C$ and using right composition and right weak union of $\mathcal{I}(\mathcal{G})$, we obtain $A \perp_\mu B \cup \{\gamma\} \mid C \cup \{\gamma\}$. Using the induction assumption we have that $\langle A, B \mid C \cup \{\gamma\} \rangle \in \mathcal{I}$. Assume to obtain a contradiction that $A \not\perp_\mu \delta \mid C \cup \gamma$ and $\gamma \not\perp_\mu \delta \mid C \cup A$

for some $\delta \in C$. We know that $A \perp_\mu \gamma \mid C$ and by using the contrapositive of Lemma 19 this means that $A \not\perp_\mu \delta \mid C$. Similarly, we obtain that $\gamma \not\perp_\mu \delta \mid C$. We note that $\gamma \not\perp_\mu B \mid C$ and by Lemma 20 this means that $A \not\perp_\mu B \mid C$ which is a contradiction. Therefore, we have that for each $\delta \in C$, either $A \perp_\mu \delta \mid C \cup \gamma$ (and therefore also $A \setminus C \perp_\mu \delta \mid C \cup \gamma$) or $\gamma \perp_\mu \delta \mid C \cup A$. Using the induction assumption, right composition of $\mathcal{I}$, the cancellation property and left weak composition of $\mathcal{I}$ we arrive at the conclusion.

Case (ii): If one cannot choose a $\gamma \in \text{an}(A \cup B \cup C)$ such that $\gamma \notin C$ and $\gamma \neq \alpha$, then $\text{an}(A \cup B \cup C) = C \cup \{\alpha\}$. Assume this and furthermore assume that $\gamma \notin \text{an}(A \cup B \cup C)$. We will first argue that $A \perp_\mu B \mid C \cup \{\gamma\}$. If this was not the case there would be a μ-connecting walk, ω, from A to $\beta \in B$ given $C \cup \{\gamma\}$ on which γ was a collider and furthermore every collider was in $C \cup \{\gamma\}$. Consider now the last occurrence of γ on this walk, and the subwalk of ω, $\gamma \sim \ldots \sim \theta \sim \ldots \to \beta$. Let θ be the node in $\text{an}(A \cup B \cup C)$ which is the closest to γ on the walk. Then there must be a tail at θ, and this means that $\theta = \alpha$ as otherwise the walk would be closed. In this case, the subwalk from α to β would also be μ-connecting given C which is a contradiction.

It also holds that $\gamma \perp_\mu B \mid C \cup A$ as every parent of a node in B is in $C \cup A$. Using the induction assumption we have that $\langle A, B \mid C \cup \{\gamma\}\rangle \in \mathcal{I}$ and $\langle \gamma, B \mid C \cup A\rangle \in \mathcal{I}$ and using Lemma 21 and left decomposition of $\mathcal{I}$ we obtain $\langle A, B \mid C\rangle \in \mathcal{I}$.

□

C PROOF OF LEMMA 11

Lemma 11. Let $\mathcal{I}$ be a local independence model. Then it satisfies left {redundancy, decomposition, weak union, weak composition} and right {decomposition, composition} and furthermore $\langle A, B \mid C\rangle \in \mathcal{I}$ whenever $B = \emptyset$. If $\mathcal{F}_t^A \cap \mathcal{F}_t^C = \mathcal{F}_t^{A \cap C}$ holds for all $A, C \subseteq V$ and $t \in [0, T]$, then left intersection holds.

Proof. Left redundancy: We note that $\mathcal{F}_t^{A \cup C} = \mathcal{F}_t^C$ from which the result follows.

Left decomposition: Assume that $A_1 \cup A_2 \not\to_\lambda B \mid C$. We wish to show that $A_1 \not\to_\lambda B \mid C$.

$$\begin{aligned} E(\lambda_t^\beta \mid \mathcal{F}_t^{A_1 \cup C}) &= E\big(\underbrace{E(\lambda_t^\beta \mid \mathcal{F}_t^{A_1 \cup A_2 \cup C})}_{=E(\lambda_t^B \mid \mathcal{F}_t^C)} \,\big|\, \mathcal{F}_t^{A_1 \cup C}\big) \\ &= E(\lambda_t^\beta \mid \mathcal{F}_t^C) \end{aligned}$$

Left weak union: Simply note that the conditioning σ-algebra stays the same in the conditional expectation which is assumed to be $\mathcal{F}_t^C$-adapted and therefore also $\mathcal{F}_t^{C \cup D}$-adapted.

Left weak composition: The conditioning σ-algebra again stays the same in the conditional expectation.

Right decomposition and right composition follow directly from the coordinate-wise definition of local independence.

