Faithful graphical representations of local independence

Søren Wengel Mogensen
Department of Automatic Control, Lund University

Abstract

Graphical models use graphs to represent conditional independence structure in the distribution of a random vector. In stochastic processes, graphs may represent so-called local independence or conditional Granger causality. Under some regularity conditions, a local independence graph implies a set of independences using a graphical criterion known as δ-separation, or using its generalization, μ-separation. This is a stochastic process analogue of d-separation in DAGs. However, there may be more independences than implied by this graph and this is a violation of so-called *faithfulness*. We characterize faithfulness in local independence graphs and give a method to construct a faithful graph from any local independence model such that the output equals the true graph when Markov and faithfulness assumptions hold. We discuss various assumptions that are weaker than faithfulness, and we explore different structure learning algorithms and their properties under varying assumptions.

1 INTRODUCTION

Graphical models are widely used and so-called *Markov properties* are essential as they describe how graphs encode conditional independence (Lauritzen, 1996). While such Markov properties hold under fairly general conditions, it is well-understood that conditional independence models are too complicated to be described completely by these properties. One particular issue is the potential lack of *faithfulness* such that the graph encodes a dependence which is not in the probability distribution (Spirtes and Zhang, 2018).

Proceedings of the 27$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

In models of multivariate stochastic processes, tests of *local independence* or *Granger causality* may be used to learn a causal graph in which each node represents a coordinate process. Most prior work assumes that the causal graph is Markov and faithful with respect to the observed independences. This might not hold, even in the theoretical distribution from which we sample data. Moreover, when presented with real data, we need statistical tests of local independence and therefore wrong test results will also distort the output.

In this paper, we characterize faithfulness and discuss a hierarchy of faithfulness assumptions that are relevant in this context. We describe differences between structure learning in DAG-based models and in stochastic process models. We compare different algorithms for use in stochastic process models and highlight how to minimize the impact of faithfulness issues. We start by defining the two independence relations that we will use.

1.1 Local Independence

Local independence is a ternary independence relation (Schweder, 1970; Aalen, 1987; Didelez, 2008) and we will use graphs to represent local independence in a multivariate stochastic process, analogously to how graphs may encode conditional independence in the distribution of a random vector.

The definition of local independence will depend on the class of stochastic processes we consider. We follow the definition in Mogensen et al. (2018). Let $X_t = (X_t^1, \dots, X_t^n)$ be a continuous-time stochastic process. We say that X_t^i is a *coordinate process*. We let $V = \{1, 2, \dots, n\}$. For $D \subseteq V$, we define $\mathcal{F}_t^D$ as the completed and right-continuous version of $\sigma(\{X_s^\alpha : s < t, \alpha \in D\})$.

Definition 1.1 (Local independence)**.** *Let $\lambda_t = (\lambda_t^1, \dots, \lambda_t^n)$ be a stochastic process. Let $A, B, C \subseteq V$. We say that X^B is* locally independent *of X^A given X^C, or simply that B is* locally independent *of A given C, if for all $\beta \in B$*

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

has an $\mathcal{F}_t^C$-adapted version.

The above definition does not answer the important question: What should the λ-process be? This will depend on the class of processes. For stochastic differential equations, λ is the drift (Mogensen et al., 2018). For point processes, it is the conditional intensity. We give a detailed point process example in Appendix A. The λ-process should essentially describe how the immediate evolution of the multivariate process depends on the past. If so, B is locally independent of A given C if predicting the immediate future of B can be done equally well using the past of process C only or the past of processes A and C.

1.2 Granger Causality

Granger causality (Granger, 1969) is at times treated with some suspicion as it is said to not be 'true' causality. In this paper, we only use Granger causality as an independence relation, analogously to how conditional independence is used in causal models of random vectors. In this way, tests of Granger causality can help us identify certain features of the underlying causal graph, and *(conditional) Granger independence* would in fact be a better term for our usage of Granger causality. In this context, $X = (X_t^1, \dots, X_t^n)$ is a multivariate time series, that is, a stochastic process in discrete time. We let $X_{<t}^D$ denote the set $\{X_s^\alpha : s < t, \alpha \in D\}$.

Definition 1.2 (Granger causality)**.** *Let $A, B, C \subseteq V$. We say that X^B is* Granger-noncausal *for X^A given X^C if for all t and all $\beta \in B$,*

$$X_t^\beta \perp\!\!\!\perp X_{<t}^A \mid X_{<t}^C$$

where $\cdot \perp\!\!\!\perp \cdot \mid \cdot$ denotes conditional independence.

Example 1.3 (VAR)**.** *As an example of a time series model, we consider a* vector-autoregressive process *of order 1. For each t, we have*

$$X_t = AX_{t-1} + \varepsilon_t$$

such that (ε_t) is a sequence of independent random vectors. Moreover, the entries of ε_t are independent. In this case, the zeroes of the $n \times n$ matrix A encode which variables at time $t-1$ directly influence the variables at time t. We can construct an intuitive graphical representation with nodes $V = \{1, 2, \dots, n\}$ by including the edge $\alpha \to \beta$ if and only if $A_{\beta\alpha} \neq 0$. Assume now that $n = 4$, $V = \{1, 2, 3, 4\}$, and

$$A = \begin{bmatrix} a_{11} & 0 & 0 & 0 \\ a_{21} & a_{22} & a_{23} & a_{2n} \\ 0 & 0 & a_{23} & 0 \\ 0 & a_{42} & a_{43} & a_{44} \end{bmatrix}$$

where the entries of A are nonzero if not indicated as zero in the above equation. The corresponding graph is in Figure 1.

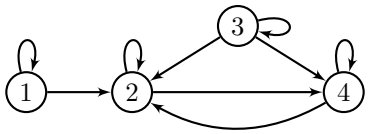

Figure 1: Graph from Example 1.3. In this graph, 4 is μ-separated from 1 given $\{2, 3, 4\}$. Under the global Markov property, this implies that 1 is Granger noncausal for 4 given $\{2, 3, 4\}$. This means that we are able to predict the present of variable 4, X_t^4, equally well using the past of processes $\{2, 3, 4\}$, $X_{<t}^{\{2,3,4\}}$, and using the past of processes $\{1, 2, 3, 4\}$, $X_{<t}^{\{1,2,3,4\}}$. That is, conditionally on the past of $\{2, 3, 4\}$, the past of process 1 does not add any information on the present value of 4. On the other hand, 4 is not μ-separated from 1 given $\{2, 4\}$ as the path $1 \to 2 \leftarrow 3 \to 4$ is μ-connecting.

1.3 Graph Prerequisites

A graph is an ordered pair (V, E) where V is a finite set of *nodes* (also known as *vertices*) and E is a finite set of *edges*. In this paper, we will mostly consider *directed graphs* in which E can be thought of as a subset of $V \times V$. For $\alpha, \beta \in V$, the edge $\alpha \to \beta$ is in the graph if $(\alpha, \beta) \in E$. We always include all *self-edges*, i.e., edges $\alpha \to \alpha$ for $\alpha \in V$.

For graphs $\mathcal{D}_1 = (V, E_1)$ and $\mathcal{D}_2 = (V, E_2)$, we say that $\mathcal{D}_1$ is a *(proper) subgraph* of $\mathcal{D}_2$ if $E_1 \subseteq E_2$ ($E_1 \subsetneq E_2$), and we denote this by $\mathcal{D}_1 \subseteq \mathcal{D}_2$ ($\mathcal{D}_1 \subsetneq \mathcal{D}_2$). We also say that $\mathcal{D}_2$ is a *(proper) supergraph* of $\mathcal{D}_1$. A *walk*, ω, is an ordered, alternating sequence of nodes and edges, $\alpha_1, e_1, \alpha_2, \dots, \alpha_n, e_l, \alpha_{l+1}$, such that each edge is between its adjacent nodes. The *length* of the walk ω is l. A *path* is a walk such that no node is repeated. For nodes $\alpha, \beta \in V$, we say that a walk from α to β is *directed* if every edge points towards β, $\alpha \to \dots \to \beta$. If there exists a directed walk from α to β, we say that α is an *ancestor* of β. We let $\mathrm{an}_\mathcal{D}(\beta)$ denote the set of ancestors of β, and we let $\mathrm{an}_\mathcal{D}(B) = \cup_{\beta \in B} \mathrm{an}_\mathcal{D}(\beta)$. By convention, we say that a *trivial walk* (a walk with

no edges) is directed and therefore $B \subseteq \text{an}_{\mathcal{D}}(B)$. The *complete graph* on nodes V is the graph (V, E) such that $(\alpha, \beta) \in E$ for all α and β such that $\alpha \neq \beta$.

We will use *μ-separation* to encode local independence or Granger noncausality. This is analogous to how d-separation in DAGs may encode conditional independence. We use the convention that endpoint nodes of a walk are neither colliders nor noncolliders. Additional introduction to the relevant graph theory can be found in Mogensen and Hansen (2020).

Definition 1.4 (μ-separation, Mogensen et al. (2018); Mogensen and Hansen (2020))**.** *Let $\mathcal{D} = (V, E)$ be a graph, let $\alpha, \beta \in V$, and let $A, B, C \subseteq V$. We say that a nontrivial walk between α and β is* μ-connecting *from α to β given C if $\alpha \notin C$, all colliders are in $an_{\mathcal{D}}(C)$, no noncolliders are in C, and the final edge has a head at β. We say that B is μ-separated from A given C if there is no μ-connecting walk from any node $\alpha \in A$ to any node $\beta \in B$ given C.*

The notion of μ-separation is a generalization of δ-separation (Didelez, 2000, 2008).

1.4 Independence Models

In this paper, we will use an abstract *independence model*, $\mathcal{I}$, which is simply a set of triples, (A, B, C), $A, B, C \subseteq V$, and we say that this is an independence model *over* V. Such an independence model may represent the local independences that hold in a multivariate, continuous-time stochastic process or the conditional Granger-noncausalities that hold in a discrete-time stochastic process, i.e., $(A, B, C) \in \mathcal{I}$ if and only if B is locally independent of A given C, for example. Using an abstract independence model, there is no need to distinguish between independence models representing local independences and independence models representing Granger noncausalities. In the remainder of the paper, we will often refer to both types of independences as simply 'local independences'.

For a graph $\mathcal{D}$, we define $\mathcal{I}(\mathcal{D})$ as the set of triples (A, B, C) such that B is μ-separated from A given C in $\mathcal{D}$. *Markov and faithfulness properties* describe how $\mathcal{I}$ and $\mathcal{I}(\mathcal{D})$ are related.

