Graphical modeling in dynamical systems

PhD thesis

Søren Wengel Mogensen

Department of Mathematical Sciences,
University of Copenhagen

This thesis has been submitted to the PhD School of the Faculty of Science,
University of Copenhagen

SØREN WENGEL MOGENSEN
SWENGEL@MATH.KU.DK

DEPARTMENT OF MATHEMATICAL SCIENCES
UNIVERSITY OF COPENHAGEN
UNIVERSITETSPARKEN 5
DK-2100 COPENHAGEN
DENMARK

Academic advisor:	Professor Niels Richard Hansen University of Copenhagen
Assessment committee:	Professor Steffen Lauritzen University of Copenhagen Professor Vanessa Didelez Leibniz Institute for Prevention Research and Epidemiology Associate Professor Robin Evans University of Oxford
Submission date:	April 30, 2020
ISBN:	978-87-7078-897-7

Preface

This thesis was written while I was a PhD student at the University of Copenhagen. My studies were funded by VILLUM FONDEN (research grant 13358) and I was working under the supervision of Professor Niels Richard Hansen at the Department of Mathematical Sciences.

The main topic of the thesis is graphical modeling of dynamical systems. Graphical modeling offers a wide range of methods that could also be studied in dynamical systems. However, there are only so many roads a single PhD thesis can go down, and our focus is on the graphs themselves as mathematical objects and on the task of structure learning.

There are many people that, in one way or the other, have helped shape this thesis. First and foremost, I thank Niels for his patient supervision, for the many conversations we have had over the years, and for his help as an advisor and a collaborator. Around the time when I enrolled as a PhD student, the Copenhagen Causality Lab was founded by Niels, Professor Steffen Lauritzen, and Professor Jonas Peters. It has been a pleasure to see the group grow and expand and to be a part of a friendly community of people working on causal inference problems. I thank everyone in the lab for answering my often half-baked questions on graphical modeling and causal inference and for being great co-workers. I also thank everyone else in the statistics and probability group for creating an enjoyable environment.

One of the papers in this thesis was co-authored with Daniel Malinsky and I thank him for our many discussions on causal inference and for his outstanding hospitality during my visits to Pittsburgh and Baltimore. In the fall of 2019, I visited the Department of Statistics at the University of Washington. I sincerely thank Professor Thomas Richardson for hosting my visit, working with me, and for sharing his thoughts on science. I also thank Emilija Perković and Vincent Roulet for welcoming me into their social lives and for helping me find my feet in a new city.

Finally, I am extremely grateful to Karen who has been an exceptional source of encouragement and support, in particular on the final stretch to the finish line.

Søren Wengel Mogensen
April 2020

Til Karen

Abstract

This thesis studies graphical modeling of local independence in stochastic processes. By applying a separation criterion to graphs, we can obtain a graphical representation of a local independence structure which describes how the system evolves over time. We describe a class of graphs which facilitates graphical modeling of partially observed stochastic processes driven by correlated error processes.

Within this graphical framework, we prove some Markov properties for specific classes of stochastic processes which provides a link between a graph and the local independence structure of a stochastic process. Graphs that encode the same independence structure are said to be Markov equivalent. Characterizations of Markov equivalence are interesting as they allow us to understand which graphical structures are indistinguishable in terms of the separation models that they encode. We prove several results relating to Markov equivalence in different classes of graphs. We also consider various computational problems and show that many of the naturally occuring problems are hard in these classes of graphs.

We consider structure learning in the case of partially observed stochastic processes, i.e., the task of recovering a graphical representation from an observational distribution. Exact structure learning based on tests of local independence is also hard and we suggest an approximation algorithm which is computationally feasible.

Resumé

Denne afhandling beskriver grafiske modeller af lokal uafhængighed for stokastiske processer. Ved hjælp af et separationskriterium kan vi anvende grafer som en repræsentation af lokal uafhængighed, der beskriver, hvordan et system udvikler sig over tid. Vi beskriver en klasse af grafer, som gør det muligt at lave grafiske modeller af delvist observerede stokastiske processer, der drives af korrelerede fejlprocesser.

Inden for denne familie af grafer beviser vi nogle Markovegenskaber for specifikke klasser af stokastiske processer, hvilket skaber en forbindelse mellem graferne og lokal uafhængighed af stokastiske processer. Grafer, der repræsenterer den samme uafhængighedsstruktur, siges at være Markovækvivalente. Karakteriseringer af Markovækvivalens er interessante, da de giver os en forståelse for, hvilke grafiske strukturer vi ikke er i stand til at skelne ud fra de separationsmodeller, som de repræsenterer. Vi beviser adskillige resultater vedrørende Markovækvivalens i forskellige klasser af grafer. Vi undersøger også flere beregningsmæssige problemer og viser, at mange af de naturligt forekommende problemer i disse klasser af grafer er beregningsmæssigt svære.

Vi undersøger strukturlæring af delvist observerede stokastiske processer, altså hvordan man kan lære en grafisk repræsentation ud fra en observationel fordeling. Eksakt strukturlæring ved hjælp af tests af lokal uafhængighed er også beregningsmæssigt svært, og vi foreslår en approksimativ algoritme, som er beregningsmæssigt hensigtsmæssig.

Contents

Chapter 1

Introduction

Stochastic processes are often used as formal means of reasoning about the evolution of a dynamical system. In complicated systems with many components one needs some type of sparsity assumption to easily make sense of the behavior of the system. In multivariate stochastic processes, we can induce such a sparsity structure by using *local independence* (Schweder, 1970; Aalen, 1987). In words, a set of coordinate processes, B, is locally independent of another set, A, given a third set, C, if the prediction of what happens in B at any time point does not improve when learning about the past of A whenever the past of C is already known. This allows us to describe a certain sparsity in the evolution of the system as at any time point a coordinate process need not depend on the past of every other coordinate process. Local independence can be used as a formalization of the notion that in a large, sparse system, the influence of some processes (A) on others (B) is mediated entirely through some third set of processes (C). Starting from the notion of local independence, we can ask for representations of local independence structure, for instance, a *graph* representing this structure (Didelez, 2000).

Local independence can be thought of as a dynamical version of classical conditional independence of random variables. There is a rich literature on graphical models of conditional independence (Lauritzen, 1996; Maathuis et al., 2018) which allows, e.g., modeling marginal distributions and systems with correlated errors. Much of this thesis is concerned with finding analogous concepts and methods to use in models of local independence. While the starting point for graphical modeling has most often been conditional independence of random variables, our starting point will be local independence of stochastic processes, building on work by, e.g., Schweder (1970); Aalen (1987); Didelez (2000); Eichler (2013). This will naturally lead to a theory that differs from the classical one, even though it is analogous in many ways. A central difference is the fact that conditional independence of random variables, A is conditionally independent of B given C, is symmetric in arguments A and B. This is not the case for local independence and taking this step away from symmetry we will also see that some problems become more difficult in the asymmetric case, while others become easier. We aim to contribute to a general theory which can be applied to any, discrete-time or continuous-time, stochastic process in which local independence can be defined. On the other hand, continuous-time processes have been the motivating case for us and where we have invested most of our efforts.

While the first part of the thesis describes a graphical framework and relates this framework to local independence, the second part of the thesis looks into structure learning in this framework. Instead of considering a known graph and relating it to local independences of a stochastic process, we will attempt to choose a graph from a set of candidate graphs based on the observed local independences of a system. This is mostly interesting if we actually believe that the ‘connections’ that we recover when doing so are *causal*, e.g., in the sense that they would remain the same under interventions in the system.

Notes to the reader

This thesis builds on work by many other people. We reference that work by (*authors, year*) or by [*number*]. The papers that are a part of the thesis are referenced by capital letters.

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

B Søren Wengel Mogensen and Niels Richard Hansen. Graphical modeling of stochastic processes driven by correlated errors. 2020b

C 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

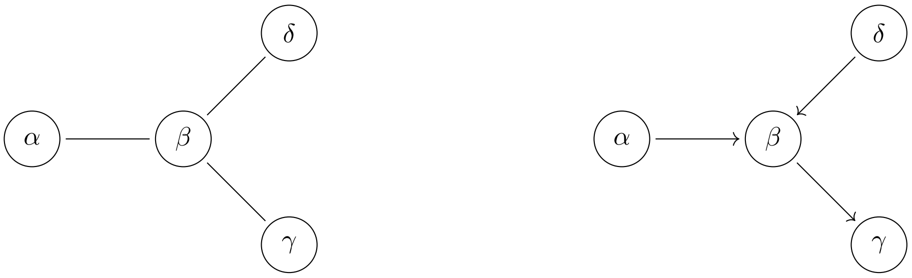


Figure 1.1: Examples of an *undirected graph* (left) and a *directed graph* (right). Both graphs have node sets $\{\alpha, \beta, \gamma, \delta\}$.

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

Within each paper, references to other work are made to bibliographies at the end of the paper. In the rest of the thesis, references are made to the bibliography at the end of the thesis. When referencing results in the papers, we prepend the paper letter, i.e., Definition **A**.3.1 refers to Definition 3.1 in Paper **A**. If the entire thesis is read, we suggest reading Papers **A**-**D** when they appear as subsequent parts of the thesis will assume knowledge of the contents of the papers. The papers themselves are self-contained.

Structure

We give here a short description of the central topics and objects that we will study in the subsequent chapters.

Graphs and separation A *graph* is a discrete, mathematical structure consisting of a node set and an edge set. In Figure 1.1, we give an example of an *undirected* graph (the nodes are α, β, γ, and δ and the edges are drawn as connections between pairs of nodes) and a *directed graph.* A graph often represents some other object, e.g., a multivariate distribution of random variables, interconnected tasks of a project, or a physical network. In this chapter, we describe the classes of graphs that we will be using in the thesis. We also introduce some computational problems relating to graphs. This chapter contains many graph-theoretical details without revealing what the purpose of any of this really is. The impatient reader may want to only skim this chapter and return if needed.

Directed mixed graphs While directed mixed graphs are introduced in Chapter 2, this chapter dives deeper into the theory of this class of graphs. We will use them to represent local independences in partially observed systems of stochastic processes. We study equivalence classes of directed mixed graphs, and we study algorithms for working with these graphs.

Directed correlation graphs Directed graphs can be used to represent local independence structure when the error processes driving the system are independent. In this chapter, we relax this assumption to allow correlated error processes. In this case, we can use directed correlation graphs to represent the local independences of such processes. We prove a global Markov property in a specific model class and study equivalence classes of these graphs.

Structure learning In the above graphical framework, where a directed mixed graph represents the local independences of a multivariate stochastic process (admittedly, in a sense that we still have not made precise), one can ask if it is possible to learn a graphical representation from testing local independence. This chapter studies this problem. The problem is computationally hard, and a learning algorithm which attempts to recover only parts of the structure is also introduced.

A motivating example

Caenorhabditis elegans (C. elegans) is a roundworm and a particularly well-studied organism. It was the first multicellular organism to have its entire genome sequenced. As of April 2020, it is also the only organism to have had its entire *connectome* mapped (White et al., 1986; Achacoso et al., 1989; Varshney et al., 2011; Cook et al., 2019). The connectome describes the neural connections in a nervous system and, at a microscopic level, provides the entire topology of neurons (computational ‘units’) and synapses (their connections). This particular roundworm has approx. 300 neurons and through anatomical studies, synapses between these neurons have been mapped.