Left intersection: We note that $E(\lambda_t^\beta \mid \mathcal{F}_t^{A \cup C})$ by assumption has an $\mathcal{F}_t^A$-adapted and an $\mathcal{F}_t^C$-adapted version, thus it has a version, which is adapted w.r.t. the filtration $\mathcal{F}_t^A \cap \mathcal{F}_t^C = \mathcal{F}_t^{A \cap C}$.

Finally, it is clear that $\langle A, B \mid C\rangle \in \mathcal{I}$ if $B = \emptyset$ as this makes the condition void. □

D PROOFS, SECTION 5

Lemma 16. Subalgorithm 1 outputs the separability graph of $\mathcal{I}$, $\mathcal{S}$, and furthermore $\mathcal{N} \subseteq \mathcal{S}$.

Proof. In Subalgorithm 1, we only remove edges $\alpha \ast\!\!\to \beta$ when we have found a set $C \subseteq V \setminus \{\alpha\}$ that separates β from α. The DMGs $\mathcal{G}_0$ and $\mathcal{N}$ are Markov equivalent and therefore the same separation holds in $\mathcal{I}(\mathcal{N})$. Such an edge would always be μ-connecting from α to β given C as $\alpha \notin C$ and therefore we know it to be absent in $\mathcal{N}$. This means that the output of the algorithm is a supergraph of $\mathcal{N}$.

The graph $\mathcal{G}$ in Subalgorithm 1 is always a supergraph of $\mathcal{G}_0$ and therefore $D_{\mathcal{G}_0}(\alpha, \beta) \subseteq D_{\mathcal{G}}(\alpha, \beta)$. If there exists a set that separates β from α then $D_{\mathcal{G}_0}(\alpha, \beta)$ does and by the above inclusion we are always sure to test this set. This means that the output is the separability graph. □

Lemma 17. Subalgorithm 2 outputs a supergraph of $\mathcal{N}$.

Proof. By Lemma 16, $\mathcal{N} \subseteq \mathcal{S}$. We also know that if there is an edge $\alpha \to \beta$ in $\mathcal{S}$ then $\alpha \in u(\beta, \mathcal{I}(\mathcal{G}_0)) = u(\beta, \mathcal{I}(\mathcal{N})) = u(\beta, \mathcal{I})$. Assume there is an unshielded W-structure ${}_w(\alpha, \beta, \gamma)$ in $\mathcal{S}$. The edge between α and β in $\mathcal{S}$ means that β cannot be separated from α in $\mathcal{I}(\mathcal{N})$ and therefore there exists for every $C \subseteq V \setminus \{\alpha\}$ a μ-connecting walk from α to β given C. By definition of μ-connecting walks this has a head at (the final) β. The W-structure is unshielded, that is, $\alpha \to \gamma$ is not in $\mathcal{S}$. This means that we have previously found a separating set $S_{\alpha,\gamma}$, such that $\langle \alpha, \gamma \mid S_{\alpha,\beta}\rangle \in \mathcal{I}(\mathcal{N})$ and $\alpha \notin S_{\alpha,\gamma}$. We know that there exists a μ-connecting walk ω, from α to β given $S_{\alpha,\gamma}$ in $\mathcal{N}$ as $\alpha \in u(\beta, \mathcal{I}(\mathcal{N}))$. If $\beta \notin S_{\alpha,\gamma}$ then we can compose ω with the edge $\beta \to \gamma$ which

gives a μ-connecting walk from α to γ given $S_{\alpha,\gamma}$ which is a contradiction, and therefore the edge $\beta \to \gamma$ cannot be in $\mathcal{N}$. If $\beta \in S_{\alpha,\gamma}$ then we can argue analogously and obtain that $\beta \leftrightarrow \gamma$ cannot be in $\mathcal{N}$. □

Theorem 18. The algorithm defined by first doing the separation step, then the pruning, and finally the potential step outputs $\mathcal{N}$, the maximal element of $[\mathcal{G}_0]$.

Proof. By Lemma 17, the output after the first two steps is a supergraph of $\mathcal{N}$. In the potential step, an edge $\alpha \to \beta$ is only removed if α is not a potential parent of β in $\mathcal{I}$. We know that if the edge is in $\mathcal{N}$ then α is a potential parent of β in $\mathcal{I}(\mathcal{N}) = \mathcal{I}(\mathcal{G}_0) = \mathcal{I}$ (Mogensen and Hansen, 2018) and by contraposition of this result it follows that every directed edge removed is not in $\mathcal{N}$. The same argument applies in the case of a bidirected edge and therefore the output is a supergraph of $\mathcal{N}$.