1.5 Markov Properties and Faithfulness

Markov properties describe how graphs encode independence by relating properties of a graph, $\mathcal{D}$, to an independence model, $\mathcal{I}$. We use the notation $\alpha \to_{\mathcal{D}} \beta$ to indicate that the edge $\alpha \to \beta$ is in $\mathcal{D}$.

Definition 1.5 (Pairwise Markov property)**.** *We say that $\mathcal{I}$ satisfies the* pairwise Markov property *with respect to $\mathcal{D}$ if for all $\alpha, \beta \in V$*

$$\alpha \not\to_{\mathcal{D}} \beta \Rightarrow (\alpha, \beta, V \setminus \{\alpha\}) \in \mathcal{I}.$$

Definition 1.6 (Global Markov property)**.** *We say that $\mathcal{I}$ satisfies the* global Markov property *with respect to $\mathcal{D}$, or simply that $\mathcal{I}$ is* Markov *with respect to $\mathcal{D}$, if for all $A, B, C \subseteq V$,*

$$(A, B, C) \in \mathcal{I}(\mathcal{D}) \Rightarrow (A, B, C) \in \mathcal{I}.$$

The global Markov property may also be written as $\mathcal{I}(\mathcal{D}) \subseteq \mathcal{I}$.

The global and pairwise Markov properties are equivalent under fairly general assumptions, see, e.g., Didelez (2000, 2008); Eichler (2012); Mogensen et al. (2018) for related results in different model classes. Some of these results restrict the sets A,B, and C, e.g., such that $B \subseteq C$ in our notation.

Definition 1.7 (Faithfulness)**.** *We say that $\mathcal{I}$ is* faithful *with respect to $\mathcal{D}$ if for all $A, B, C \subseteq V$,*

$$(A, B, C) \in \mathcal{I} \Rightarrow (A, B, C) \in \mathcal{I}(\mathcal{D}),$$

that is, if $\mathcal{I} \subseteq \mathcal{I}(\mathcal{D})$.

Note that, in our terminology, faithfulness corresponds to the statement $\mathcal{I} \subseteq \mathcal{I}(\mathcal{G})$, not to the stronger statement $\mathcal{I} = \mathcal{I}(\mathcal{G})$.

Example 1.8. *We consider the example graph, $\mathcal{D}$, shown in Figure 2, and we assume that the independence model $\mathcal{I}$ satisfies the global Markov property with respect to $\mathcal{D}$, i.e., $\mathcal{I}(\mathcal{D}) \subseteq \mathcal{I}$.*

We consider the walk $1 \to 2 \to 4$ and sets $A = \{1\}$, $B = \{4\}$, $C = \{4\}$. This walk is between $1 \in A \setminus C = \{1\}$ and $4 \in B$, all colliders are in $an_{\mathcal{D}}(C)$ (there are none), no noncolliders are in C (endpoint nodes are not colliders), and the walk has a head at 4. This means that $\{4\}$ is not μ-separated from $\{1\}$ given $\{4\}$ in the graph $\mathcal{D}$ shown in Figure 2, i.e., $(\{1\}, \{4\}, \{4\}) \notin \mathcal{I}(\mathcal{D})$. If $(\{1\}, \{4\}, \{4\}) \in \mathcal{I}$, then $\mathcal{I}$ is not faithful with respect to $\mathcal{D}$.

1.6 Structure Learning

There is a large literature on structure learning from multivariate stochastic processes, often assuming *causal sufficiency*, i.e., that every relevant coordinate process is observed, and assuming some specific parametric or semiparametric class of stochastic processes.

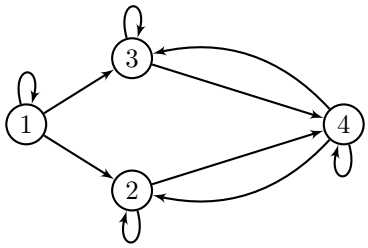

Figure 2: Graph from Example 1.8.

We will also make the assumption of causal sufficiency in this paper, however, we will take a nonparametric approach. For parametric model classes, and assuming causal sufficiency, one may also, e.g., use methods that are specific to the model class to learn a causal graph from data. Our approach is completely nonparametric in that it only uses tests of local independence. Examples A.1 and 1.3 are therefore mostly meant as an illustration.

In the next section, we give a characterization of faithfulness which allows us to construct faithful representations of local independence models.

1.6.1 Causal Interpretation

Structure learning is often done from a causal perspective. The causal interpretation will also depend on the model class. In this paper, we assume that $\mathcal{D}$ is a *causal graph* which summarizes the cause-effect relations between the coordinate processes of the system. The exact meaning of this is discussed by, e.g., Eichler and Didelez (2007); Røysland et al. (2023).

2 TRANSITIVITY CONDITIONS

We define a set of *transitivity conditions*.

Definition 2.1 (Transitivity conditions)**.** *Let $\mathcal{D} = (V, E)$ and let $\mathcal{I}$ be an independence model over V. Let $C \subseteq V$. We say that $\mathcal{I}$ is C-transitive with respect to $\mathcal{D}$ if for each edge $\alpha \to \beta$ in $\mathcal{D}$, conditions D0-D3 hold.*

D0 if $\alpha \notin C$, then $(\alpha, \beta, C) \notin \mathcal{I}$,

D1 if $\alpha \notin C$, then for all γ:
$(\gamma, \alpha, C) \notin \mathcal{I}(\mathcal{D}) \Rightarrow (\gamma, \beta, C) \notin \mathcal{I}$,

D2 if $\alpha \notin C, \beta \in C$, then for all γ, δ:
$(\gamma, \beta, C) \notin \mathcal{I}(\mathcal{D}), (\alpha, \delta, C) \notin \mathcal{I}(\mathcal{D}) \Rightarrow (\gamma, \delta, C) \notin \mathcal{I}$,

D3 if $\alpha \notin C$, then for all γ:
$(\alpha, \gamma, C) \notin \mathcal{I}(\mathcal{D}) \Rightarrow (\beta, \gamma, C) \notin \mathcal{I}$.

We say that an independence model $\mathcal{I}$ is transitively closed *with respect to a graph $\mathcal{D}$ if $\mathcal{I}$ is C-transitive with respect to $\mathcal{D}$ for all $C \subseteq V$.*

A simpler version of the conditions in Definition 2.1 are also found in Mogensen and Hansen (2020) where the authors used them to prove that every Markov equivalence class of partially observed local independence graphs have a greatest element. We can recover their version by using $\mathcal{I} = \mathcal{I}(\mathcal{D})$ in the above definition. For our result, the generalization is important as it connects an arbitrary independence model, $\mathcal{I}$, to a graphical representation, $\mathcal{D}$.

The conditions in Definition 2.1 are in a certain sense rewriting the definition of μ-separation. This has three purposes. First, this assigns faithfulness violations to specific edges that can be removed to obtain faithful representations (Subsection F.2). Second, it allows us to construct a (nontrivial) faithful representation directly from the independence model (Section 3). Third, we will reformulate these conditions slightly to see that they correspond to different notions of faithfulness (Section 4).

Proposition 2.2. *The independence model $\mathcal{I}(\mathcal{D})$ is transitively closed with respect to the graph $\mathcal{D}$.*

Proposition 2.3. *Let $\mathcal{I}_1 \subseteq \mathcal{I}_2$. If $\mathcal{I}_2$ is transitively closed with respect to $\mathcal{D}$, then $\mathcal{I}_1$ is transitively closed with respect to $\mathcal{D}$.*

3 CHARACTERIZATION OF FAITHFULNESS

Definition 2.1 gives a characterization of faithfulness as described in the next theorem.

Theorem 3.1. *An independence model $\mathcal{I}$ is transitively closed with respect to a graph $\mathcal{D}$ if and only if $\mathcal{I}$ is faithful with respect to $\mathcal{D}$.*

The above characterizes the set of graphs, $\mathcal{D}$, that are faithful with respect to the independence model $\mathcal{I}$. However, the conditions in Definition 2.1 use both $\mathcal{I}$ and $\mathcal{I}(\mathcal{D})$, and it is therefore not immediately clear how to construct these graphs if we only have access to the independence model $\mathcal{I}$. The next definition defines a graph from an independence model, $\mathcal{I}$, only, and Theorem 3.3 proves that $\mathcal{I}$ in fact is faithful with respect to the graph that we obtain from the definition.

Definition 3.2 (Edge-transitive graph)**.** *Let $\mathcal{I}$ be an independence model over V. We define a graph $\mathcal{F}_I = (V, E_I)$ by including the edge $\alpha \to \beta$, $\alpha, \beta \in V$, $\alpha \neq \beta$, if and only if E0-E3 hold for all C.*

E0 if $\alpha \notin C$, then $(\alpha, \beta, C) \notin \mathcal{I}$,

E1 *if* $\alpha \notin C$, *then for all* γ:
$(\gamma, \alpha, C) \notin \mathcal{I} \Rightarrow (\gamma, \beta, C) \notin \mathcal{I}$,

E2 *if* $\alpha \notin C, \beta \in C$, *then for all* γ, δ:
$(\gamma, \beta, C) \notin \mathcal{I}, (\alpha, \delta, C) \notin \mathcal{I} \Rightarrow (\gamma, \delta, C) \notin \mathcal{I}$,

E3 *if* $\alpha \notin C$, *then for all* γ:
$(\alpha, \gamma, C) \notin \mathcal{I} \Rightarrow (\beta, \gamma, C) \notin \mathcal{I}$.

For an independence model, $\mathcal{I}$, *we say that* $\mathcal{F}_I$ *defined above is the* edge-transitive graph *corresponding to* $\mathcal{I}$.

Theorem 3.3. *The independence model* $\mathcal{I}$ *is faithful with respect to* $\mathcal{F}_I$.

We say that an independence model, $\mathcal{I}$, is *graphical* if there exists a graph $\mathcal{D}$ such that $\mathcal{I} = \mathcal{I}(\mathcal{D})$. The following proposition simply states that if the independence model is Markov and faithful with respect to a graph, i.e., is graphical, then $\mathcal{F}_I$ as defined in Definition 3.2 is equal to $\mathcal{D}$.

Proposition 3.4. *Assume* $\mathcal{I}$ *is graphical, that is,* $\mathcal{I} = \mathcal{I}(\mathcal{D})$ *for a graph* $\mathcal{D}$. *In this case,* $\mathcal{F}_I = \mathcal{D}$.

Any independence model is faithful with respect to the empty graph, and it is useful to introduce the concept of *maximal faithfulness*. We say that $\mathcal{I}$ is *maximally faithful* with respect to $\mathcal{D}$ if it is faithful with respect to $\mathcal{D}$ and it is not faithful with respect to any proper supergraph of $\mathcal{D}$.