Mathematically, we can describe the neuronal activity as a stochastic process $X_t = (X_t^1, \dots, X_t^n)$. A neuron is the basic building block of this system and each neuron corresponds to a single coordinate process, X_t^i. We will relate the multivariate process to a graph, and in this case each coordinate process will be identified with a *node* (see Figure 1.1). The activity of each neuron depends on the activity of other neurons, and this will be represented by *edges* between nodes. Local independence can be used to describe dependence and independence between processes in the system. It will also give us a clear interpretation of the graphical structure and will serve as a testable implication of this structure.

When we turn to structure learning, one may ask why recovering a graph is interesting. In this example, the answer is straightforward. If we, from data, can recover a graph that correctly describes where the synapses are found, then we have in fact learned something about the real world. In the case of a nervous system, we may have reason to believe that these connections are stable under interventions. If we were to intervene on a particular neuron this signal would propagate through the system via the same synapses as when we are not intervening. Much of the contents of this thesis revolves around using graphs to represent local independence of stochastic processes and in many applications there is as such no reason to expect that the graphs are also causal or structural as in the case of neuronal connectivity. In some cases, they may be, though, and this motivates the structure learning endeavor. Learning the connectome of C. elegans from data is, of course, prohibitively difficult. For one reason, because data describing single-neuron activity is not available. Paper **D** returns briefly to C. elegans to use the topology of this system as an example of a large, real-world system. While most readers will probably agree that not every page of this thesis seems to bring us closer to a data-driven unraveling of the C. elegans connectome, we hope that this system can serve as an example of an application where our contributions may offer new perspectives.

Chapter 2

Graphs and separation

Graphs are convenient as mathematical representations of *structure* and are used throughout the sciences (Gross et al., 2013). In statistics, the field of graphical modeling relates properties of graphs to properties of probabilistic models (Maathuis et al., 2018). We will use graphs to represent data-generating mechanisms and as representations of local independence structures. In this section, we will give various graph-theoretical definitions and introduce the classes of graphs that we will need in subsequent chapters. Other graph-theoretical notions are introduced in Papers **A**-**D** as we need them.

Graph-theoretical prerequisites

A *graph* is a pair $\mathcal{G} = (V, E)$ where V is a finite set of *nodes* (or *vertices*) and E is a finite set of *edges*. One could also consider infinite node and edge sets which could be relevant for applications in time series. However, in this thesis we will avoid this complication and settle for the finite case. A graph can have edges of several types as introduced below. Every edge is *between* a pair of nodes (not necessarily distinct), that is, there is a known map which assigns an edge to a pair of nodes.

Edges, types and orientations We will consider graphs with *undirected* (—), *directed* (→), *blunt* ($\mapsfromchar\!\!\mapsto$), *bidirected* (↔), and *semidirected* (↦), edges, and we say that these are five edge *types*. The typography of the edges show if they are *symmetric* or not in the following sense. Let α and β be nodes in a graph. Undirected, blunt, and bidirected edges are symmetric in that α — β, $\alpha \mapsfromchar\!\!\mapsto \beta$, $\alpha \leftrightarrow \beta$ equal the edges β — α, $\beta \mapsfromchar\!\!\mapsto \alpha$, $\beta \leftrightarrow \alpha$, respectively. On the other hand, the edge $\alpha \rightarrow \beta$ is different from the edge $\alpha \leftarrow \beta$, and analogously for semidirected edges. We say that, e.g., edges $\alpha \rightarrow \beta$ and $\alpha \leftarrow \beta$ are of the same type, but have different *orientations*. We will throughout use $\alpha \sim \beta$ to denote a generic edge of any type (or of a context-specific, allowed subset of types). In the classes of graphs that we will consider, multiple edges between a pair of nodes can be allowed, however, only if they are of different types or orientations. Self-edges are also allowed. When $\mathcal{G} = (V, E)$ is a graph, $\alpha, \beta \in V$, we use $\alpha \sim_\mathcal{G} \beta$ to denote that this edge is in E, and $\alpha \not\sim_\mathcal{G} \beta$ to denote that it is not. We say that $\alpha \sim \alpha$ is a *loop*, or a *self-edge*.

Walks and paths A *walk* in a graph (V, E) is an alternating sequence of nodes and (oriented) edges,

$$\alpha_1 \overset{e_1}{\sim} \alpha_2 \overset{e_2}{\sim} \dots \overset{e_{n-1}}{\sim} \alpha_n \overset{e_n}{\sim} \alpha_{n+1},$$

such that $\alpha_i \in V$ for all $i = 1, \dots, n+1$ and $e_j \in E$ for all $j = 1, \dots, n$. We say that α_1 and α_{n+1} are *endpoints* or *endpoint nodes*. A *path* is a walk such that no node is repeated. A walk is *directed* if every edge is directed and points towards the same endpoint. A directed *cycle* is a directed path $\alpha \rightarrow \dots \rightarrow \beta$ along with an edge $\beta \rightarrow \alpha$.

Classes of dynamical graphs

We define the following classes of graphs that we will use to represent dynamical systems. In this thesis, we say that a graph is *dynamical* if it is a member of one of the following classes. Note that these are all subclasses of *directed mixed correlation graphs* (Definition 2.4).

In the next subsection, we describe some classes of graphs that we will say are *nondynamical*. These are really subclasses of dynamical classes so this distinction is somewhat artificial. We only use it to distinguish between classes of graphs that are meant to represent dynamical systems and classes of graphs that are not.

All these classes of graphs have been used in the literature on graphical models, some more than others, and we will give relevant references to previous work in subsequent sections and chapters.

Definition 2.1 (Directed graph, DG)**.** Let $\mathcal{G} = (V, E)$ be a graph. We say that $\mathcal{G}$ is a *directed graph* if every edge in E is directed.

Definition 2.2 (Directed correlation graph, cDG)**.** The graph $\mathcal{G}$ is a *directed correlation graph* if every edge is directed or blunt.

Definition 2.3 (Directed mixed graph, DMG)**.** We say that a graph is a *directed mixed graph* if every edge is directed or bidirected.

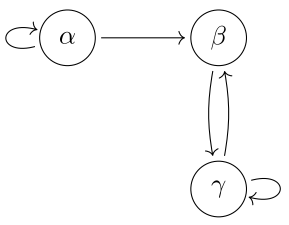


Figure 2.1: A directed graph on nodes $\{\alpha, \beta, \gamma\}$.

We could equivalently define a DMG as a graph such that every pair of nodes are connected by a subset of the edges $\{\alpha \to \beta, \alpha \leftarrow \beta, \alpha \leftrightarrow \beta\}$. The following is the largest class of graphs that we will consider in this thesis. In a sense which will be made mathematically precise in subsequent chapters, we will use directed edges to describe a direct influence of one process on another, bidirected edges to describe an unobserved confounding process, and blunt edges to describe a correlation between error processes. We will return to the interpretation of semidirected edges. The class of *directed mixed correlation graphs* combines all of these edge types and will allow us to represent partially observed dynamical systems driven by correlated error processes.

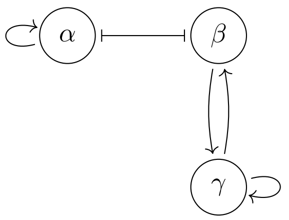


Figure 2.2: A directed correlation graph on nodes $\{\alpha, \beta, \gamma\}$.

Definition 2.4 (Directed mixed correlation graph, cDMG)**.** We say that $\mathcal{G}$ is a *directed mixed correlation graph* if its edges are directed ($\to$), bidirected ($\leftrightarrow$), semidirected ($\mapsto$), or blunt ($\vdash\!\dashv$).

Again, one can equivalently define the class of cDMGs as the graphs such that every pair of nodes $\{\alpha, \beta\}$ is joined by a subset of the edges $\{\alpha \to \beta, \alpha \leftarrow \beta, \alpha \leftrightarrow \beta, \alpha \mapsto \beta, \alpha \mapsfrom \beta, \alpha \vdash\!\dashv \beta\}$. Note that Definitions 2.1-2.4 all allow both the absence and presence of loops, $\alpha \sim \alpha$, of the appropriate types. In applications, it may be natural to impose restrictions on these classes of graphs, e.g., on loops that they may or must contain.

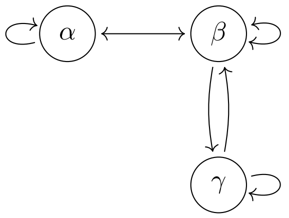


Figure 2.3: A directed mixed graph on nodes $\{\alpha, \beta, \gamma\}$.

Classes of nondynamical graphs

In the literature, it is not common to say that the following classes of graphs are *nondynamical*. However, we only use them for comparison with the above dynamical classes, and in this context the terminology seems befitting.

Definition 2.5 (Undirected graph)**.** Let $\mathcal{G} = (V, E)$ be a graph. We say that $\mathcal{G}$ is *undirected* if every edge in E is undirected.

While undirected graphs are interesting in their own right, we will mostly use them as technical tools, derived from other graphs. The two following definitions are restrictions of DGs and DMGs to disallow directed cycles and loops.

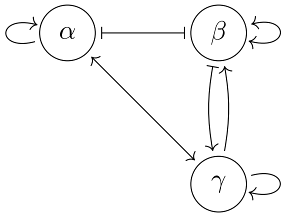


Figure 2.4: A directed mixed correlation graph on nodes $\{\alpha, \beta, \gamma\}$.

Definition 2.6 (Directed acyclic graph, DAG)**.** A *directed acyclic graph* is a DG with no directed cycles.

Note that by considering a trivial path directed, it follows that a loop, $\alpha \to \alpha$, creates a directed cycle from α to α, and therefore the definition of a DAG also excludes loops.

Definition 2.7 (Acyclic directed mixed graph, ADMG)**.** An *acyclic directed mixed graph* is a DMG with no directed cycles and no loops.

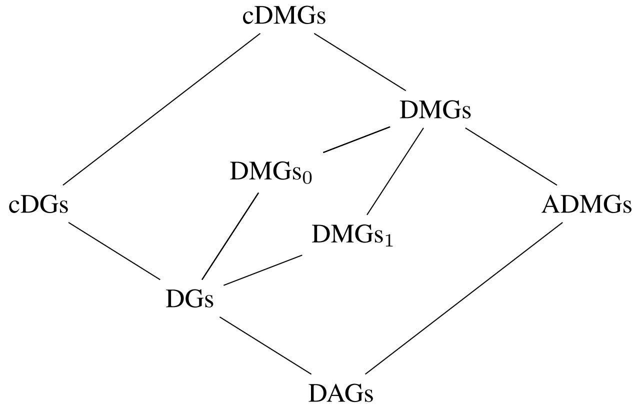


Figure 2.5: A Hasse diagram of classes of graphs. Consider a fixed node set V. Let cDMGs denote the set of cDMGs on nodes V, and similarly for the other classes of graphs. Edges denote set inclusion such that the set of graphs at the lower endpoint is a subset of the set of graphs at the upper endpoint. DMGs$_0$ and DMGs$_1$ are *reachable* and *reduced* DMGs and will be introduced in Chapter 3.

Separation

The graphs introduced above are all used for encoding an independence structure using a *separation criterion*.