If we consider some edge $\alpha \overset{e}{\to} \beta$ in the output graph, then either α is a potential parent of β, in which case e is also in $\mathcal{N}$, or $\mathcal{I}(\mathcal{G} - e) \cap \mathcal{L}_n \neq \emptyset$. Assume the latter. We have that $\mathcal{G}_0 \subseteq \mathcal{G}$, and therefore $\mathcal{I}(\mathcal{G} - e) \subseteq \mathcal{I}(\mathcal{G}_0)$ if e is not in $\mathcal{G}_0$. The above intersection is non-empty and therefore there is some triple which is in both $\mathcal{I}(\mathcal{G} - e)$ and $\mathcal{L}_n$, and by $\mathcal{I}(\mathcal{G} - e) \subseteq \mathcal{I}(\mathcal{G}_0)$ it is also in $\mathcal{I}(\mathcal{G}_0)$. But by definition $\mathcal{L}_n$ contains only triples not in $\mathcal{I}(\mathcal{G}_0)$, so this is a contradiction. Therefore, e must be in $\mathcal{G}_0$ and also in $\mathcal{N}$ as $\mathcal{G}_0 \subseteq \mathcal{N}$. One can argue analogously for the bidirected edges. We conclude that the output graph is equal to $\mathcal{N}$, the maximal element of $[\mathcal{G}_0]$. □

E POTENTIAL PARENT/SIBLINGS

Consider an independence model, $\mathcal{I}$, over V and let $\alpha, \beta \in V$. The set $u(\beta, \mathcal{I})$ is defined in Subsection 5.1.1. As described in Subsection 5.1 the below definitions define a list of independence tests which one can conduct to directly construct $\mathcal{N}$. This was proven by Mogensen and Hansen (2018). However, the list is very large and one can construct $\mathcal{N}$ in a more efficient manner. If e.g. $|V| = 10$, then for each choice of γ in (s2) we can choose C in 2^8 different ways (omitting sets C containing γ as such an independence would hold trivially for any independence model satisfying left redundancy and left decomposition).

Definition 23. We say that α and β are *potential siblings* in the independence model $\mathcal{I}$ if (s1)-(s3) hold:

(s1) $\beta \in u(\alpha, \mathcal{I})$ and $\alpha \in u(\beta, \mathcal{I})$,

(s2) for all $\gamma \in V$, $C \subseteq V$ such that $\beta \in C$,

$$\langle \gamma, \alpha \mid C \rangle \in \mathcal{I} \Rightarrow \langle \gamma, \beta \mid C \rangle \in \mathcal{I},$$

(s3) for all $\gamma \in V$, $C \subseteq V$ such that $\alpha \in C$,

$$\langle \gamma, \beta \mid C \rangle \in \mathcal{I} \Rightarrow \langle \gamma, \alpha \mid C \rangle \in \mathcal{I}.$$

Definition 24. We say that α is a *potential parent* of β in the independence model $\mathcal{I}$ if (p1)-(p4) hold:

(p1) $\alpha \in u(\beta, \mathcal{I})$,

(p2) for all $\gamma \in V$, $C \subseteq V$ such that $\alpha \notin C$,

$$\langle \gamma, \beta \mid C \rangle \Rightarrow \langle \gamma, \alpha \mid C \rangle,$$

(p3) for all $\gamma, \delta \in V$, $C \subseteq V$ such that $\alpha \notin C, \beta \in C$,

$$\langle \gamma, \delta \mid C \rangle \Rightarrow \langle \gamma, \beta \mid C \rangle \vee \langle \alpha, \delta \mid C \rangle,$$

(p4) for all $\gamma \in V, C \subseteq V$, such that $\alpha \notin C$,

$$\langle \beta, \gamma \mid C \rangle \Rightarrow \langle \beta, \gamma \mid C \cup \{\alpha\} \rangle.$$

F SIMULATION STUDY

We conducted a small simulation study to empirically evaluate the cost and impact of the third step in the learning algorithm, the potential step. This step is computationally expensive as it involves testing the potential parent/siblings conditions, see above.

We simulated a random DMG on 5 nodes by first drawing p_d from a uniform distribution on $[0, 1/2]$ and p_b from a uniform distribution on $[0, 1/4]$. We then generated independent Bernoulli random variates, $\{b_{\langle \alpha,\beta \rangle}\}$, each with success parameter p_d, and one for each ordered pair of nodes, $\langle \alpha, \beta \rangle$. The edge $\alpha \to \beta$ was included if $b_{\langle \alpha,\beta \rangle} = 1$. For each unordered pair of nodes, $\{\alpha, \beta\}$, we did analogously, using p_b as success parameter. We discarded graphs for which the maximal Markov equivalent graph had more then 15 edges.

Simulating 800 random DMGs, we saw that on average the first step required 90 independence tests and removed 26 edges. The second step removed 1.1 edge on average (it does not use any additional independence tests), while the third required an additional 77 independence tests. On average the third step removed 0.8 edge. This simulation is very limited and simple, however, it does indicate that the potential step of the learning algorithm constitutes a substantial part of the computational cost while not removing a lot of edges.

References