4 WEAKER NOTIONS OF FAITHFULNESS

The Markov condition holds under fairly general assumptions, however, some version of a faithfulness-like assumption is needed for structure learning. It is possible to define such notions that are weaker than faithfulness, yet useful in the context of structure learning. In DAG-based models, Zhang and Spirtes (2008); Ramsey et al. (2006) study such notions, and Lam et al. (2022); Andrews et al. (2023) discuss how to weaken the faithfulness assumption in permutation algorithms for causal discovery. In local independence models, Mogensen (2020a) gives the following definition.

Definition 4.1 (Ancestor faithfulness, Mogensen (2020a))**.** *We say that* $\mathcal{I}$ *is* ancestor faithful *with respect to* $\mathcal{D}$ *if, for all* A, B, *and* C, *the existence of a directed and* μ*-connecting path from* $\alpha \in A$ *to* $\beta \in B$ *given* C *implies* $(A, B, C) \notin \mathcal{I}$.

The following is a weaker notion than that of ancestor faithfulness.

Definition 4.2 (Parent faithfulness)**.** *We say that* $\mathcal{I}$ *is* parent faithful *with respect to* $\mathcal{D}$ *if, for all* A, B, *and* C, *the existence of a directed edge* $\alpha \to \beta$ *such that* $\alpha \in A \setminus C$ *and* $\beta \in B$ *implies* $(A, B, C) \notin \mathcal{I}$.

We say that β is *inseparable* from α if there is no $C \subseteq V \setminus \{\alpha\}$ such that $(\alpha, \beta, C) \in \mathcal{I}(\mathcal{D})$. Parent faithfulness can be seen as an analogue of adjacency faithfulness in DAG-based models: In a DAG, nodes are inseparable if and only if they are adjacent. In a local independence graph, a node β is inseparable from a node α if and only if the edge $\alpha \to \beta$ is in the graph. One should note that the notion of inseparability is symmetric in DAGs, but asymmetric in local independence graphs. This means that in a local independence graph, α need not be inseparable from β even if β is inseparable from α.

In cases where faithfulness is not violated, there may still be near-violations of faithfulness. In this setting, learning methods that only assume weaker notions of faithfulness may show better performance (Ramsey et al., 2006; Zhalama et al., 2017).

We define an even weaker faithfulness-like assumption.

Definition 4.3 (Parent dependence)**.** *We say that* $\mathcal{I}$ *satisfies* parent dependence *with respect to* $\mathcal{D}$, *if* $\alpha \to \beta$ *implies* $(\alpha, \beta, \beta) \notin \mathcal{I}$ *for all* $\alpha \neq \beta$.

Example 4.4. *This example is a continuation of Example 1.8. We consider again the graph* $\mathcal{D}$ *in Figure 2, and we let* $\mathcal{I} = \mathcal{I}(\mathcal{D}) \cup \{(\{1\}, \{4\}, \emptyset), (\{1\}, \{4\}, \{4\})\}$. *As argued in Example 1.8,* $\mathcal{I}$ *is not faithful with respect to* $\mathcal{D}$ *as the walk* $1 \to 2 \to 4$ *is* μ*-connecting from* 1 *to* 4 *given* $\{4\}$.

The walk $1 \to 2 \to 4$ *in Figure 2 is a directed path and it is* μ*-connecting from* 1 *to* 4 *given* $\{4\}$. *This means that* $\mathcal{I}$ *is not ancestor faithful with respect to* $\mathcal{D}$.

On the other hand, we see that $\mathcal{I}$ *is parent faithful with respect to* $\mathcal{D}$. *Let* $A, B, C \subseteq V$, *and assume that* $\alpha \to \beta$ *where* $\alpha \in A \setminus C$ *and* $\beta \in B$. *In this case,* B *is not* μ*-separated from* A *given* C. *Moreover,* $A = \{1\}$ *and* $B = \{4\}$ *cannot both be true as* 1 *is not a parent of* 4.

4.1 Causal Minimality

Faithfulness, and similar assumptions, are common for structure learning. In the context, of local independence there is a far weaker notion which is in fact sufficient for structure learning.

The concept of a maximally faithful graph is essentially dual to the concept of *causal minimality*. We say that $\mathcal{I}$ is *causally minimal* with respect to $\mathcal{D}$ if it is Markov with respect to $\mathcal{D}$ and there is no proper subgraph of $\mathcal{D}$, $\mathcal{D}'$, such that $\mathcal{I}$ is Markov with respect to $\mathcal{D}'$. (Peters et al., 2017). In symbols, $\mathcal{I}(\mathcal{D}) \subseteq \mathcal{I}$ and there is no $\mathcal{D}_0 \subsetneq \mathcal{D}$ such that $\mathcal{I}(\mathcal{D}_0) \subseteq \mathcal{I}$. Causal minimality is also known as *minimal Markovness* (Sadeghi,

2017).

Proposition 4.5. *Let $\mathcal{I}$ be an independence model and $\mathcal{D}$ be a graph. If $\mathcal{I}$ is faithful with respect to $\mathcal{D}$, then it is ancestor faithful with respect to $\mathcal{D}$. If $\mathcal{I}$ is ancestor faithful with respect to $\mathcal{D}$, then it is parent faithful with respect to $\mathcal{D}$. If $\mathcal{I}$ is parent faithful and Markov with respect to $\mathcal{D}$, then $\mathcal{I}$ is causally minimal with respect to $\mathcal{D}$.*

The above describes a hierarchy of faithfulness assumptions for local independence graphs. Local independence is not symmetric, i.e., $(A, B, C) \in \mathcal{I}$ does not imply $(B, A, C) \in \mathcal{I}$, and therefore μ-separation is also not symmetric. On the other hand, Lam (2023) describes a hierarchy of faithfulness assumptions in conditional independence models, and that hierarchy is different from the above. This is quite natural as conditional independence and d-separation are symmetric ternary independence relations.

Definition 3.2 allows us to construct a faithful graph from an independence model (Theorem 3.3). We can also directly construct a causally minimal graph. For an independence model, $\mathcal{I}$, we define a graph, $\mathcal{D}_I$, such that $\alpha \to \beta$ is in $\mathcal{D}_I$ if and only if $(\alpha, \beta, V \setminus \{\alpha\}) \notin \mathcal{I}$ and we say that $\mathcal{D}_I$ is the *induced local independence graph* corresponding to $\mathcal{I}$.

Proposition 4.6 (Mogensen (2020b))**.** *Assume equivalence of pairwise and global Markov properties. The induced local independence graph corresponding to $\mathcal{I}$, $\mathcal{D}_I$, is causally minimal with respect to $\mathcal{I}$.*

Proposition 4.7. *The graph $\mathcal{D}_I$ is the only causally minimal graph with respect to $\mathcal{I}$.*

In other words, assuming the equivalence of pairwise and global Markov properties, $\mathcal{I}$ is causally minimal with respect to $\mathcal{D}$ if and only if $\alpha \to \beta$ is in $\mathcal{D}$ exactly when $(\alpha, \beta, V \setminus \{\alpha\}) \notin \mathcal{I}$.

Theorem C.1 argues that violations of faithfulness are detectable with infinite data under Markov and causal minimality assumptions on $\mathcal{D}$ and $\mathcal{I}$ in the sense that there is no other graph, $\mathcal{D}'$, such that $\mathcal{I}$ is Markov and faithful with respect to $\mathcal{D}'$ (see Appendix C for further explanation).

The next proposition uses *asymmetric graphoid properties* that hold in local independence models, see, e.g., Didelez (2006); Mogensen et al. (2018) and Appendix B.

Proposition 4.8. *Assume $\mathcal{I}$ is causally minimal with respect to $\mathcal{D}$. If $\alpha \notin pa_{\mathcal{D}}(\beta)$, $\alpha \neq \beta$, and $pa_{\mathcal{D}}(\beta) \subseteq C$, then $(\alpha, \beta, C) \in \mathcal{I}$. Assume that $\mathcal{I}$ satisfies left weak union, left decomposition, and left contraction, that we have equivalence of pairwise and global Markov properties, $\alpha \in pa_{\mathcal{D}}(\beta)$, $\alpha \neq \beta$, and that $pa_{\mathcal{D}}(\beta) \setminus \{\alpha\} \subseteq C$. In this case, $(\alpha, \beta, C) \notin \mathcal{I}$.*

4.2 Hierarchy of Faithfulness Assumptions

In this subsection, we rewrite the conditions in Definition 2.1 to illustrate how they correspond to different faithfulness assumptions. We first define the notion of *trek faithfulness*.

Definition 4.9 (Trek faithfulness)**.** *We say that a walk is a* trek *if it has no colliders. We say that $\mathcal{I}$ is* trek faithful *with respect to $\mathcal{D}$, if for all A, B, and C, the existence of a μ-connecting trek from A to B given C implies $(A, B, C) \notin \mathcal{I}$.*

It is immediate that faithfulness implies trek faithfulness, and that trek faithfulness implies ancestor faithfulness, noting that a directed walk is also a trek.

Lemma 4.10 reformulates the conditions from Definition 2.1 to provide an equivalent set of conditions. These conditions correspond to the hierarchical nature of the faithfulness conditions: D0 is equivalent to parent faithfulness, the combination of D0 and D1' is equivalent to ancestor faithfulness, and the combination of D0, D1', and D3' is equivalent to trek faithfulness. The combination of D0, D1', D2, and D3' is equivalent to faithfulness. This is the content of Theorem 4.11.

Lemma 4.10. *For an edge $\alpha \to \beta$ and a set C, we define the following conditions.*

D1' If there is a directed path which is μ-connecting from γ to α given C in $\mathcal{D}$, then $(\gamma, \beta, C) \notin \mathcal{I}$.

D3' If there is a trek which is μ-connecting from α to γ given C in $\mathcal{D}$, then $(\beta, \gamma, C) \notin \mathcal{I}$.

An independence model $\mathcal{I}$ and a graph $\mathcal{D}$ satisfy D0, D1, D2, and D3 for every edge in $\mathcal{D}$ and set C if and only if they satisfy D0, D1', D2, and D3' for every edge in $\mathcal{D}$ and set C.