Definition 2.8 (Separation criterion)**.** A separation criterion is a function which maps a graph, $\mathcal{G} = (V, E)$, and three sets $A, B, C \subseteq V$ (possibly under some restrictions on the sets) to *true* or *false*. We write a class of graphs, $\mathbb{G}$, with a separation criterion, s, as a pair $(\mathbb{G}, s)$.

Separation is defined for a graph and three subsets of its node set. When A or B are singletons, $A = \{\alpha\}, B = \{\beta\}$, we will often write α instead of $\{\alpha\}$ and β instead of $\{\beta\}$.

Definition 2.9 (Separation model)**.** Let $\mathcal{G}$ be a graph and s a separation criterion. We let $\mathcal{I}(\mathcal{G}, s)$, or just $\mathcal{I}(\mathcal{G})$, denote the collection of triplets of sets such that $s(\mathcal{G}, A, B, C)$ is true.

Definition 2.10 (Markov equivalence)**.** Consider a class of graphs with a separation criterion, s. We say that graphs $\mathcal{G}_1 = (V, E_1)$ and $\mathcal{G}_2 = (V, E_2)$ are *s-Markov equivalent* (or simply *Markov equivalent*) if for all allowed triplets of subsets, (A, B, C), it holds that $(A, B, C) \in \mathcal{I}(\mathcal{G}_1, s)$ if and only if $(A, B, C) \in \mathcal{I}(\mathcal{G}_2, s)$. If $(A, B, C) \in \mathcal{I}(\mathcal{G}, s)$, we say that B is s-separated from A by (or given) C in the graph $\mathcal{G}$. Note that separation does not need to be symmetric in arguments A and B.

Any separation criterion makes Markov equivalence an equivalence relation on a set of graphs with node sets V as it is reflexive (a graph is Markov equivalent with itself), symmetric, and transitive. However, one needs to fix the class of graphs within which one considers Markov equivalence. As an example, if we let $\mathcal{D}$ denote the complete DG on nodes V, i.e., $\alpha \to_{\mathcal{D}} \alpha$ for every $\alpha \in V$, we will later argue that no other DG is Markov equivalent with $\mathcal{D}$ when using the relevant notion of separation. However, within the class of DMGs, this is no longer the case as there exist DMGs that are Markov equivalent with $\mathcal{D}$. Most often the class within which to consider Markov equivalence will be obvious from the context.

Let $\mathcal{G}_1 = (V, E_1)$ and $\mathcal{G}_2 = (V, E_2)$. If $E_1 \subseteq E_2$, then we say that $\mathcal{G}_1$ is a *subgraph* of $\mathcal{G}_2$, and if $E_2 \subseteq E_1$, then we say that $\mathcal{G}_1$ is a *supergraph* of $\mathcal{G}_2$. In case of proper inclusion, we say that $\mathcal{G}_1$ is a *proper* subgraph and a *proper* supergraph, respectively. The separation criteria we will consider are monotone in the sense that more edges will lead to fewer separations. Under such monotonicity, we define the following.

Definition 2.11 (Maximality)**.** Let $(\mathbb{G}, s)$ be a class of graphs endowed with a separation criterion. We say that a graph, $\mathcal{G} = (V, E)$, is *maximal* within $\mathbb{G}$ if for any proper supergraph of $\mathcal{G}$, $\mathcal{G}^+ = (V, E^+) \in \mathbb{G}$,

$$\mathcal{I}(\mathcal{G}^+, s) \subsetneq \mathcal{I}(\mathcal{G}, s)$$

where $\subsetneq$ denotes proper set inclusion.

We can think of an equivalence class of graphs as a partially ordered set in which the partial order is induced by set inclusion of the edge sets. Using standard terminology of partially ordered sets, a *greatest* element is a graph in the Markov equivalence class which is a supergraph of all graphs in the equivalence class. A *least* element is a graph which is a subgraph of all elements in the equivalence class. A *maximal* element is a graph which is not a proper subgraph of any element in the equivalence class, and a *minimal* graph is not a proper supergraph of any element in the equivalence class. Greatest and least elements need not exist in every equivalence class. When these do exist, they are the unique maximal and minimal elements, respectively. We prove that equivalence classes of DMGs have a greatest element which we in that case call the maximal element without ambiguity.

Definition 2.12 (Simple graph)**.** We say that a graph $\mathcal{G} = (V, E)$ is *simple* if it has no loops and for each pair of nodes $\alpha, \beta \in V$, there is at most one edge between α and β.

When $\mathcal{G}$ is a graph, we let $G_A = (A, E_A)$ denote the subgraph induced by the set $A \subseteq V$ where $E_A = \{e \in E \mid \alpha \overset{e}{\sim} \beta,\ \alpha, \beta \in A\}$.

Definition 2.13 (Separability)**.** Let $\mathcal{G} = (V, E)$ be a graph, $\alpha, \beta \in V$, and let s be a separation criterion. We say that β is *s-separable* from α if there exists $C \subseteq V \setminus \{\alpha\}$ such that $(\mathcal{G}, \{\alpha\}, \{\beta\}, C)$ is in the domain of s and $s(\mathcal{G}, \{\alpha\}, \{\beta\}, C)$ is true, and otherwise we say that β is (s-)*inseparable* from α.

Definition 2.14 (Adjacency)**.** Let $\mathcal{G} = (V, E)$ be a graph, and let $\alpha, \beta \in V$. We say that α and β are *adjacent* if there exists $e \in E$ such that e is between α and β.

Definition 2.15 (Skeleton)**.** The *skeleton* of $\mathcal{G} = (V, E)$ is the undirected graph $\mathcal{U} = (V, F)$ such that for $\alpha, \beta \in V$, the edge $\alpha — \beta$ is in F if and only if α and β are adjacent in $\mathcal{G}$ (Definition 2.14). We denote the skeleton of $\mathcal{G}$ by $\text{sk}(\mathcal{G})$.

μ-separation

In dynamical graphs, we will mostly apply *μ-separation*. Before defining it, we need some more definitions. When $\mathcal{G} = (V, E)$, $\alpha, \gamma \in V$, we let $\gamma \to \ldots \to \alpha$ denote the existence of a directed path from γ to α. We define

$$\text{an}_{\mathcal{G}}(\alpha) = \{\gamma \in V : \gamma \to \ldots \to \alpha\}$$

and $\text{an}_{\mathcal{G}}(C) = \cup_{\alpha \in C} \text{an}_{\mathcal{G}}(\alpha)$. At times we omit the subscript and write simply $\text{an}(C)$. We say that $\gamma \in \text{an}(\alpha)$ is an *ancestor* of α, and we use the convention that a *trivial* path (a path with no edges) is directed which means that $C \subseteq \text{an}(C)$. We say that $\delta \to \varepsilon$ has a *tail* at δ and a *head* at ε. Consider a walk in a cDMG,

$$\alpha \sim \ldots \overset{e_1}{\sim} \gamma \overset{e_2}{\sim} \ldots \sim \beta.$$

We say that a nonendpoint node, γ, is a *noncollider* if e_1 or e_2 has a tail at γ, and otherwise we say that γ is a collider. Note that these properties are really properties of *instances* of a node on a walk. A node may be repeated on a walk and may both be a collider and a noncollider on a walk, as well as both an endpoint node and a nonendpoint node. We say that edges $\alpha \to \beta$, $\alpha \mapsto \beta$, and $\alpha \leftrightarrow \beta$ have a *head* at β.

Definition 2.16 (μ-separation, **A**.3.2)**.** Let $\mathcal{G} = (V, E)$ be a cDMG, and $\alpha, \beta \in V$. We say that a walk

$$\alpha \overset{e_1}{\sim} \gamma_1 \overset{e_2}{\sim} \ldots \overset{e_{n-1}}{\sim} \gamma_{n-1} \overset{e_n}{\sim} \beta$$

is *μ-connecting* from α to β given $C \subseteq V$ if it is nontrivial, $\alpha \notin C$, no noncollider is in C, every collider is in $\text{an}_{\mathcal{G}}(C)$, and the edge e_n has a head at β. We say that B is *μ-separated* from A by C, or in shorthand $A \perp_\mu B \mid C$, if there is no μ-connecting walk from any $\alpha \in A$ to any $\beta \in B$ in $\mathcal{G}$. Alternatively, we write $A \perp_\mu B \mid C\ [\mathcal{G}]$ to emphasize which graph the statement applies to.

μ-separation builds on the notions of δ-separation (Didelez, 2000, 2008) and δ^*-separation (Meek, 2014). We discuss the exact relationship in Appendix A of Paper **A**. In that paper, μ-separation is described in the class of DMGs, however, the extension to cDMGs is straightforward.

δ-, δ^*-, and μ-separation are not symmetric notions of separation. From the definition of μ-separation we see that a walk which is μ-connecting from α to β given C is not necessarily μ-connecting from β to α given C. This is the central difference between these notions of separation and d- and m-separation which are often used in nondynamical classes of graphs. We will see that this asymmetry allows us to model so-called *local independence* in multivariate stochastic processes as this independence relation is also asymmetric.

Graphical marginalization

We will be interested in using graphs to describe independence structure in systems where we only have partial observation in the sense that some coordinate processes are fully unobserved. For this purpose, we use *graphical marginalization*. Let $(\mathbb{G}, s)$ be a pair such that $\mathbb{G}$ is class of graphs and s is a separation criterion. For a graph $\mathcal{G} = (V, E) \in \mathbb{G}$ and a subset $O \subseteq V$, we can ask if there exists a graph $\mathcal{G}_O = (O, E_O)$ such that for all $A, B, C \subseteq O$

$$A \perp_s B \mid C\ [\mathcal{G}] \Leftrightarrow A \perp_s B \mid C\ [\mathcal{G}_O].$$

In this case, we say that $\mathcal{G}_O$ is a *marginal graph of $\mathcal{G}$ over O*, or a *marginalization* of $\mathcal{G}$ over O. If it is possible for any $\mathcal{G} = (V, E) \in \mathbb{G}$ and any $O \subseteq V$ to find such $\mathcal{G}_O \in \mathbb{G}$ satisfying the above, then we say that the pair $(\mathbb{G}, s)$ is *closed under marginalization*.

Directed graphs

Directed graphs (Definition 2.1) are the fundamental graphical objects of this thesis and much of the theory revolves around this class of graph. Let us consider a fixed node set V. It is clear that knowing the edge set uniquely determines the μ-separations of the graph. The following proposition shows that the opposite also holds true, and we say that the separation model *identifies* the DG. The proposition is similar to Proposition 3.6 in Paper **A**.

Proposition 2.17. Let $\mathcal{D} = (V, E)$ be a DG such that $\alpha, \beta \in V$. Then $\alpha \perp_\mu \beta \mid V \setminus \{\alpha\}$ if and only if there is no directed edge from α to β.

Proof. Assume first that there is no directed edge from α to β and consider a walk from α to β. This walk must have length at least two, $\alpha \sim \ldots \sim \gamma \to \beta$, and $\gamma \neq \alpha$. We see that this walk is closed as γ is in the conditioning set. On the other hand, if there is a directed edge, then $\alpha \to \beta$ is a μ-connecting walk from α to β given $V \setminus \{\alpha\}$. □

Corollary 2.18. If $\mathcal{D}_1 = (V, E_1)$ and $\mathcal{D}_2 = (V, E_2)$ are DGs, then they are Markov equivalent if and only if they are equal.

The corollary is a simple consequence of Proposition 2.17. It follows that every Markov equivalence class of a DG is a singleton within the class of DGs. However, if we consider a DG as a DMG, then this will not hold anymore. For instance, the complete DG on nodes V and the complete DMG on nodes V (i.e., the graph such that $\alpha \to \beta$ and $\alpha \leftrightarrow \beta$ for all $\alpha, \beta \in V$) are Markov equivalent. The class of DGs are in many ways, as we will see, analogous to DAGs. Corollary 2.18 is a simple result, however, a similar statement does not hold for DAGs equipped with d-separation.

Directed mixed correlation graphs

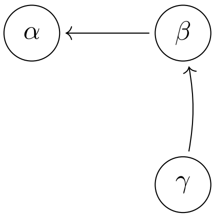


Figure 2.6: There is a μ-connecting route from α to β given $\emptyset$, but no μ-connecting path.

In this section, we will look further into the large class of *directed mixed correlation graphs* (cDMGs) and describe various properties of these graphs. In the next chapter, we will look at the subclass of *directed mixed graphs* (DMGs) in which we can obtain some interesting and stronger results that do not hold in the class of cDMGs.

The set of walks in a cDMG, $\mathcal{G}$, uniquely determines the μ-separation model of the graph, $\mathcal{I}(\mathcal{G}, \mu)$. However, this set is infinite unless the graph has no edges. One can ask if there exists a true subset of walks that also identifies the separations. To show that this is the case, we can use *routes*.

Definition 2.19 (Route, Paper **A**)**.** A *route* from α to β is a walk

$$\alpha \sim \gamma_1 \sim \ldots \sim \gamma_k \sim \beta$$

such that no node different from β is repeated on the walk and such that β occurs at most twice.

If we let $\mathcal{P}(\mathcal{G}), \mathcal{R}(\mathcal{G}), \mathcal{W}(\mathcal{G})$ denote the paths, routes, and walks, respectively, of a cDMG, $\mathcal{G}$, then

$$\mathcal{P}(\mathcal{G}) \subsetneq \mathcal{R}(\mathcal{G}) \subsetneq \mathcal{W}(\mathcal{G})$$

unless $\mathcal{G}$ has no edges. The following proposition shows that routes completely characterize the μ-separations in a cDMG. Paper **A** states the same result in the class of DMGs (Proposition **A**.3.5).

Proposition 2.20 (μ-connecting routes)**.** Let $\mathcal{G} = (V, E)$ be a cDMG, and let $\alpha, \beta \in V, C \subseteq V$. There is a μ-connecting route from α to β given C if and only if there is a μ-connecting walk from α to β given C.

Proof. The proof of this is identical to the proof in the case of DMGs (proof of Proposition **A**.3.5 in the supplementary material of Paper **A**). □

The example in Figure 2.6 shows that paths cannot be used for characterizing μ-separation models. This is different from the nondynamical graphs where d- and m-separation can be defined using either paths or walks. Note that an edge can be repeated on a route, but only in the configuration $\gamma_{k-1} \overset{e_1}{\sim} \gamma_k \overset{e_2}{\sim} \beta$ such that e_1 and e_2 is the same edge and $\gamma_{k-1} = \beta$.

Maximality and adjacency

For some classes of graphs and separation criteria, one can show that in a maximal graph (Definition 2.11) every pair of inseparable nodes are adjacent, and this has also been used as a definition of *maximality* of a graph (Richardson and Spirtes, 2002). In the case of ancestral graphs (with m-separation) it is then a theorem that an ancestral graph is maximal if and only if no edges can be added Markov equivalently (Richardson and Spirtes, 2002). In the case of cDMGs with μ-separation a graph may be maximal, i.e., no edge can be added without changing the independence model, yet there exists a pair of nonadjacent nodes, (α, β), such that β is not separable from α (see Example **A**.4.12).

Properties of cDMGs

In this section, we will see that the class of cDMGs is closed under marginalization and that, loosely speaking, the fact that cDMGs contain cDGs as a subclass lead to semidirected edges, $\mapsto$, via marginalization. We start, however, by noting that there is a certain hierarchy between edges of the types $\alpha \leftrightarrow \beta, \alpha \mapsto \beta$, and $\alpha \mathrel{\vdash\!\dashv} \beta$ such that adding edges lower in the hierarchy will not change the separation model.

Proposition 2.21. Let $\mathcal{G} = (V, E)$ be a cDMG, and let $\alpha, \beta \in V$. If $\alpha \leftrightarrow_{\mathcal{G}} \beta$, then adding $\alpha \mapsto \beta$ gives a Markov equivalent graph.

Proposition 2.22. Let $\mathcal{G} = (V, E)$ be a cDMG, and let $\alpha, \beta \in V$. If $\alpha \mapsto_{\mathcal{G}} \beta$, then adding $\alpha \mathrel{\vdash\!\dashv} \beta$ gives a Markov equivalent graph.

Proof of Propositions 2.21 and 2.22. Let $\mathcal{G}^+ = \mathcal{G} + e$ denote the larger graph. We just need to argue that if there is a μ-connecting walk from α to β given C in $\mathcal{G}^+$, then we can also find a μ-connecting walk from α to β given C in $\mathcal{G}$. If e is not on the walk, then it is also connecting in $\mathcal{G}$ as e does not change the ancestry. If e is on the graph, then substitute it in each instance with $\alpha \leftrightarrow \beta$ or $\alpha \mapsto \beta$ to obtain a connecting walk. □

In cDMGs, the tip (closest to α) of an edge can either be a tail $\alpha \to \beta$, a head $\alpha \leftarrow \beta$, $\alpha \leftrightarrow \beta$, $\alpha \mapsfrom \beta$, or a stump $\alpha \mathrel{\vdash\!\dashv} \beta$, $\alpha \mapsto \beta$. Jointly, we call tails, heads, and stumps *edge marks*. We say that walks

$$\alpha \overset{e_1^\alpha}{\sim} \gamma_1^1 \sim \dots \gamma_1^{k_1} \overset{e_1^\beta}{\sim} \beta$$
$$\alpha \overset{e_2^\alpha}{\sim} \gamma_2^1 \sim \dots \gamma_2^{k_2} \overset{e_2^\beta}{\sim} \beta$$

are *endpoint-identical* if e_1^α and e_2^α have the same edge mark at α and e_1^β and e_2^β have the same edge mark at β. We say that an edge, e, between α and β, is endpoint-identical with a walk, ω, from α to β, if ω is endpoint-identical with the walk

$$\alpha \overset{e}{\sim} \beta.$$

Latent projection is a form of graphical marginalization that has been used in different classes of graphs (Verma and Pearl, 1991; Koster, 1999; Sadeghi, 2013; Richardson et al., 2017). Paper **A** uses it in DMGs and below we define a latent projection in cDMGs. Eichler (2012) studies a class of graphs very similar to cDMGs and defines a latent projection in this class of graphs.

Definition 2.23 (Latent projection)**.** Let $\mathcal{G} = (V, E)$ be a cDMG, and let $O \subseteq V$. The *latent projection* of $\mathcal{G}$ on O is the cDMG (O, F) such that for all $\alpha, \beta \in O$, the edge $\alpha \overset{e}{\sim} \beta$ is in F if and only if there is an endpoint-identical (and nontrivial) walk between α and β in $\mathcal{G}$ with no colliders and such that every nonendpoint node is in $V \setminus O$. We denote the latent projection of $\mathcal{G}$ on O by $m(\mathcal{G}, O)$.

Note that the latent projection is always a cDMG. The cDMGs are therefore closed under latent projection, and we will see that they can represent independence structure in partially observed dynamical systems that are driven by correlated noise. Eichler (2013) used them for the same purpose.

The following theorem justifies thinking about the latent projection as a graphical marginalization of a cDMG and a similar result was provided by Eichler (2013). We will provide a proof, though, as the formalism we use is somewhat different.

Theorem 2.24. Let $\mathcal{G} = (V, E)$ be a cDMG and let $O \subseteq V$. Let $\mathcal{M} = m(\mathcal{G}, O)$ denote the latent projection of $\mathcal{G}$ on O. For all $A, B, C \subseteq O$,

$$A \perp_\mu B \mid C \; [\mathcal{G}] \Leftrightarrow A \perp_\mu B \mid C \; [\mathcal{M}].$$

Proof. Theorem **A**.3.12 gives this result in case of a DMG, $\mathcal{G}$. The same proof applies to this more general case. The proof in the supplementary material of Paper **A** uses Propositions **A**.3.3 and **A**.3.11. Those results also hold in cDMGs, and the same proofs apply. □

Loops in cDMGs We argue that directed loops, $\alpha \to \alpha$, and bidirected loops, $\alpha \leftrightarrow \alpha$, are interchangeable in DMGs, but not in cDMGs in general. In the case of DMGs, we can first note that directed loops never change the ancestry of the graph (which holds in general cDMGs as well). μ-separation is characterized by routes, so consider a route in a DMG. If a loop is present on the route, then it must be the final edge. This means that either the route has length 1, in which case it is immaterial if the loop is directed or bidirected, or

$$\alpha \sim \ldots \gamma \overset{e}{\sim} \beta \sim \beta.$$

If e has a head at β, then the subroute from α to the first instance of β is connecting. If it has a tail instead, then the type of loop at β does not matter for the connectivity. This means that whenever there is a directed loop at β in a DMG, we can Markov equivalently add a bidirected loop at β, and vice versa. This is no longer true in general cDMGs. Consider simply the graph

$$\alpha \mathrel{\vdash\!\!\dashv} \beta.$$

If we add a directed loop at β, then $\alpha \perp_\mu \beta \mid \beta$. This is not true if we add a bidirected loop. On the other hand, if we add a bidirected loop at β, then $\alpha \perp_\mu \beta \mid \varnothing$, and this is not true if we add a directed loop.

Computational complexity

In this section, we give a short and informal introduction to the computational complexity theory that we need in this thesis. We also relate this topic to the graphs described above and the questions that graphical modeling asks about such graphs, especially in relation to separation and Markov equivalence. More background can be found in Garey and Johnson (1979); Goldreich (2010); Sipser (2013).

Complexity classes We say that an algorithm is of *polynomial time* if we can bound its worst-case running time by a polynomial function in the size of the input. A *decision problem* is a computational problem such that the answer is *yes* or *no*, in contrast to, e.g., optimization problems. The complexity class **P** consists of the decision problems that can be solved by a deterministic Turing machine (an abstract computer) in polynomial time. **NP** is the set of decision problems such that the yes-instances have *certificates* that can be verified in polynomial time by a deterministic Turing machine. As an example of a problem in **NP**, we can take separability; given nodes α and β in a DMG, does there exists a $C \subseteq V \setminus \{\alpha\}$ such that β is μ-separated from α by C? This problem is seen to be in **NP**: in a yes-instance (i.e., they are separable) there exists a separating set C_0, and given this (polynomially sized) certificate, C_0, we will show later that we can verify that β is μ-separable from α by testing $\alpha \perp_\mu \beta \mid C_0$ in polynomial time. In fact, this problem is also in **P**.

coNP is the set of decision problems such that the no-instances have certificates that can be verified in polynomial time. As an example of a problem in **coNP**, we can consider Markov equivalence; given two DMGs, are they Markov equivalent? In a no-instance, there is some triplet (A, B, C) such that B is μ-separated from A given C in one graph, but not in the other. Given a certificate indicating these three sets, we can again easily (i.e., in polynomial time) verify that this is a no-instance. On the other hand, in a yes-instance, there is no obvious certificate which allows us to verify Markov equivalence in polynomial time.