Theorem 4.11. *Let $\mathcal{I}$ be an independence model which satisfies left and right decomposition, and let $\mathcal{D}$ be a graph.*

- *Condition D0 holds for all edges $\alpha \to \beta$ in $\mathcal{D}$ and sets C if and only if $\mathcal{I}$ is parent faithful with respect to $\mathcal{D}$.*
- *Conditions D0 and D1' hold for all edges $\alpha \to \beta$ in $\mathcal{D}$ and sets C if and only if $\mathcal{I}$ is ancestor faithful with respect to $\mathcal{D}$.*
- *Conditions D0, D1', and D3' hold for all edges $\alpha \to \beta$ in $\mathcal{D}$ and sets C if and only if $\mathcal{I}$ is trek faithful with respect to $\mathcal{D}$.*

5 STRUCTURE LEARNING

In graphical structure learning, the task is to recover a graphical representation from tests of local independence. In this section, we describe how the above theory relates to structure learning algorithms. It is common to assume faithfulness in the context of structure learning, see, e.g., Meek (2014); Mogensen et al. (2018); Absar and Zhang (2021) for examples in structure learning based on local independence/Granger noncausality. Mogensen (2020a) uses a weaker notion of faithfulness.

We assume causal sufficiency except in Appendix G where we describe some results assuming only partial observation. As is common in the literature, we will at times assume that we have access to an *independence oracle*, i.e., instead of inputting, e.g., p-values from tests of local independence, our algorithm simply has access to the actual independence model and therefore always gets the right answer to an independence query. This is mostly done to separate algorithmic issues from testing issues. In practical applications of the learning algorithms, the test $(\alpha, \beta, C) \in \mathcal{I}$ is replaced by a p-value and a significance threshold.

5.1 Comparison with DAG-based Models

There is a large literature on learning causal graphs based on tests of conditional independence (see Spirtes and Zhang (2018) and references therein). One example of an algorithm is the PC-algorithm (Spirtes et al., 2001). In the adjacency phase of this algorithm, larger and larger conditioning sets are used to look for separating sets. One motivation is to use tests with small conditioning sets to achieve larger power of the statistical tests (Spirtes and Zhang, 2018). Meek (2014) and Absar and Zhang (2021) proceed by checking larger and larger sets of potential separating sets and remove an edge when one is found, essentially using this basic idea of the PC-algorithm in the stochastic process-setting. However, there are a number of important differences between constraint-based learning in DAG-based models and constraint-based learning in stochastic process models. First, in the case, of DAG-based model several graphs may encode the same conditional independences. On the other hand, for fully observed stochastic processes and under quite general assumptions the causal graph is actually identified from the local independence model (see Section 5.4) in the sense that the Markov equivalence class of a directed graph is a singleton when using μ-separation. In the case of partial observation, this is no longer true and Mogensen and Hansen (2020) characterize the Markov equivalence classes of *directed mixed graphs* that represent partially observed local independence graphs. Second, the set $V \setminus \{\alpha\}$ μ-separates β from α if and only if $\alpha \to \beta$ is not in the graph. For this, we do not need to know the graph and essentially this means that we can construct a separating set, if one exists, without any knowledge of the graph.

5.2 The CA-algorithm

We briefly describe the CA-algorithm from Meek (2014). This is also similar to the algorithm in Absar and Zhang (2021). In this algorithm, for each ordered pair (α, β), larger and larger conditioning sets are tried to find a separating set, i.e., a set, C, such that $(\alpha, \beta, C) \in \mathcal{I}$. This is similar to the classical PC-algorithm for DAGs (Spirtes et al., 2001). The details of the algorithm can be found in Meek (2014).

5.3 The CS-algorithm

The CS-algorithm (*causal screening*) was introduced in Mogensen (2020a) as a fast screening approach for partially observed systems. In its first step, it tests $(\alpha, \beta, \beta) \in \mathcal{I}$ for all ordered pairs (α, β). In its second step, it tests $(\alpha, \beta, \mathrm{pa}_{\mathcal{D}_1}(\beta) \setminus \{\alpha\}) \in \mathcal{I}$ when $\alpha \to \beta$ is in $\mathcal{D}_1$ where $\mathcal{D}_1$ is a subgraph of the output from the first step. The idea is to use a superset of the actual parent set of β as a conditioning set.

Proposition 5.1. *In the oracle case, the CS-algorithm (Algorithm 1) outputs the true graph under Markov and parent faithfulness assumptions.*

Proof. If $\alpha \to \beta$ is not in the true graph, then it is also not in the output (Proposition 2 in the supplementary material of (Mogensen, 2020a)). If is in the true graph, then parent faithfulness implies that it is also in the output. □

Proposition 5.2. *Assume causal sufficiency, left weak union, left decomposition, and left contraction of $\mathcal{I}$, and equivalence of pairwise and global Markov properties. If $\mathcal{I}$ is causally minimal with respect to $\mathcal{D}$ and satisfies parent dependence with respect to $\mathcal{D}$, then causal screening outputs $\mathcal{D}$ in the oracle setting.*

Many other algorithms will only be correct in the oracle case under stronger assumptions, one reason being that they test more 'small' sets which may lead to a faulty edge removal due to a violation of faithfulness.

5.4 Learning with Minimal Assumptions

If we take the Markov property for granted, the four conditions outlined above, faithfulness, ancestor faithfulness, parent faithfulness, and causal minimality, are in this list ordered from strongest to weakest. In this subsection, we assume the weakest condition, that of

Algorithm 1: Causal screening algorithm (CS)

input : $\mathcal{I}$ over V such that $|V| = n$
for $\beta \in V$ **do**
  for $\alpha \in V \setminus \{\alpha\}$ **do**
    if $(\alpha, \beta, \beta) \in \mathcal{I}$ **then**
      $E \leftarrow E \setminus \{\alpha \to \beta\}$;
      $\mathcal{D} \leftarrow (V, E)$;
    end
  end
end
for $\beta \in V$ **do**
  for $\alpha \in pa_{\mathcal{D}}(\beta)$ **do**
    if $(\alpha, \beta, pa_{\mathcal{D}}(\beta) \setminus \{\alpha\}) \in \mathcal{I}$ **then**
      $E \leftarrow E \setminus \{\alpha \to \beta\}$;
      $\mathcal{D} \leftarrow (V, E)$;
    end
  end
end
$\mathcal{D}_{cs} \leftarrow \mathcal{D}$;
output: $\mathcal{D}_{cs}$

causal minimality, to discuss how structure learning can be achieved with this minimal assumption.

Under Markov and causal minimality assumptions, the induced local independence graph, $\mathcal{D}_I$, equals the true graph (in the oracle case): Let $\mathcal{D}$ be the true graph such that $\mathcal{D}$ and $\mathcal{I}$ are causally minimal. The graph $\mathcal{D}_I$ satisfies the pairwise Markov property by definition. Under equivalence of pairwise and global Markov properties, we have that $\mathcal{D}_I$ is Markov with respect to $\mathcal{D}$. If e is in $\mathcal{D}_I$, then $(\alpha, \beta, V \setminus \{\alpha\}) \notin \mathcal{I}$. Using Markovness, e must be in $\mathcal{D}$ as well, so $\mathcal{D}_I \subseteq \mathcal{D}$. If e is not in $\mathcal{D}_I$, then e also not in $\mathcal{D}$ due to causal minimality. The CM-algorithm (Causal Minimality) is the algorithm which outputs a graph, $\mathcal{D}_{cm}$, such that $\alpha \to_{\mathcal{D}_{cm}} \beta$ if and only if $(\alpha, \beta, V \setminus \{\alpha\}) \notin \mathcal{I}$.

However, the above may not be practical if there are many coordinates processes as this may require very large conditioning sets, $V \setminus \{\alpha\}$, and therefore tests with poor performance. On the other hand, it avoids many tests, all of which have a risk of faithfulness violations, or near-violations, which leads to worse output. These observations may motivate the use of the CS-algorithm in Subsection 5.3.

If we instead test all subsets, and include $\alpha \to \beta$ if and only if there is no separating set, we obtain a subgraph of the true graph, only assuming Markovness (see details in Algorithm 2 in Appendix E). This algorithm is a local independence version of the SGS algorithm (Spirtes et al., 2001). If we assume that there are at most k parents, we test all subsets of size at most k. This also returns a subgraph of the true graph under the Markov assumption. Appendix E defines this algorithm and gives states this result formally.

5.5 Learning and Faithfulness

The previous section describes general structure learning algorithms. Appendix F connects structure learning with the faithfulness results in the previous sections. In Subsection F.1, we consider the *edge-transitive graph*. In Subsection F.2, we argue that one may trim the output of a learning algorithm to obtain a faithful representation.

6 NUMERICAL EXAMPLES

We compare the algorithms to investigate their properties when using actual data.[1] We repeatedly generated a true graph, $\mathcal{D} = (V, E)$, and observations from a corresponding VAR(1)-process. Using tests of Granger causality, we computed an output graph, $\mathcal{D}_a = (V, E_a)$ for each algorithm a. We computed the *surplus edges*, $E_a \setminus E$, the *missing edges* $E \setminus E_a$, and the *difference*, $(E_a \setminus E) \cup (E \setminus E_a)$ between $\mathcal{D}$ and $\mathcal{D}_a$. In Figure 3, we report the mean number of surplus edges, the mean number of missing edges, and the mean of $|(E_a \setminus E) \cup (E \setminus E_a)|$ for each algorithm a, and for different values of significance threshold and $n = |V|$. More details are in Section H.

The dSGS algorithm is seen to have the lowest number of surplus edges which is not surprising as it tests every possible set and removes the edge if any test is nonsignificant. We also know from Proposition E.1 that, in the oracle case, it outputs a subgraph of the true graph under minimal assumptions.

We observe that the CM-algorithm, simply using a single test for each ordered pair (α, β) does surprisingly well, e.g., in comparison with the CA-algorithm. An important point is the fact that using more tests increase the risk of making errors due to faithfulness violations or near-violations. It is essential that, in this context, the set $V \setminus \{\alpha\}$ always separated β from α if such separation is possible, and this can be tested with no prior knowledge of the graph. For this reason, testing smaller conditioning sets is not needed, at least for n of moderate size.