Encodings Above we have tacitly used a *size* of the input without specifying what this means. In our case, the input will often be a graph in which case we can think of the size of the input as simply the number of nodes. Formally, it is the size of a string over some alphabet, however, the precise encoding is not important and any 'reasonable' encoding will do (Sipser, 2013, Chapter 7).

Reductions A *reduction* is a function that transforms one problem, A, into another, B. If this transformation is easily computable, i.e., in polynomial time, and yes/no instances are preserved under the transformation, then we can use an algorithm for solving B to also solve A. Assume we have access to an oracle for the problem B, i.e., a mechanism that will provide us with the correct answer to an instance of problem B. A *many-one* reduction (also known as a *Karp* reduction) is a particularly simple type of polynomial-time reduction which transforms an instance of problem A into an instance of problem B and returns the solution (yes/no) to problem B as obtained from the oracle. A *Turing* reduction (also known as a *Cook* reduction) is a more general reduction which as a subroutine can query the oracle multiple times, though only polynomially many, and which is a polynomial-time algorithm outside the calls to the oracle.

In the supplementary material of Paper **A**, it is shown that one can decide μ-separation in a DMG by deciding separation in an auxiliary undirected graph. Similar results hold in other classes of graphs (Lauritzen, 1996; Didelez, 2000; Richardson and Spirtes, 2002; Richardson, 2003). This transformation can be done in polynomial time, and it gives an example of a many-one reduction, reducing the problem of deciding μ-separation in DMGs to deciding separation in an undirected graph.

Hardness We say that a problem is NP-*hard* if it is as hard as any problem in **NP**, more precisely, if any problem in **NP** can be reduced in polynomial time to this problem using a many-one reduction. If any problem in **NP** is reducible to a problem using a Turing reduction then we say that the problem is *Turing* NP-hard. We say that a problem is NP-*complete* if it is in **NP** and NP-hard. Analogously, we define a problem to be coNP-*hard* if it is at least as hard as the hardest problems in **coNP**, and to be coNP-*complete* if it is coNP-hard and also in **coNP**. The class **P** is a subclass of **NP** and of **coNP**. It is generally believed (though not proven) that **P** ≠ **NP** and **P** ≠ **coNP** and this would imply that there are no polynomial-time algorithms for solving NP- and coNP-hard problems.

Computational problems We end this section by defining the problems that we will consider later in the thesis. First, the following decision problem for a class of graphs $\mathbb{G}$ and a separation criterion s.

Markov equivalence in $(\mathbb{G}, s)$
Instance: $\mathcal{G}_1 = (V, E_1),\ \mathcal{G}_2 = (V, E_2) \in \mathbb{G}$
Question: Is $\mathcal{I}(\mathcal{G}_1, s) = \mathcal{I}(\mathcal{G}_2, s)$?

Similarly, we define the following search problems.

Minimal Markov equivalent graph in $(\mathbb{G}, s)$
Instance: $\mathcal{G} = (V, E) \in \mathbb{G}$
Question: Find a minimal $\mathcal{G}^-$ such that $\mathcal{G}^- \in [\mathcal{G}]$

Maximal Markov equivalent graph in $(\mathbb{G}, s)$
Instance: $\mathcal{G} = (V, E) \in \mathbb{G}$
Question: Find a maximal $\mathcal{G}^+$ such that $\mathcal{G}^+ \in [\mathcal{G}]$

Learn maximal Markov equivalent graph in $(\mathbb{G}, s)$
Instance: An oracle for $\mathcal{I}(\mathcal{G})$ such that $\mathcal{G} = (V, E) \in \mathbb{G}$
Question: Find a maximal graph $\mathcal{G}^+$ such that $\mathcal{I}(\mathcal{G}^+) = \mathcal{I}(\mathcal{G})$

Smallest Markov equivalent graph in $(\mathbb{G}, s)$
Instance: $\mathcal{G} = (V, E) \in \mathbb{G}$
Question: Find a $\mathcal{G}^- = (V, E^-) \in [\mathcal{G}]$ such that for all $\tilde{\mathcal{G}} = (V, \tilde{E}) \in [\mathcal{G}]$, it holds that $|E^-| \leq |\tilde{E}|$

Smallest Markov equivalent subgraph in $(\mathbb{G}, s)$
Instance: $\mathcal{G} = (V, E) \in \mathbb{G}$
Question: Find a $\mathcal{G}^- = (V, E^-) \in [\mathcal{G}],\ \mathcal{G}^- \subseteq \mathcal{G}$, such that for all $\tilde{\mathcal{G}} = (V, \tilde{E}) \in [\mathcal{G}],\ \tilde{\mathcal{G}} \subseteq \mathcal{G}$, it holds that $|E^-| \leq |\tilde{E}|$

For these search problems, we can define analogous decision problems by asking if there exists a solution with less/more than k edges. As the very last step in this chapter, we introduce two well-known hard problems. Let $X = \{x_1, \ldots, x_n\}$ be a set of Boolean variables. A *literal* is either a variable, x_l, or its negation, $\neg x_l$. A *term* is a set of literals $\{z_1, \ldots, z_k\}$ which represents their conjunction,

$$z_1 \wedge \ldots \wedge z_k.$$

A *3-term* is a term consisting of at most three literals. A Boolean formula which is a disjunction of 3-terms,

$$(z_1^1 \wedge z_2^1 \wedge z_3^1) \vee \ldots \vee (z_1^N \wedge z_2^N \wedge z_3^N),$$

is said to be in *3-disjunctive normal form* (3DNF). N is the number of terms.

3DNF tautology
Instance: A disjunction of 3-terms over variables X
Question: Does the formula evaluate to 1 for all inputs?

3DNF tautology is known to be coNP-complete (Garey and Johnson, 1979). We will also use the following problem. We consider a *universe*, $\mathcal{U} = \{u_1, \ldots, u_n\}$, that is, a finite set, and a family, $\mathcal{F}$, of subsets of $\mathcal{U}$ such that $\cup_{S \in \mathcal{F}} S = \mathcal{U}$. We say that $\mathcal{C} \subseteq \mathcal{F}$ is a *covering* if $\cup_{S \in \mathcal{C}} S = \mathcal{U}$. The *size* of $\mathcal{C} = \{S_1, \ldots, S_m\}$ is its cardinality, m.

Set-covering problem
Instance: A universe, $\mathcal{U}$, and a collection, $\mathcal{F}$, of subsets of $\mathcal{U}$ such that $\cup_{S \in \mathcal{F}} S = \mathcal{U}$, $k \in \mathbb{N}$
Question: Does there exists a covering of size at most k?

The set-covering problem is NP-complete (Cormen et al., 2009).

Chapter 3

Directed mixed graphs

This chapter goes into depth with *directed mixed graphs* (DMGs). We use these graphs to represent *local independence* in stochastic processes in which we only have partial observation, that is, some coordinate processes are unobserved. Local independence is an asymmetric independence relation which will also be introduced in this chapter. We first compare DMGs with the larger class of cDMGs. Paper **A** then describes DMGs in more details and proves a central result on Markov equivalence in this class of graphs. Some more results are presented after the paper, including a description of *reduced directed mixed graphs* that constitute simpler graphical means for representing the independence models encoded by the class of DMGs.

Definition 3.1 (Canonical DMG)**.** Let $\mathcal{G} = (V, E)$ be a cDMG. The *canonical DMG* of $\mathcal{G}$, $\mathcal{B}(\mathcal{G}) = (V, E_{\mathcal{B}})$, is the DMG obtained by changing every blunt or semidirected edge to a bidirected edge. That is, for all $\alpha, \beta \in V$,

$$\begin{aligned} \alpha \rightarrow_{\mathcal{B}(\mathcal{G})} \beta \quad &\text{if and only if} \quad \alpha \rightarrow_{\mathcal{G}} \beta, \\ \alpha \leftrightarrow_{\mathcal{B}(\mathcal{G})} \beta \quad &\text{if and only if} \quad \alpha \leftrightarrow_{\mathcal{G}} \beta, \alpha \mapsto_{\mathcal{G}} \beta, \alpha \leftarrow\!\dashv_{\mathcal{G}} \beta, \text{ or } \alpha \vdash\!\dashv_{\mathcal{G}} \beta. \end{aligned}$$

The canonical DMG is a coarser description of the separation model in the following sense.

Proposition 3.2. Let $\mathcal{G}$ be a cDMG, and let $\mathcal{B}(\mathcal{G})$ be its canonical DMG. It holds that $\mathcal{I}(\mathcal{B}(\mathcal{G}), \mu) \subseteq \mathcal{I}(\mathcal{G}, \mu)$.

Proof. Propositions 2.21 and 2.22 give the result as we can add blunt and semidirected edges Markov equivalently to the canonical DMG to obtain a supergraph of $\mathcal{G}$. □

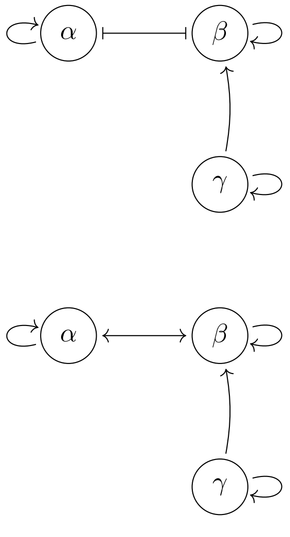


Figure 3.1: A cDG (top) and its canonical DMG (bottom). See Example 3.3.

Example 3.3. We illustrate that not every independence encoded in a cDMG is retained in its canonical DMG. We consider the cDG in Figure 3.1 and its canonical DMG. The two are not Markov equivalent as β is μ-separated from α by $\{\beta, \gamma\}$ in the cDG whereas this is not the case in the canonical DMG. We can ask if there is any DMG on nodes $\{\alpha, \beta, \gamma\}$ which is Markov equivalent with this cDG. If so, then α and β cannot be adjacent in this DMG as this would make one of them inseparable from the other. Similarly, for α and γ. In the cDG, α is not separated from γ given $\{\beta\}$, and this means that β must be a collider on a path between α and γ. In this case, β is inseparable from α in the DMG. This is a contradiction, and it shows that no such Markov equivalent DMG exists. In conclusion, the independence models of cDGs are not contained in the independence models of DMGs and therefore it follows that the independence models of cDMGs are a proper superset of those of DMGs.

The above example illustrates that the restriction to DMGs is in fact reducing the expressive power of the graphs in terms of which μ-separation structures they may represent. However, this restriction also enables stronger results. Paper **A** studies directed mixed graphs and the central result shows that every equivalence class has a greatest element. This is useful as this fact allows us to visualize and understand some aspects of the entire equivalence class very easily. This result does not hold in the larger class of cDMGs, and in fact, it does not even hold for cDGs as shown in Paper **B**.