A key weakness of the CM-algorithm is, of course, the fact that it uses large conditioning sets for large values of n and such tests are expected to have low power. As a remedy, one may use the CS-algorithm which tries to reduce the set of potential parents. We see that the CS- and CM-algorithms have similar perfor-

[1] Code can be found at `https://github.com/soerenwengel/faithfulLIGsCode`

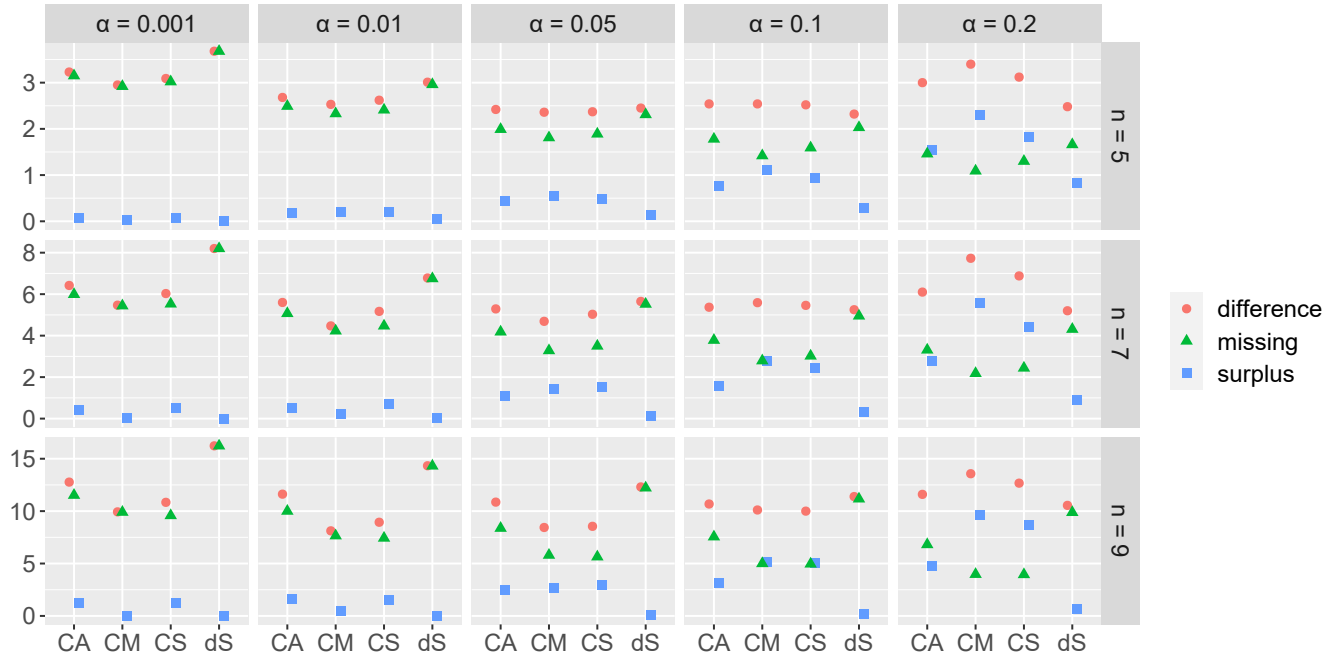

Figure 3: Comparison of algorithms. Points indicate mean over $M = 100$ repetitions. Red circles indicate mean difference between true graph and output graph. Green triangles indicate mean number of surplus edges, and blue squares indicate mean number of missing edges. Section 6 explains this experiment in more detail.

mances, and the CS-algorithm may be viable alternative for large values of n.

Appendix G provides additional results in the case of partial observation.

7 DISCUSSION

Constraint-based learning in stochastic processes is still lacking some of the tools that are available for constraint-based learning from random vectors. This paper studies notions of faithfulness and discusses differences between the two frameworks, e.g., the fact that starting from small conditioning sets may not always be preferable in the stochastic process-setting.

The use of score-based methods, or methods that aggregate the information across edges, is an interesting topic for future research.

Acknowledgments

The author was supported by the Independent Research Fund Denmark (DFF-International Postdoctoral Grant 0164-00023B). The author is a member of the ELLIIT Strategic Research Area at Lund University. The author is grateful to the reviewers for their comments and suggestions.

References

Odd O. Aalen. Dynamic modelling and causality. *Scandinavian Actuarial Journal*, 1987(3-4):177–190, 1987.

Saima Absar and Lu Zhang. Discovering time-invariant causal structure from temporal data. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2807–2811, 2021.

Bryan Andrews, Joseph Ramsey, Ruben Sanchez Romero, Jazmin Camchong, and Erich Kummerfeld. Fast scalable and accurate discovery of DAGs using the best order score search and grow shrink trees. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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

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

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

Michael Eichler. Graphical modelling of multivariate time series. *Probability Theory and Related Fields*, 153(1):233–268, 2012.

Michael Eichler and Vanessa Didelez. Causal reasoning in graphical time series models. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence (UAI)*, 2007.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

Wai-yin Lam. Causal razors: A comparative study of simplicity assumptions in causal discovery. 2023.

Wai-yin Lam, Bryan Andrews, and Joseph Ramsey. Greedy relaxations of the sparsest permutation algorithm. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2022.

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

Christopher Meek. Toward learning graphical and causal process models. In *CI at UAI*, pages 43–48, 2014.

Søren Wengel Mogensen. Causal screening in dynamical systems. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2020a.

Søren Wengel Mogensen. *Graphical modeling in dynamical systems*. PhD thesis, University of Copenhagen, 2020b.

Søren Wengel Mogensen. Weak equivalence of local independence graphs. 2023.

Søren Wengel Mogensen and Niels Richard Hansen. Markov equivalence of marginalized local independence graphs. *The Annals of Statistics*, 48(1):539–559, 2020.

Søren Wengel Mogensen, Daniel Malinsky, and Niels Richard Hansen. Causal learning for partially observed stochastic dynamical systems. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, MA: MIT Press, 2017.

Joseph Ramsey, Jiji Zhang, and Peter L Spirtes. Adjacency-faithfulness and conservative causal inference. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI)*, 2006.

Kjetil Røysland, Pål Ryalen, Mari Nygård, and Vanessa Didelez. Graphical criteria for the identification of marginal causal effects in continuous-time survival and event-history analyses. 2023.

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.

Peter Spirtes and Kun Zhang. Search for causal models. In *Handbook of Graphical Models*, pages 439–470. CRC Press, 2018.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, second edition, 2001.

Zhalama, Jiji Zhang, and Wolfgang Mayer. Weakening faithfulness: Some heuristic causal discovery algorithms. *International Journal of Data Science and Analytics*, 3:93–104, 2017.

Jiji Zhang and Peter Spirtes. Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18:239–271, 2008.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes: Sections 1 and 4]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes: Subsection H.2]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes: attached]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes: provided in theorems, propositions, etc]
   (b) Complete proofs of all theoretical results. [Yes: in the main paper, or in the supplementary material]
   (c) Clear explanations of any assumptions. [Yes: e.g., in Section 4]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes: code attached]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes: Sections 6 and H describe this. Code to reproduce the results is provided.]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes: Section 6]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable: runs on a standard laptop]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Not Applicable]

   (b) The license information of the assets, if applicable. [Not Applicable]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Material for Faithful graphical representations of local independence

Søren Wengel Mogensen
Department of Automatic Control, Lund University

A LINEAR HAWKES PROCESSES

Example A.1 (Linear Hawkes process)**.** Linear Hawkes processes *are a class of point processes. A (multivariate) point process,* $X_t = (X_t^1, \dots, X_t^n)$*, consists of a set of events,* (t, α)*, such that* t *is a time point and* $\alpha \in V = \{1, 2, \dots, n\}$ *is a coordinate process. Point processes may be described using the conditional intensity which we will denote* $\lambda_t = (\lambda_t^1, \dots, \lambda_t^n)$*. It holds that*

$$\lambda_t^\beta = \frac{1}{h} \lim_{h \downarrow 0} P(\textit{there is a } \beta\textit{-event in } (t, t+h] \mid \mathcal{F}_t^V),$$

and λ_t^β *can therefore be interpreted as describing how likely it is to observe a* β*-event in the immediate future given the past of the process. A point process is a* linear Hawkes process *if for all* β

$$\lambda_t^\beta = \mu_\beta + \sum_{\alpha \in V} \sum_{(s,\alpha): s < t} f_{\beta\alpha}(t - s)$$

where μ_β *is a nonnegative constant,* $f_{\beta\alpha}$ *is a nonnegative function, and the sum is over all events of type* α *until time* t*. In this example, the* λ*-process in Definition 1.1 can be chosen as the conditional intensity.*

When the function $f_{\beta\alpha}$ *is zero there is no direct dependence of* λ_t^β *on the past of the* α*-process. We construct a graph with nodes* $V = \{1, 2, \dots, n\}$ *such that for* $\alpha, \beta \in V$*, we include* $\alpha \to \beta$ *if and only* $f_{\beta\alpha} \neq 0$*. This graph encodes a set of local independences as described by the global Markov property (Definition 1.6). When the graph is unknown, we can use tests of local independence to learn about the graph. We will say that the graph defined above is the* causal graph*, see also Subsection 1.6.1.*

B ASYMMETRIC GRAPHOIDS

Graphoid properties are often used in the context of graphical models of random variables (Lauritzen, 1996). Analogously, *asymmetric graphoid properties* may be defined (Didelez, 2006; Mogensen et al., 2018). These have *left* and *right* versions as *symmetry*, $(A, B, C) \in \mathcal{I} \Rightarrow (B, A, C) \in \mathcal{I}$, is not assumed.

Definition B.1 (Asymmetric graphoid properties)**.** *Let $\mathcal{I}$ be an independence model over V. We say that $\mathcal{I}$ satisfies* left decomposition *if*

$$(A, B, C) \in \mathcal{I} \Rightarrow (D, B, C) \in \mathcal{I} \text{ whenever } D \subseteq A.$$

We say that $\mathcal{I}$ satisfies right decomposition *if*

$$(A, B, C) \in \mathcal{I} \Rightarrow (A, D, C) \in \mathcal{I} \text{ whenever } D \subseteq B.$$

We say that $\mathcal{I}$ satisfies left weak union *if*

$$(A, B, C) \in \mathcal{I} \Rightarrow (A, B, C \cup D) \in \mathcal{I} \text{ whenever } D \subseteq A.$$

We say that $\mathcal{I}$ satisfies left contraction *if*

$$(A, B, C) \in \mathcal{I},\ (D, B, A \cup C) \in \mathcal{I} \Rightarrow (A \cup D, B, C) \in \mathcal{I}.$$

The following is similar to results in Didelez (2006) and Mogensen et al. (2018).

Proposition B.2. *Let $\mathcal{I}$ be a local independence model, or a Granger causality model, i.e., an independence model constructed using Definition 1.1 or Definition 1.2. The independence model $\mathcal{I}$ satisfies left and right decomposition.*

Proof. Left and right decomposition follow immediately from the definitions. □

C DETECTION OF FAITHFULNESS VIOLATIONS

Assume that $\mathcal{I}$ is not faithful with respect to the causal graph $\mathcal{D}$. We say that a failure of faithfulness is *detectable* if there is no other graph, $\mathcal{D}'$, such that $\mathcal{I} = \mathcal{I}(\mathcal{D}')$, i.e., no other graph, $\mathcal{D}'$, such that $\mathcal{I}$ is Markov and faithful with respect to $\mathcal{D}'$ (Zhang and Spirtes, 2008). Detectability implies that we, in principle and for infinite data, will realize that the independence model we are observing is not graphical.

Theorem C.1. *If we assume Markovness and causal minimality, and the faithfulness assumption fails, then the failure is detectable.*

Proof. Assume that $\mathcal{I}(\mathcal{D}) \subsetneq \mathcal{I}$. If the failure is undetectable, there exists $\mathcal{D}'$ such that $\mathcal{I}(\mathcal{D}') = \mathcal{I}$. In this case, $\mathcal{I}(\mathcal{D}) \subsetneq \mathcal{I}(\mathcal{D}')$. Proposition C.2 gives $\mathcal{D}' \subseteq \mathcal{D}$, and therefore $\mathcal{D}' \subsetneq \mathcal{D}$. However, this is a violation of causal minimality. Alternatively, this follows also from uniqueness in Proposition 4.7. □

Proposition C.2. *Let $\mathcal{D}_1 = (V, E_1)$ and $\mathcal{D}_2 = (V, E_2)$ be directed graphs. If $\mathcal{I}(\mathcal{D}_1) \subseteq \mathcal{I}(\mathcal{D}_2)$, then $\mathcal{D}_2 \subseteq \mathcal{D}_1$.*

Proof. Let $\alpha \to \beta$ be an edge which is not in $\mathcal{D}_1$. In this case, $(\alpha, \beta, \mathrm{pa}_{\mathcal{D}_1}(\beta)) \in \mathcal{I}(\mathcal{D}_1)$ and $\mathrm{pa}_{\mathcal{D}_1}(\beta) \subseteq V \setminus \{\alpha\}$. Therefore, $(\alpha, \beta, \mathrm{pa}_{\mathcal{D}_1}(\beta)) \in \mathcal{I}(\mathcal{D}_2)$, and $\alpha \to \beta$ is not in $\mathcal{D}_2$ as $\alpha \notin \mathrm{pa}_{\mathcal{D}_1}(\beta)$. □

D EDGE ORDER

We define the order of the pair (α, β).

Definition D.1 (Order of an ordered pair of nodes)**.** *Let $\mathcal{D} = (V, E)$ be a directed graph, and $\alpha, \beta \in V$, $\alpha \neq \beta$. The order of (α, β) relative to $\mathcal{D}$ is $\inf\{|C| : (\alpha, \beta, C) \in \mathcal{I}(\mathcal{D}), C \subseteq V \setminus \{\alpha\}\}$, and we denote this by $o(\alpha, \beta, \mathcal{D})$.*

By convention $o(\alpha, \beta, \mathcal{D}) = \infty$ if and only if there is no set $C \subseteq V \setminus \{\alpha\}$ such that $(\alpha, \beta, C) \in \mathcal{I}(\mathcal{D})$. The *order* of a graph, $\mathcal{D} = (V, E)$, is the largest, finite order $o(\alpha, \beta, \mathcal{D})$ if such a finite order exists.

Proposition D.2. *If $o(\alpha, \beta, \mathcal{G}) < \infty$, then $o(\alpha, \beta, \mathcal{G}) \leq |pa_{\mathcal{D}}(\beta)|$*

Proof. If $o(\alpha, \beta, \mathcal{G}) < \infty$, then $\alpha \to \beta$ is not in $\mathcal{D}$, and therefore β is μ-separated from α given $\mathrm{pa}_{\mathcal{D}}(\beta) \subseteq V \setminus \{\alpha\}$. □

E THE dSGS-ALGORITHM

Algorithm 2: Dynamical SGS (dSGS)

```
input  : I over V such that |V| = n, and k such that 0 ≤ k ≤ n − 1
i ← 0;
for i = 0, 1, ..., k do
    for β ∈ V do
        for α ∈ V \ {β} do
            for C ⊆ V \ {α} : |C| = k do
                if (α, β, C) ∈ I then
                    E ← E \ {α → β};
                    D ← (V, E);
                end
            end
        end
    end
end
D_sgs ← D;
output: D_sgs
```

The following result uses only the Markov assumption. This is analogous to the SGS-algorithm for DAGs (Spirtes et al., 2001). The order of a graph is defined in Appendix D.

Proposition E.1. *Assume that $\mathcal{D}$ is of order less than or equal to m, and that $\mathcal{I}$ is Markov with respect to $\mathcal{D}$. In the oracle case, the output of Algorithm 2 (dSGS), using $k = m$ as the integer parameter, is a subgraph of $\mathcal{D}$.*

Proof. Assume $\alpha \to \beta$ is not $\mathcal{D}$. In this case, $(\alpha, \beta, \mathrm{pa}_{\mathcal{D}}(\beta)) \in \mathcal{I}(\mathcal{D})$, $\alpha \notin \mathrm{pa}_{\mathcal{D}}(\beta)$. We have $o(\alpha, \beta, \mathcal{D}) < \infty$, and therefore $o(\alpha, \beta, \mathcal{D}) \leq m$ by assumption. There exists a C, $|C| \leq m$, and $C \subseteq V \setminus \{\alpha\}$ such that $(\alpha, \beta, C) \in \mathcal{I}(\mathcal{D})$. Using the Markov property, $(\alpha, \beta, C) \in \mathcal{I}$. In the oracle case, this edge is therefore removed using the set C, and $\alpha \to \beta$ is not in $\mathcal{D}_{sgs}$. □

F LEARNING AND FAITHFULNESS

In this section, we relate the contents of Section 3 to structure learning.

F.1 Learning the Edge-Transitive Graph

From a collection of test results, i.e., an empirical independence model, we may output the corresponding edge-transitive graph. This graph is defined for *any* independence model and Algorithm 3 (for $k = n$) therefore

outputs a graph which is faithful to the observed test results, regardless of whether there are statistical errors in the test results.

We say that $\mathcal{I}$ is k-faithful with respect to $\mathcal{D}$ if for all C such that $|C| \leq k$

$$(A, B, C) \in \mathcal{I} \Rightarrow (A, B, C) \in \mathcal{I}(\mathcal{D}).$$

One should note that n-faithfulness, and $(n-1)$-faithfulness, is the same as faithfulness. The idea of k-faithfulness is similar in nature to Mogensen (2023) which defines a weak notion of Markov equivalence by restricting the size of the conditioning sets.

Algorithm 3: Edge-Transitive Graph

input : $\mathcal{I}$ over V such that $|V| = n$, and k such that $0 \leq k \leq n-1$
$i \leftarrow 0$;
$\mathcal{D}$ is the complete graph on nodes V;
for $i = 0, 1, \ldots, k$ **do**
  for $\beta \in V$ **do**
    for $\alpha \in V \setminus \{\beta\}$ **do**
      for $C \subseteq V \setminus \{\alpha\} : |C| = k$ **do**
        if *E0, E1, E2, or E3 is violated* **then**
          $E \leftarrow E \setminus \{\alpha \to \beta\}$;
          $\mathcal{D} \leftarrow (V, E)$;
        end
      end
    end
  end
end
output: $\mathcal{D}$

The next proposition follows immediately from the proof of Theorem 3.3.

Proposition F.1. *The output of Algorithm 3 is k-faithful.*

F.2 Trimming the Output of a Learning Algorithm

Constraint-based learning algorithms proceed by testing a number of conditional independences. These tests results may be reused to check conditions D0, D1, D2, and D3, and remove any edges that are in violation.

Algorithm 4: Trimming

input : $\mathcal{I}$ over V such that $|V| = n$, and a graph $\mathcal{D} = (V, E)$
for $\beta \in V$ **do**
  for $\alpha \in V \setminus \{\beta\}$ **do**
    for $C \subseteq V \setminus \{\alpha\}$ **do**
      if *D0, D1, D2, or D3 is violated* **then**
        $E \leftarrow E \setminus \{\alpha \to \beta\}$;
        $\mathcal{D} \leftarrow (V, E)$;
      end
    end
  end
end
$\mathcal{D}_{tr} \leftarrow \mathcal{D}$;
output: $\mathcal{D}_{tr}$

Proposition F.2. *The independence model $\mathcal{I}$ (input in Algorithm 4) is faithful with respect to $\mathcal{D}$ (output in Algorithm 4).*

A constraint-based learning algorithm need not test all possible independences. If we let $\mathcal{I}$ be an 'empirical' independence model, i.e., a set of local independences that are believed to hold/not hold based on statistical tests, we may not have access to the entire $\mathcal{I}$ after running a learning algorithm. In that case, the trimming would need to be restricted to the observed part of $\mathcal{I}$.

Proof. Let $\mathcal{D}_{tr}$ be the output of Algorithm 4. If $\alpha \to \beta$ is in $\mathcal{D}_{tr}$, then this edges satisfies the conditions in Definition 2.1 for some graph $\mathcal{D}_i$ such that $\mathcal{D}_{tr} \subseteq \mathcal{D}_i \subseteq \mathcal{D}$. Therefore, the conditions are also satisfied for $\mathcal{D}_{tr}$, and by Theorem 3.1, $\mathcal{I}$ is faithful with respect to $\mathcal{D}_{tr}$. □

G PARTIAL OBSERVATION

In this section, we will turn our attention to the setting where we only assume *partial observation*. We will assume that there exists an underlying causal graph, $\mathcal{D} = (V, E)$, however, we only observe the coordinate processes in the set O, $O \subseteq V$. In this case, one can use a so-called *latent projection* to compute a *directed mixed graph*, $\mathcal{G}$, such that $\mathcal{I}(\mathcal{G}) = \mathcal{I}(\mathcal{D})_O$ where $\mathcal{I}(\mathcal{D})_O = \{(A, B, C) \in \mathcal{I}(\mathcal{D}) : A, B, C \subseteq O\}$ (Mogensen and Hansen, 2020). A directed mixed graph may have both directed, $\to$, and bidirected edges, $\leftrightarrow$. In case of partial observation, we only have access to $\mathcal{I}_O = \{(A, B, C) \in \mathcal{I} : A, B, C \subseteq O\}$ as we can only test local independence among the observed coordinate processes. It is important to note that the *partial observation* in this paper refers to the fact that some coordinate processes are fully unobserved.

A first observation is the fact that the corresponding induced local independence graph, $\mathcal{D}_{I_O}$, may still be useful, even if its interpretation is slightly different. Assuming the equivalence of pairwise and global Markov properties, we still have $\mathcal{I}(\mathcal{D}_{I_O}) \subseteq \mathcal{I}_O$ such that μ-separation in the induced local indepedence graph implies local independence.