Paper **A**

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

+ supplementary material

The Annals of Statistics
2020, Vol. 48, No. 1, 539–559
https://doi.org/10.1214/19-AOS1821

MARKOV EQUIVALENCE OF MARGINALIZED LOCAL INDEPENDENCE GRAPHS

BY SØREN WENGEL MOGENSEN[*] AND NIELS RICHARD HANSEN[**]

Department of Mathematical Sciences, University of Copenhagen, [*]swengel@math.ku.dk; [**]niels.r.hansen@math.ku.dk

Symmetric independence relations are often studied using graphical representations. Ancestral graphs or acyclic directed mixed graphs with m-separation provide classes of symmetric graphical independence models that are closed under marginalization. Asymmetric independence relations appear naturally for multivariate stochastic processes, for instance, in terms of local independence. However, no class of graphs representing such asymmetric independence relations, which is also closed under marginalization, has been developed. We develop the theory of directed mixed graphs with μ-separation and show that this provides a graphical independence model class which is closed under marginalization and which generalizes previously considered graphical representations of local independence.

Several graphs may encode the same set of independence relations and this means that in many cases only an equivalence class of graphs can be identified from observational data. For statistical applications, it is therefore pivotal to characterize graphs that induce the same independence relations. Our main result is that for directed mixed graphs with μ-separation each equivalence class contains a maximal element which can be constructed from the independence relations alone. Moreover, we introduce the directed mixed equivalence graph as the maximal graph with dashed and solid edges. This graph encodes all information about the edges that is identifiable from the independence relations, and furthermore it can be computed efficiently from the maximal graph.

1. Introduction. Graphs have long been used as a formal tool for reasoning with independence models. Most work has been concerned with symmetric independence models arising from standard probabilistic independence for discrete or real-valued random variables. However, when working with dynamical processes it is useful to have a notion of independence that can distinguish explicitly between the present and the past, and this is a key motivation for considering local independence.

The notion of local independence was introduced for composable Markov processes by Schweder [37] who also gave examples of graphs describing local independence structures. Aalen [1] discussed how one could extend the definition of local independence in the broad class of semimartingales using the Doob–Meyer decomposition. Several authors have since then used graphs to represent local independence structures of multivariate stochastic process models, in particular for point process models; see, for example, [4, 11–13, 35]. Local independence takes a dynamical point of view in the sense that it evaluates the dependence of the present on the past. This provides a natural link to statistical causality as cause must necessarily precede effect [1, 2, 28, 37]. Furthermore, recent work argues that for some applications it can be important to consider continuous-time models, rather than only cross-sectional models, when trying to infer causal effects [3].

Received February 2018; revised January 2019.

MSC2010 subject classifications. Primary 62M99; secondary 62A99.

Key words and phrases. Directed mixed graphs, independence model, local independence, local independence graph, Markov equivalence, μ-separation.

Local independence for point processes has been applied for data analysis (see, e.g., [2, 23, 44]), but in applications a direct causal interpretation may be invalid if only certain dynamical processes are observed while other processes of the system under study are unobserved. Allowing for such latent processes is important for valid causal inference, and this motivates our study of representations of marginalized local independence graphs.

Graphical representations of independence models have also been studied for time series [14–17]. In the time series context—using the notion of Granger causality—Eichler [15] gave an algorithm for learning a graphical representation of local independence. However, the equivalence class of graphs that yield the same local independences was not identified, and thus the learned graph does not have any clear causal interpretation. Related research has been concerned with inferring the graph structure from subsampled time series, but under the assumption of no latent processes; see, for example, [9, 22].

In this paper, we give a formal, graphical framework for handling the presence of unobserved processes and extend the work on graphical representations of local independence models by formalizing marginalization and giving results on the equivalence classes of such graphical representations. The graphical framework that we propose is a generalization of that of Didelez [11–13]. This development is analogous to work on marginalizations of graphical models using directed acyclic graphs, DAGs. Starting from a DAG, one can find graphs (e.g., maximal ancestral graphs or acyclic directed mixed graphs) that encode marginal independence models [8, 18, 19, 25, 33, 34, 36, 39]. One can then characterize the equivalence class of graphs that yield the same independence model [5, 45]—the so-called Markov equivalent graphs—and construct learning algorithms to find such an equivalence class from data. The purpose of this paper is to develop the necessary theoretical foundation for learning local independence graphs by developing a precise characterization of the learnable object: the class of Markov equivalent graphs.

The paper is structured as follows: in Section 2, we discuss abstract independence models, relevant graph-theoretical concepts and the notion of local independence and local independence graphs. In Section 3, we introduce μ-separation for directed mixed graphs, which will be used to represent marginalized local independence graphs, and we describe an algorithm to marginalize a given local independence graph. In Sections 4 and 5, we develop the theory of μ-separation for directed mixed graphs further, and we discuss, in particular, Markov equivalence of such graphs. All proofs of the main paper are given in the Supplementary Material [29]. Sections A to F are in the Supplementary Material.

2. Independence models and graph theory. Graphical separation criteria as well as probabilistic models give rise to abstract conditional independence statements. Graphical modeling is essentially about relating graphical separation to probabilistic independence. We will consider both as instances of abstract *independence models*.

Consider some set $\mathcal{S}$. An *independence model*, $\mathcal{I}$, on $\mathcal{S}$ is a set of triples (A, B, C) where $A, B, C \in \mathcal{S}$, that is, $\mathcal{I} \subseteq \mathcal{S} \times \mathcal{S} \times \mathcal{S}$. Mathematically, an independence model is a ternary relation. In this paper, we will consider independence models *over* a finite set V which means that $\mathcal{S} = \mathcal{P}(V)$, the power set of V. In this case, an independence model $\mathcal{I}$ is a subset of $\mathcal{P}(V) \times \mathcal{P}(V) \times \mathcal{P}(V)$. We will call an element $s \in \mathcal{P}(V) \times \mathcal{P}(V) \times \mathcal{P}(V)$ an *independence statement* and write s as $\langle A, B \mid C\rangle$ for $A, B, C \subseteq V$. This notation emphasizes that s is thought of as a statement about A and B conditionally on C.

Graphical and probabilistic independence models have been studied in very general settings, though mostly under the assumption of symmetry of the independence model, that is,

$$\langle A, B \mid C\rangle \in \mathcal{I} \quad \Rightarrow \quad \langle B, A \mid C\rangle \in \mathcal{I};$$

see, for example, [7, 10, 26] and references therein. These works take an abstract axiomatic approach by describing and working with a number of properties that hold in, for example, models of conditional independence. In this paper, we consider independence models that do not satisfy the symmetry property as will become evident when we introduce the notion of local independence.

2.1. *Local independence.* We consider a real-valued, multivariate stochastic process

$$X_t = (X_t^1, X_t^2, \dots, X_t^n), \quad t \in [0, T]$$

defined on a probability space $(\Omega, \mathcal{F}, P)$. In this section, the process is a continuous-time process indexed by a compact time interval. The case of a discrete time index, corresponding to $X = (X_t)$ being a time series, is treated in Section C in the Supplementary Material. We will later identify the coordinate processes of X with the nodes of a graph; hence, both are indexed by $V = \{1, 2, \dots, n\}$. As illustrated in Example 2.3 below, the index set may be chosen in a more meaningful way for a specific application. In that example, $X_t^I \geq 0$ is a price process, $X_t^L \in \mathbb{N}_0$ is a counting process of events, and the remaining four processes take values in $\{0, 1\}$ indicating if an individual at a given time is a regular user of a given substance. Figure 1 shows examples of sample paths for three individuals.

To avoid technical difficulties, irrelevant for the present paper, we restrict attention to right-continuous processes with coordinates of finite and integrable variation on the interval $[0, T]$. This includes most nonexplosive multivariate counting processes as an important special case, but also other interesting processes such as piecewise-deterministic Markov processes.

To define local independence below, we need a mathematical description of how the stochastic evolution of one coordinate process depends infinitesimally on its own past and the past of the other processes. To this end, let $\mathcal{F}_t^{C,0}$ denote the σ-algebra generated by $\{X_s^\alpha : s \leq t, \alpha \in C\}$ for $C \subseteq V$. For technical reasons, we need to enlarge this σ-algebra, and we define $\mathcal{F}_t^C$ to be the completion of $\bigcap_{s>t} \mathcal{F}_s^{C,0}$ w.r.t. P. Thus $(\mathcal{F}_t^C)$ is a right-continuous and complete filtration which represents the history of the processes indexed by $C \subseteq V$ until time t. Figure 2 illustrates, in the context of Example 2.3, the filtrations $\mathcal{F}_t^V$, $\mathcal{F}_t^{\{L,M,H\}}$ and $\mathcal{F}_t^{\{T,A,M,H\}}$.

For $\beta \in V$ and $C \subseteq V$, let $\Lambda^{C,\beta}$ denote an $\mathcal{F}_t^C$-predictable process of finite and integrable variation such that

$$E(X_t^\beta \mid \mathcal{F}_t^C) - \Lambda_t^{C,\beta}$$

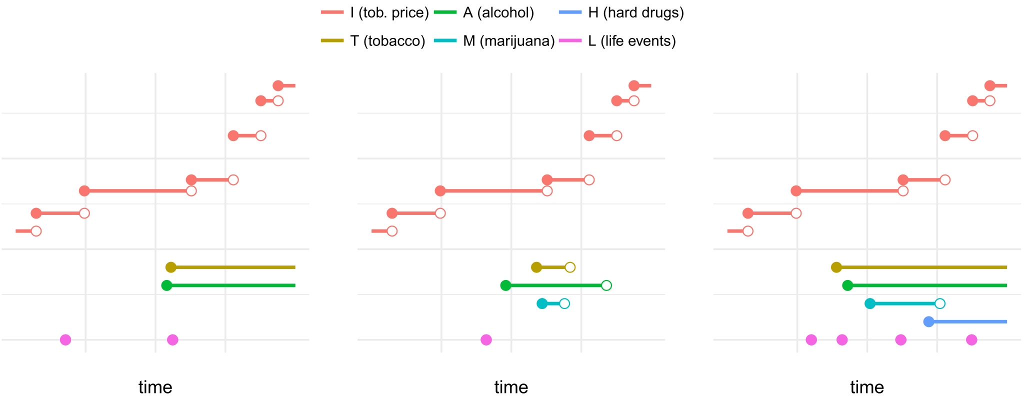


FIG. 1. *Sample paths for three individuals of the processes considered in Example 2.3. The price process (I) is a piecewise constant jump process and the life event process (L) is illustrated by the event times. The remaining four processes are illustrated by the segments of time where the individual is a regular user of the substance. The absence of a process, for example, the hard drug process (H) in the left and middle samples, means that the individual never used that substance.*

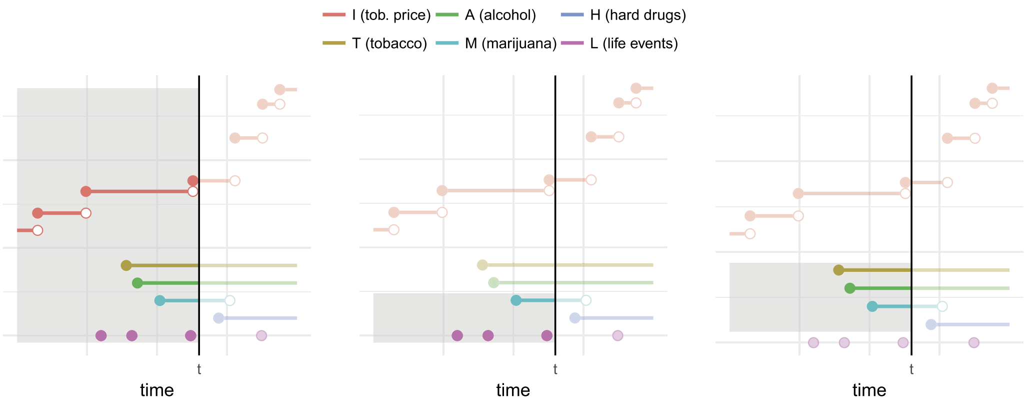


FIG. 2. *Illustration of the past at time t as captured by different filtrations for a single sample path of processes from Example 2.3. The filtration $\mathcal{F}_t^V$ (left) captures the past of all processes, while $\mathcal{F}_t^{\{L,M,H\}}$ (middle) captures the past of L, M and H only, and $\mathcal{F}_t^{\{T,A,M,H\}}$ (right) captures the past of T, A, M and H.*

is an $\mathcal{F}_t^C$ martingale. Such a process exists (see Section E for the technical details), and is usually called the compensator or the dual predictable projection of $E(X_t^\beta \mid \mathcal{F}_t^C)$. It is in general unique up to evanescence.