Proposition G.1. *Let $\mathcal{D} = (V, E)$, and let $\mathcal{G} = (O, E)$ be the latent projection of $\mathcal{G}$ over O, $O \subseteq V$. If $\mathcal{I}$ is faithful (ancestor faithful) with respect to $\mathcal{D}$, then $\mathcal{I}_O$ is faithful (ancestor faithful) with respect to $\mathcal{G}$.*

Proof. If $\mathcal{I}$ is faithful with respect to $\mathcal{D}$, then $\mathcal{I}_O$ is clearly faithful with respect to $\mathcal{G}$ using the fact that $\mathcal{I}(\mathcal{G}) = \mathcal{I}(\mathcal{D})_O$.

Assume $\mathcal{I}$ is ancestor faithful with respect to $\mathcal{D}$. If there is a directed path from A to B in $\mathcal{G}$ which is μ-connecting given C, then there is also a directed path from A to B in $\mathcal{D}$ which is μ-connecting given C, and we see that $(A, B, C) \notin \mathcal{I}_O$. □

On the other hand,´parent faithfulness or causal minimality of $\mathcal{I}$ and $\mathcal{D}$ is not inherited by $\mathcal{I}_O$ and $\mathcal{G}$ in this way.

Note that the next proposition does not assume causal sufficiency.

Proposition G.2 (Mogensen (2020a))**.** *Assume ancestor faithfulness of $\mathcal{I}$ with respect to $\mathcal{D}$. If $\alpha \to \beta$ is not in the output of the CS-algorithm (in the oracle case), then $\alpha \to \beta$ is not in the latent projection of the causal graph, $\mathcal{G}$.*

The edge $\alpha \to \beta$, $\alpha \neq \beta$, is in the latent projection of $\mathcal{D}$ if and only if there is a directed path from α to β in $\mathcal{D}$ such that all nonendpoint nodes are unobserved, i.e., not in O.

H SIMULATIONS

We generated data from a VAR(1)-process in the following way. We first generated a graph by sampling a parameter, a, from a uniform distribution on the interval $[0, 0.5]$, and then we sampled edges independently using a as the success parameter. Given the graphical structure, we sampled the nonzero regression parameters independently and uniformly on $[-1, 1]$. We kept sampling until the result was a stable VAR(1)-process. We sampled data from this VAR(1)-process (100 observed time points). We repeated the entire procedure M times (see Figures 3 and 4).

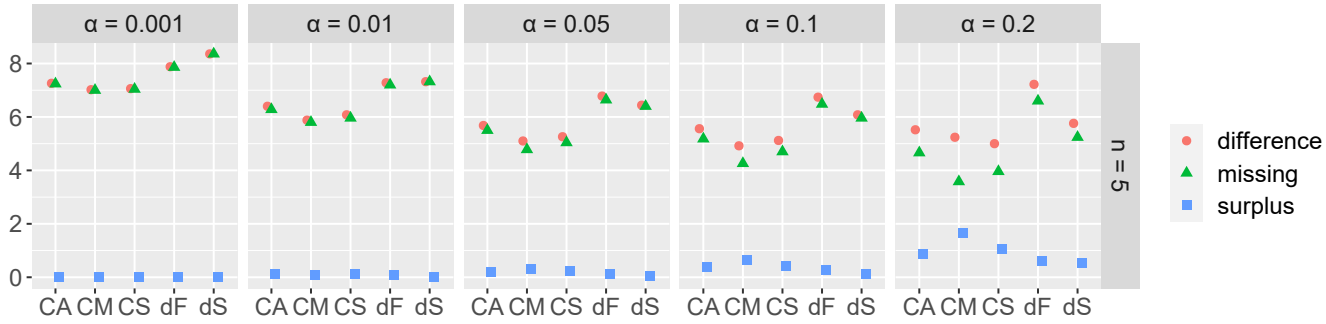

Figure 4: Comparison of algorithms in the case of partial observation (five coordinate processes, $n = 5$). Points indicate mean over $M = 50$ repetitions (see caption of Figure 3 for a description of the symbols). Subsection H.1 provides more details.

The simulations were implemented in `R` and we used the Granger causality test´in the FIAR package (`condGranger`). Code is available along with this paper.

H.1 Partial Observation

We also compare the algorithms in the case of partial observation (Appendix G), also including the dFCI-algorithm from Mogensen et al. (2018). This algorithm is the only algorithm of the five in Figure 4 which is sound and complete in the case of partial observation, i.e., outputs the true graph in the oracle case. In the case of partial observation, the learning target is the greatest element of the Markov equivalence class of the true graph (Mogensen and Hansen, 2020) as the true graph itself is not necessarily identifiable from tests of local independence. We compare the output of the learning algorithms only to the *directed* part of the learning target, i.e., we ignore bidirected edges in the learning target.

For this experiment, we sampled the number of unobserved nodes uniformly on $\{0, 1, \dots, n\}$. The true (and fully observed) graph was then sampled as in Figure 3 (see above description). We marginalized the graph using the latent projection and computed the greatest element of the Markov equivalence class of the latent projection.

As seen from Figure 4, the dFCI does not fare better than the simpler algorithms. Most likely this is due to the fact that it uses a large number of tests and makes decisions sequentially based on these test results. This may lead to propagation of statistical errors.

H.2 Number of Tests

Of the algorithms reported, only the CA-, the CS-, and the dFCI-algorithms are 'adaptive' in the sense that they use different numbers of tests depending on the test results. The dSGS-algorithm uses all possible tests, the CM-algorithm uses a single test for each ordered pair of nodes, and the CS-algorithm uses at most two tests for each ordered pair of nodes. In the experiment reported in Figure 3, the number of tests used by the CA-algorithm was in the ranges $20 - 142$ ($n = 5$), $45 - 621$ ($n = 7$), and $79 - 1509$ ($n = 9$), respectively.

I PROOFS

Proof of Proposition 2.2. We should show that the conditions D0-D3 in Definition 2.1 hold for every $C \subseteq V$ when $\mathcal{I} = \mathcal{I}(\mathcal{D})$. D0 holds as $\alpha \to \beta$ is μ-connecting for all C such that $\alpha \notin C$. In D1, if $(\gamma, \alpha, C) \notin \mathcal{I}(\mathcal{D})$, then there is a μ-connecting walk from γ to α given C, and if $\alpha \notin C$, then the composition of this walk with the edge $\alpha \to \beta$ is μ-connecting from γ to β given C. Conditions D2 and D3 follow similarly. This is clear from the definition of μ-separation. □

Lemma I.1 (Mogensen and Hansen (2020))**.** *If there is a μ-connecting walk from α to β given C, then there is a μ-connecting walk from α to β given C such that all colliders are in C.*

Proof of Proposition 2.3. Assume $\alpha \to \beta$ is in $\mathcal{D}$. If $\alpha \notin C$, then $(\alpha, \beta, C) \notin \mathcal{I}_2$, and therefore $(\alpha, \beta, C) \notin \mathcal{I}_1$. The other conditions follow similarly using the fact that $\mathcal{I}_1 \subseteq \mathcal{I}_2$. □

We use the notation $\alpha \sim \beta$ to indicate an edge, $\alpha \to \beta$ or $\alpha \leftarrow \beta$, between nodes α and β.

Proof of Theorem 3.1. Assume first that $\mathcal{D}$ is transitively closed with respect to $\mathcal{I}$, and assume $(A, B, C) \notin \mathcal{I}(\mathcal{G})$. Let $\tilde{\omega}$ be a μ-connecting walk from $\gamma \in A$ to $\delta \in B$ given C. We can find a walk, ω, which is μ-connecting from γ to δ given C such that all colliders on ω are in C (Proposition I.1). If ω has length 1, then $\gamma \to \delta$, $\gamma \notin C$, and from D0 we have $(\gamma, \delta, C) \notin \mathcal{I}$. Otherwise, the walk has a nonendpoint node, ε, $\gamma \sim \ldots \sim \varepsilon \to \delta$, and ω is of one of the three types in Lemma I.2. If it is of type 1, then there is a μ-connecting walk from γ to ε given C and $\varepsilon \notin C$ as ω is μ-connecting. D1 gives that $(\gamma, \delta, C) \notin \mathcal{I}$. If ω is of type 2, there is an edge $\alpha \to \beta$ on ω such that β is in C (all colliders on ω are in C), $\alpha \notin C$, the subwalk from γ to β is μ-connecting given C, and the subwalk from α to δ is μ-connecting given C. D2 gives that $(\gamma, \delta, C) \notin \mathcal{I}$. If ω is of type 3, there must be a head at γ, $\gamma \leftarrow \alpha \leftarrow \ldots \leftarrow \varepsilon \to \delta$. The subwalk from α to δ is μ-connecting given C as $\alpha \notin C$, and D3 gives that $(\gamma, \delta, C) \notin \mathcal{I}$. This means that in each case $(\gamma, \delta, C) \notin \mathcal{I}$. From the left and right decomposition properties of local independence, this means that $(A, B, C) \notin \mathcal{I}$. Note that the right decomposition property is immediate from the definition of local independence/Granger noncausality that we use (see also Section B).

Assume now that $\mathcal{I}$ is faithful with respect to $\mathcal{D}$, that is, $\mathcal{I} \subseteq \mathcal{I}(\mathcal{D})$. Proposition 2.2 gives that $\mathcal{I}(\mathcal{D})$ is transitively closed with respect to $\mathcal{D}$, and Proposition 2.3 gives that $\mathcal{I}$ is transitively closed with respect to $\mathcal{D}$. □

For convenience, we say that a μ-connecting walk of length strictly greater than 1 is of type 1 if $\alpha \cdots \to \gamma \to \beta$. We say that it is of type 2 if $\alpha \cdots \leftarrow \gamma \to \beta$ and it contains a collider, and we say that it is of type 3 if $\alpha \cdots \leftarrow \gamma \to \beta$ and it does not contain a collider. The following lemma helps clarify the contents of Definition 2.1: D0-D3 are essentially sufficient to characterize the μ-connecting walks.

Lemma I.2. *Any μ-connecting walk of length strictly greater than 1 is of type 1, 2, or 3.*

Proof. Let ω be a μ-connecting walk of length strictly greater than 1. In this case, there is a nonendpoint node, γ, $\alpha \sim \ldots \sim \gamma \to \beta$. The statement follows immediately from this. □

The next corollary follows from Theorem 3.1 and the definition of faithfulness as $\mathcal{I}(\mathcal{D}_2) \subseteq \mathcal{I}(\mathcal{D}_1)$ when $\mathcal{D}_1 \subseteq \mathcal{D}_2$.

Corollary I.3. *Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be graphs such that $\mathcal{D}_1 \subseteq \mathcal{D}_2$. If $\mathcal{D}_2$ is transitively closed with respect to $\mathcal{I}$, then $\mathcal{D}_1$ is transitively closed with respect to $\mathcal{I}$.*

Proof of Theorem 3.3. Assume $(A, B, C) \notin \mathcal{I}(\mathcal{F}_\mathcal{I})$. In this case, there is a μ-connecting walk from $\alpha \in A$ to $\beta \in B$ given C. We can then also find a μ-connecting walk in $\mathcal{F}_I$ from $\alpha \in A$ to $\beta \in B$ given C such that all colliders are in C (Proposition I.1). We show by induction on walk length that the existence of a μ-connecting walk, ω, from γ to δ given C implies that $(\gamma, \delta, C) \notin \mathcal{I}$, assuming that all colliders on ω are in C. If ω has length 1, then E0 gives the result. Assume now that it holds for all walks of lengths 1,2,…,$m-1$ and with all colliders in C that μ-connectivity implies dependence. We consider a μ-connecting walk of length m. Assume this walk is of type 1, say, $\gamma \sim \ldots \to \varepsilon \to \delta$. The subwalk from γ to ε is μ-connecting given C and has length $m-1$. From the induction assumption, $(\gamma, \varepsilon, C) \notin \mathcal{I}$. As $\varepsilon \to \beta$ is in $\mathcal{F}_I$ and $\varepsilon \notin C$, we have that $(\gamma, \delta, C) \notin \mathcal{I}$. If it is of type 2, there is an edge $\alpha \to \beta$, a μ-connecting walk from γ to $\beta \in C$ (all colliders on ω are in C), and a μ-connecting walk from $\alpha \notin C$ to δ given C such that the μ-connecting walks are both of length less than $m-1$. Using the induction hypothesis and E2, we have $(\gamma, \delta, C) \notin \mathcal{I}$. Finally, if ω is of type 3, then it follows from similar arguments and E3. □

Proof of Proposition 3.4. We have $\mathcal{I} = \mathcal{I}(\mathcal{D})$ for a graph $\mathcal{D}$. If e is in $\mathcal{D}$, then it follows from Proposition 2.2 that e is in $\mathcal{F}_I$ by comparing Definitions 2.1 and 3.2 and using $\mathcal{I} = \mathcal{I}(\mathcal{D})$. If e, say $\alpha \to \beta$, is not in $\mathcal{D}$, then there exists a set C, $\alpha \notin C$, such that β is μ-separated from α given C in $\mathcal{D}$, and $(\alpha, \beta, C) \in \mathcal{I}(\mathcal{D}) = \mathcal{I}$. Therefore, e is not in $\mathcal{F}_I$ using E0. □

Proof of Proposition 4.5. The first two implications are obvious from the definitions.

If $\mathcal{I}$ is Markov and parent faithful with respect to $\mathcal{D}$, we can consider a proper subgraph $\mathcal{D}_0$ of $\mathcal{D}$. There is some edge which is in $\mathcal{D}$, but not in $\mathcal{D}_0$, say $\alpha \to \beta$, $\alpha \neq \beta$. Using parent faithfulness of $\mathcal{I}$ with respect to $\mathcal{D}$, we have $(\alpha, \beta, C) \notin \mathcal{I}$ for all C such that $\alpha \notin C$. We have $(\alpha, \beta, \mathrm{pa}_{\mathcal{D}_0}(\beta)) \in \mathcal{I}(\mathcal{D}_0)$ and $\alpha \notin \mathrm{pa}_{\mathcal{D}_0}(\beta)$ which means that $\mathcal{I}$ is not Markov with respect to $\mathcal{D}_0$. Therefore, $\mathcal{I}$ is causally minimal with respect to $\mathcal{D}$, and we conclude that the combination of Markovness and parent faithfulness implies causal minimality. □

Proof of Proposition 4.6. By definition of the induced local independence graph, $\mathcal{I}$ satisfies the pairwise Markov property with respect to $\mathcal{D}_I$, and therefore $\mathcal{I}$ is Markov with respect to $\mathcal{D}_I$. Let $\mathcal{D}_0$ be a proper subgraph of $\mathcal{D}_I$, say $\alpha \to \beta$, $\alpha \neq \beta$, is in $\mathcal{D}$, but not in $\mathcal{D}_0$. In this case, $(\alpha, \beta, V \setminus \{\alpha\}) \in \mathcal{I}(\mathcal{D}_0)$ such that $\mathcal{I}(\mathcal{D}_0) \not\subseteq \mathcal{I}$. □

Proof of Proposition 4.7. If $\alpha \to \beta$ is not in $\mathcal{D}$, then β is μ-separated from α given $V \setminus \{\alpha\}$. If $\alpha \to \beta$ is in $\mathcal{D}_I$, then $(\alpha, \beta, V \setminus \{\alpha\}) \notin \mathcal{I}$ and $\alpha \to \beta$ must be in $\mathcal{D}$ if $\mathcal{I}(\mathcal{D}) \subseteq \mathcal{I}$. Any causally minimal graph is therefore a supergraph of $\mathcal{D}_I$. As $\mathcal{D}_I$ is causally minimal it follows that $\mathcal{D}_I$ is the only such graph. □

Proof of Proposition 4.8. Any μ-connecting walk must have a head into β, and be of length at least 2, $\alpha \sim \ldots \sim \gamma \to \beta$. We see that $\gamma \in C$, and therefore β is μ-separated from α given C. Using Markovness, $(\alpha, \beta, C) \in \mathcal{I}$.

Assume now that $\mathcal{I}$ satisfies left weak union, left decomposition, and left contraction. We have $\mathrm{pa}_{\mathcal{D}}(\beta) \setminus \{\alpha\} \subseteq C$, and therefore $(V \setminus \mathrm{pa}_{\mathcal{D}}(\beta), \beta, C \cup \{\alpha\}) \in \mathcal{I}$ using the global Markov property and the above argument. If $(\alpha, \beta, C) \in \mathcal{I}$, we use left contraction to obtain $(\{\alpha\} \cup V \setminus \mathrm{pa}_{\mathcal{D}}(\beta), \beta, C) \in \mathcal{I}$. Using left weak union and left decomposition we obtain $(\alpha, \beta, V \setminus \{\alpha\}) \in \mathcal{I}$ using that $\mathrm{pa}_{\mathcal{D}}(\beta) \subseteq C$. Using the equivalence of pairwise and global Markov properties, we see that $\mathcal{I}$ is Markov with respect to the graph obtained by removing the edge $\alpha \to \beta$, and this is a violation of causal minimality. □

Proof of Lemma 4.10. The conditions D1' and D3' are weaker than conditions D1 and D3, respectively, so one direction is immediate. We assume that D0, D1', D2, and D3' hold for every edge and set C. To show that D1 holds, assume there is a μ-connecting walk from γ to α given C, and we can choose this walk such that all colliders are in C. If this walk is directed, then the result follows immediately. Otherwise, if there is no colliders on the walk, it follows from D3'. If there is a collider, there is some edge on the walk which points towards γ, and let $\phi \to \psi$ such that ψ is a collider. We see that the result follows from D2. Condition D3 is shown similarly. □

Proof of Theorem 4.11. Assume that $\mathcal{I}$ is parent faithful with respect to $\mathcal{D}$, and let $\alpha \to \beta$ be an edge in $\mathcal{D}$. In this case, $(\alpha, \beta, C) \notin \mathcal{I}$, $\alpha \notin C$, and D0 holds. On the other hand, assume that D0 holds. In this case $(\alpha, \beta, C) \notin \mathcal{I}$, $\alpha \notin C$, and it follows from left and right decomposition that $(A, B, C) \notin \mathcal{I}$ for all A and B such that $\alpha \in A$ and $\beta \in B$.

Assume that $\mathcal{I}$ is ancestor faithful with respect to $\mathcal{D}$. In this case, they are also parent faithful, and D0 follows. If there is a directed path from γ to α which is μ-connecting given C, $\alpha \notin C$, and $\alpha \to \beta$, then $(\gamma, \beta, C) \notin \mathcal{I}$ using ancestor faithfulness. On the other hand, assume that D0 and D1' hold, and that there is a μ-connecting walk from α to β given C, $\alpha \notin C$. If it has length one, then it follows from D0 that $(\alpha, \beta, C) \notin \mathcal{I}$. Otherwise, it has the form $\alpha \to \ldots \to \gamma \to \beta$, and $(\alpha, \beta, C) \notin \mathcal{I}$ follows from D1' since the subwalk from α to γ is μ-connecting given C and $\gamma \notin C$. It follows from left and right decomposition of $\mathcal{I}$ that $(A, B, C) \notin \mathcal{I}$ for all A and B such that $\alpha \in A$ and $\beta \in B$.

Assume that $\mathcal{I}$ is trek faithful with respect to $\mathcal{D}$. It is also parent and ancestor faithful, and D0 and D1' follow. If $\alpha \to \beta$ and there is a μ-connecting trek from α to γ given C, then there is also a μ-connecting trek from β to γ given C, and $(\alpha, \beta, C) \notin \mathcal{I}$ using trek faithfulness. Assume now that D0, D1', and D3' hold, and assume that there is a μ-connecting trek from α to β given C. If the trek has length one, $(\alpha, \beta, C) \notin \mathcal{I}$ follows from D0. If it is a directed walk, then $(\alpha, \beta, C) \notin \mathcal{I}$ follows from D1'. If it is not a directed walk, then it must have heads at both endpoints such that $\alpha \leftarrow \gamma \sim \ldots \to \beta$. There is a μ-connecting trek from γ to β given C and using D3' gives the result. Again, $(A, B, C) \notin \mathcal{I}$ for all A and B such that $\alpha \in A$ and $\beta \in B$. □

Proof of Proposition 5.2. Under parent dependence, the first step outputs a supergraph of the causal graph, $\mathcal{D}$: If $\alpha \to \beta$ is in $\mathcal{D}$, then $(\alpha, \beta, \beta) \notin \mathcal{I}$, and this edge is not removed. Let $\mathcal{D}_1$ denote the output of the first step. We have $\mathrm{pa}_{\mathcal{D}}(\beta) \subseteq \mathrm{pa}_{\mathcal{D}_1}(\beta)$ for all $\beta \in V$. Proposition 4.8 implies that each iteration of the second step outputs a supergraph of the causal graph and that the final output is the causal graph. □