DEFINITION 2.1 (Local independence). Let $A, B, C \subseteq V$. We say that X^B is *locally independent* of X^A given X^C if there exists an $\mathcal{F}_t^C$-predictable version of $\Lambda^{A \cup C, \beta}$ for all $\beta \in B$. We use $A \not\to B \mid C$ to denote that X^B is locally independent of X^A given X^C.

In words, the process X^B is locally independent of X^A given X^C if, for each time point, the past up until time t of X^C gives us the same *predictable* information about $E(X_t^\beta \mid \mathcal{F}_t^{A\cup C})$ as the past of $X^{A\cup C}$ until time t. Note that when $\beta \in C$, $E(X_t^\beta \mid \mathcal{F}_t^C) = X_t^\beta$.

Local independence was introduced by Schweder [37] for composable Markov processes and extended by Aalen [1]. Local independence and graphical representations thereof were later considered by Didelez [11–13] and by Aalen et al. [4]. Didelez [12] also discussed local independence models of composable finite Markov processes under some specific types of marginalization. Commenges and Gégout-Petit [6, 21] discussed definitions of local independence in classes of semimartingales. Note that Definition 2.1 allows a process to be separated from itself by some conditioning set C, generalizing the definition used, for example, by Didelez [13].

Local independence defines the independence model

$$\mathcal{I} = \{\langle A, B \mid C\rangle \mid X^B \text{ is locally independent of } X^A \text{ given } X^C\}$$

such that the local independence statement $A \not\to B \mid C$ is equivalent to $\langle A, B \mid C\rangle \in \mathcal{I}$ in the abstract notation. We note that the local independence model is generally not symmetric. Using Definition 2.1, we introduce below an associated directed graph in which there is no directed edge from a node α to a node β if and only β is locally independent of α given $V \setminus \{\alpha\}$.

DEFINITION 2.2 (Local independence graph). For the local independence model determined by X, we define the *local independence graph* to be the directed graph, $\mathcal{D}$, with nodes V such that for $\alpha, \beta \in V$,

$$\alpha \not\to_{\mathcal{D}} \beta \quad \Leftrightarrow \quad \alpha \not\to \beta \mid V \setminus \{\alpha\}$$

where $\alpha \not\to_{\mathcal{D}} \beta$ denotes that there is no directed edge from α to β in the graph $\mathcal{D}$.

Didelez [11] gives almost the same definition of a local independence graph, however, in essence always assumes that there is a dependence of each process on its own past. See also Sections A and B.

The local independence graph induces an independence model by μ-separation as defined below. The main goal of the present paper is to provide a graphical representation of the induced independence model for a subset of coordinate processes corresponding to the case where some processes are unobserved. This is achieved by establishing a correspondence, which is preserved under marginalization, between directed mixed graphs and independence models induced via μ-separation. We emphasize that the correspondence only relates local independence to graphs when the local independence model satisfies the global Markov property with respect to a graph.

The local independence model satisfies the global Markov property with respect to the local independence graph if every μ-separation in the graph implies a local independence. This has been shown for point processes under some mild regularity conditions [13] using the slightly different notion of δ-separation. Section A discusses how δ-separation is related to μ-separation, and Section B shows how to translate the global Markov property of [13] into our framework. Moreover, general sufficient conditions for the global Markov property were given in [30] covering point processes as well as certain diffusion processes. Section C provides, in addition, a discussion of Markov properties in the context of time series.

To help develop a better understanding of local independence and its relevance for applications, we discuss an example of drug abuse progression.

EXAMPLE 2.3 (Gateway drugs). The theory of *gateway drugs* has been discussed for many years in the literature on substance abuse [24, 40]. In short, the theory posits that the use of "soft" and often licit drugs precedes (and possibly leads to) later use of "hard" drugs. Alcohol, tobacco and marijuana have all been discussed as candidate gateway drugs to "harder" drugs such as heroin.

We propose a hypothetical, dynamical model of transitions into abuse via a gateway drug, and more generally, a model of substance abuse progression. Substance abuse is known to be associated with social factors, genetics and other individual and environmental factors [43]. Substance abuse can evolve over time when an individual starts or stops using some drug. In this example, we consider substance processes Alcohol (A), Tobacco (T), Marijuana (M) and Hard drugs (H) modeled as zero-one processes, that is, stochastic processes that are piecewise constantly equal to zero (no substance use) or one (substance use). We also include L, a process describing life events, and a process I, which can be thought of as an exogenous process that influences the tobacco consumption of the individual, for example, the price of tobacco which may change due to changes in tobacco taxation. Let $V = \{A, T, M, H, L, I\}$.

We will visualize each process as a node in a graph and draw an arrow from one process to another if the first has a direct influence on the second. We will not go into a full discussion of how to formalize "influence" in terms of a continuous-time causal dynamical model as this would lead us astray; see instead [13, 27, 38]. The upshot is that for a (faithful) causal model, there is no direct influence if and only if $\alpha \not\rightarrow \beta \mid V \setminus \{\alpha\}$, which identifies the "influence" graph with the local independence graph.

Several formalizations of the gateway drug question are possible. We will focus on the questions "is the use of hard drugs locally independent of use of alcohol for some conditioning set?" and "is the use of hard drugs locally independent of the use of tobacco for some conditioning set?" Using the dynamical nature of local independence, we are asking if, for example, the past alcohol usage changes the hard drug usage propensity when accounting for the past of all other processes in the model. This is one possible formalization of the gateway drug question as a negative answer would mean that there exist some gateway processes through which any influence of alcohol usage on hard drug usage is mediated. If the

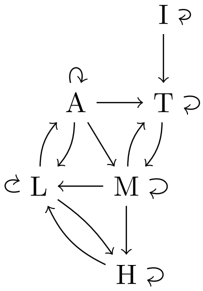


FIG. 3. *The directed graph of Example 2.3 illustrating a model where marijuana (M) potentially acts as a gateway drug, while alcohol (A) as well as tobacco (T) do not directly affect hard drug use.*

visualization in Figure 3 is indeed a local independence graph in the above sense we see that conditioning on all other processes, H is indeed locally independent of A and locally independent of T. In this hypothetical scenario, we could interpret this as marijuana in fact acting as a gateway drug to hard drugs. If the global Markov property holds, we can furthermore use μ-separation to obtain further local independences from the graph. We return to this example in Section 5.5 to illustrate how the main results of the paper can be applied. In particular, we are interested in what conclusions we can make when we do not observe all the processes but only a subset.

2.2. *Marginalization and separability.*

DEFINITION 2.4 (Marginalization). Given an independence model $\mathcal{I}$ over V, the marginal independence model over $O \subseteq V$ is defined as

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

Marginalization is defined abstractly above, though we are primarily interested in the marginalization of the independence model encoded by a local independence graph via μ-separation. The main objective is to obtain a graphical representation of such a marginalized independence model involving only the nodes O. To this end, we consider the notion of separability in an independence model.

DEFINITION 2.5 (Separability). Let $\mathcal{I}$ be an independence model over V. Let $\alpha, \beta \in V$. We say that β is *separable* from α if there exists $C \subseteq V \setminus \{\alpha\}$ such that $\langle \alpha, \beta \mid C\rangle \in \mathcal{I}$, and otherwise we say that β is *inseparable* from α. We define

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

We also define $u(\beta, \mathcal{I}) = V \setminus s(\beta, \mathcal{I})$.

We show in Proposition 3.6 that if $\mathcal{I}$ is the independence model induced by a directed graph via μ-separation, then $\alpha \in u(\beta, \mathcal{I})$ if and only if there is a directed edge from α to β. In this case, the graph is thus directly identifiable from separability properties of $\mathcal{I}$. That is, however, not true in general for a marginalization of $\mathcal{I}$, and this is the motivation for developing a theory of directed *mixed* graphs with μ-separation.

2.3. *Graph theory.* A graph, $\mathcal{G} = (V, E)$, is an ordered pair where V is a finite set of vertices (also called nodes) and E is a finite set of edges. Furthermore, there is a map that to each edge assigns a pair of nodes (not necessarily distinct). We say that the edge is *between* these two nodes. We consider graphs with two types of edges: directed (→) and bidirected

($\leftrightarrow$). We can think of the edge set as a disjoint union, $E = E_d \dot{\cup} E_b$, where E_d is a set of ordered pairs of nodes (α, β) corresponding to directed edges, and E_b is a set of unordered pairs of nodes $\{\alpha, \beta\}$ corresponding to bidirected edges. This implies that the edge $\alpha \leftrightarrow \beta$ is identical to the edge $\beta \leftrightarrow \alpha$, but the edge $\alpha \to \beta$ is different from the edge $\beta \to \alpha$. It also implies that the graphs we consider can have multiple edges between a pair of nodes α and β, but they will always be a subset of the edges $\{\alpha \to \beta, \alpha \leftarrow \beta, \alpha \leftrightarrow \beta\}$.

DEFINITION 2.6 (DMG). A *directed mixed graph* (DMG), $\mathcal{G} = (V, E)$, is a graph with node set V and edge set E consisting of directed and bidirected edges as described above.

Throughout the paper, $\mathcal{G}$ will denote a DMG with node set V and edge set E. Occasionally, we will also use $\mathcal{D}$ and $\mathcal{M}$ to denote DMGs. We use $\mathcal{D}$ only when the DMG is also a directed graph, that is, has no bidirected edges. We use $\mathcal{M}$ to stress that some DMG is obtained as a marginalization of a DMG on a larger node set. We will use notation such as $\leftrightarrow_{\mathcal{G}}$ or $\to_{\mathcal{D}}$ to denote the specific graph that an edge belongs to.

If $\alpha \to \beta$, we say that the edge has a *tail* at α and a *head* at β. Jointly tails and heads are called (edge) *marks*. An edge $e \in E$ between nodes α and β is a *loop* if $\alpha = \beta$. We also say that the edge is *incident* with the node α and with the node β and that α and β are *adjacent*.

For $\alpha, \beta \in V$, we use the notation $\alpha \sim \beta$ to denote a generic edge of any type between α and β. We use the notation $\alpha \ast\!\!\to \beta$ to indicate an edge that has a head at β and may or may not have a head at α. Note that the presence of one edge, $\alpha \to \beta$, say, does not in general preclude the presence of other edges between these two nodes. Finally, $\alpha \ast\!\!\not\to_{\mathcal{G}} \beta$ means that there is no edge in $\mathcal{G}$ between α and β that has a head at β and $\alpha \not\to_{\mathcal{G}} \beta$ means that there is no directed edge from α to β. Note that $\alpha \not\to_{\mathcal{G}} \beta$ is a statement about the absence of an edge in the graph $\mathcal{G}$ and to avoid confusion with local independence, $\alpha \not\to \beta \mid C$, we always include the conditioning set when writing local independence statements, even if $C = \emptyset$ (see also Definition 2.2).

We say that α is a *parent* of β in the graph $\mathcal{G}$ if $\alpha \to \beta$ is present in $\mathcal{G}$ and that β is a *child* of α. We say that α is a *sibling* of β (and that β is a sibling of α) if $\alpha \leftrightarrow \beta$ is present in the graph. The motivation of the term sibling will be explained in Section 3. We use $\mathrm{pa}(\alpha)$ to denote the set of parents of α.

A *walk* is an ordered, alternating sequence of vertices, γ_i, and edges, e_j, denoted $\omega = \langle \gamma_1, e_1, \dots, e_n, \gamma_{n+1} \rangle$, such that each e_i is between γ_i and γ_{i+1}, along with an orientation of each directed loop along the walk (if e_i is a loop then we also know if e_i points in the direction of γ_1 or in the direction of γ_{n+1}). Without the orientation, for instance, the walks $\alpha \to \beta \to \beta \to \gamma$ and $\alpha \to \beta \leftarrow \beta \to \gamma$ would be indistinguishable. See Figure 4 for examples. We will often present the walk ω using the notation

$$\gamma_1 \overset{e_1}{\sim} \gamma_2 \overset{e_2}{\sim} \dots \overset{e_n}{\sim} \gamma_{n+1},$$

where the loop orientation is explicit. We will omit the edge superscripts when they are not needed.

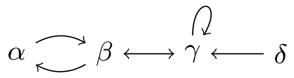


FIG. 4. *A directed mixed graph with node set $\{\alpha, \beta, \gamma, \delta\}$. Consider first the walk $\alpha \to \beta$. This is different from the walk $\beta \leftarrow \alpha$ as walks are ordered. Consider instead the two walks $\beta \leftrightarrow \gamma \leftarrow \gamma \leftarrow \delta$ and $\beta \leftrightarrow \gamma \to \gamma \leftarrow \delta$. These two walks have the same (ordered) sets of nodes and edges but are not equal as the loop at γ has different orientations between the two walks. Furthermore, one can note that for the first of the two walks, γ is a collider in the first instance, but not in the second. The walks $\alpha \to \beta \to \alpha$ and $\alpha \to \beta \leftarrow \alpha$ are both cycles, and the second is an example of the fact that the same edge can occur twice in a cycle.*

We say that the walk ω *contains* nodes γ_i and edges e_j. The *length* of the walk is n, the number of edges that it contains. We define a *trivial walk* to be a walk with no edges and, therefore, only a single node. Equivalently, a trivial walk can be defined as a walk of length zero. A *subwalk* of ω is either itself a walk of the form $\langle \gamma_k, e_k, \dots, e_{m-1}, \gamma_m\rangle$ where $1 \leq k < m \leq n+1$ or a trivial walk $\langle \gamma_k\rangle$, $1 \leq k \leq n+1$. A (nontrivial) walk is uniquely identified by its edges, and the ordering and orientation of these edges, hence the vertices can be omitted when describing the walk. At times, we will omit the edges to simplify notation, however, we will always have a specific, uniquely identified walk in mind even when the edges and/or their orientation is omitted. The first and last nodes of a walk are called *endpoint nodes* (these could be equal) or just endpoints, and we say that a walk is *between* its endpoints, or alternatively *from* its first node *to* its last node. We call the walk $\omega^{-1} = \langle \gamma_{n+1}, e_n, \dots, e_1, \gamma_1\rangle$ the *inverse* walk of ω. Note that the orientation of directed loops is also reversed in the inverse walk such that they point toward γ_1 in the inverse if and only if they point toward γ_1 in the original walk. A *path* is a walk on which no node is repeated.

Consider a walk ω and a subwalk thereof, $\langle \alpha, e_1, \gamma, e_2, \beta\rangle$, where $\alpha, \beta, \gamma \in V$ and $e_1, e_2 \in E$. If e_1 and e_2 both have heads at γ, then γ is a *collider* on ω. If this is not the case, then γ is a noncollider. Note that an endpoint of a walk is neither a collider, nor a noncollider. We stress that the property of being a collider/noncollider is relative to a walk (see also Figure 4).

Let $\omega_1 = \langle \alpha, e_1^1, \gamma_1^1, \dots, \gamma_{n-1}^1, e_n^1, \beta\rangle$ and $\omega_2 = \langle \alpha, e_1^2, \gamma_1^2, \dots, \gamma_{m-1}^2, e_m^2, \beta\rangle$ be two (nontrivial) walks. We say that they are *endpoint-identical* if e_1^1 and e_1^2 have the same mark at α and e_n^1 and e_m^2 have the same mark at β. Note that this may depend on the orientation of directed edges in the two walks. Assume that some edge e is between α and β. We say that the (nontrivial) walk ω_1 is endpoint-identical to e if it is endpoint-identical to the walk $\langle \alpha, e, \beta\rangle$. If $\alpha = \beta$ and e is directed, this should hold for just one of the possible orientations of e.

Let ω_1 be a walk between α and γ, and ω_2 a walk between γ and β. The *composition* of ω_1 with ω_2 is the walk that starts at α, traverses every node and edge of ω_1, and afterwards every node and edge of ω_2, ending in β. We say that we *compose* ω_1 with ω_2.

A *directed path* from α to β is a path between α and β consisting of edges of type $\rightarrow$ only (possibly of length zero) such that they all point in the direction of β. A *cycle* is either a loop, or a (nontrivial) path from α to β composed with $\beta \sim \alpha$. This means that in a cycle of length 2, an edge can be repeated. A *directed cycle* is either a loop, $\alpha \rightarrow \alpha$, or a (nontrivial) directed path from α to β composed with $\beta \rightarrow \alpha$. For $\alpha \in V$, we let $\mathrm{An}(\alpha)$ denote the set of *ancestors*, that is,

$$\mathrm{An}(\alpha) = \{\gamma \in V \mid \text{there is a directed path from } \gamma \text{ to } \alpha\}.$$

This is generalized to nonsingleton sets $C \subseteq V$,

$$\mathrm{An}(C) = \bigcup_{\alpha \in C} \mathrm{An}(\alpha).$$

We stress that $C \subseteq \mathrm{An}(C)$ as we allow for trivial directed paths in the definition of an ancestor. We use the notation $\mathrm{An}_{\mathcal{G}}(C)$ if we wish to emphasize in which graph the ancestry is read, but omit the subscript when no ambiguity arises.

Let $\mathcal{G} = (V, E)$ be a graph, and let $O \subseteq V$. Define the *subgraph* induced by O to be the graph $\mathcal{G}_O = (O, E_O)$ where $E_O \subseteq E$ is the set of edges that are between nodes in O. If $\mathcal{G}_1 = (V, E_1)$ and $\mathcal{G}_2 = (V, E_2)$, we will write $\mathcal{G}_1 \subseteq \mathcal{G}_2$ to denote $E_1 \subseteq E_2$ and say that $\mathcal{G}_2$ is a *supergraph* of $\mathcal{G}_1$.

A *directed graph* (DG), $\mathcal{D} = (V, E)$, is a graph with only directed edges. Note that this also allows directed loops. Within a class of graphs, we define the *complete* graph to be the graph which is the supergraph of all graphs in the class when such a graph exists. For the class of DGs on node set V, the complete graph is the graph with edge set $E = \{(\alpha, \beta) \mid \alpha, \beta \in V\}$.

A *directed acyclic graph* (DAG) is a DG with no loops and no directed cycles. An *acyclic directed mixed graph* (ADMG) is a DMG with no loops and no directed cycles.

3. Directed mixed graphs and separation. In this section, we introduce μ-separation for DMGs which are then shown to be closed under marginalization. In particular, we obtain a DMG representing the independence model arising from a local independence graph via marginalization.

The class of DMGs contains as a subclass the ADMGs that have no directed cycles [19, 32]. ADMGs have been used to represent marginalized DAG models, analogously to how we will use DMGs to represent marginalized DGs. ADMGs come with the m-separation criterion which can be extended to DMGs, but this criterion differs in important ways from the μ-separation criterion introduced below. These differences also mean that our main result on Markov equivalence does not apply to, for example, DMGs with m-separation, and thus our theory of Markov equivalence hinges on the fact that we are considering DMGs using the asymmetric notion of μ-separation.

3.1. *μ-separation.* We define μ-separation as a generalization of δ-separation introduced by Didelez [11], analogously to how m-separation is a generalization of d-separation; see, for example, [33]. In Section A, we make the connection to Didelez's δ-separation exact and elaborate further on this in Section B.

DEFINITION 3.1 (μ-connecting walk). A nontrivial walk

$$\langle \alpha, e_1, \gamma_1, \dots, \gamma_{n-1}, e_n, \beta \rangle$$

in $\mathcal{G}$ is said to be μ-connecting (or simply *open*) from α to β given C if $\alpha \notin C$, every collider is in $\mathrm{An}(C)$, no noncollider is in C, and e_n has a head at β.

When a walk is not μ-connecting given C, we say that it is *closed* or *blocked* by C. One should note that if ω is a μ-connecting walk from α to β given C, the inverse walk, ω^{-1}, is not in general μ-connecting from β to α given C. The requirement that a μ-connecting walk be nontrivial, that is, of strictly positive length, leads to the possibility of a node being separated from itself by some set C when applying the following graph separation criterion to the class of DMGs.

DEFINITION 3.2 (μ-separation). Let $A, B, C \subseteq V$. We say that B is μ-separated from A given C if there is no μ-connecting walk from any $\alpha \in A$ to any $\beta \in B$ given C and write $A \perp_\mu B \mid C$, or write $A \perp_\mu B \mid C\ [\mathcal{G}]$ if we want to stress to what graph the separation statement applies.

The above notion of separation is given in terms of walks of which there are infinitely many in any DMG with a nonempty edge set. However, we will see that it is sufficient to consider a finite subset of walks from A to B (Proposition 3.5).

Given a DMG, $\mathcal{G} = (V, E)$, we define an independence model over V using μ-separation,

$$\mathcal{I}(\mathcal{G}) = \bigl\{\langle A, B \mid C\rangle \mid (A \perp_\mu B \mid C)\bigr\}.$$

Definition 3.1 implies $A \perp_\mu B \mid C$ whenever $A \subseteq C$ and, therefore, $\mathcal{I}(\mathcal{G}) \neq \varnothing$.

Below we state two propositions that essentially both give equivalent ways of defining μ-separation. The propositions are useful when proving results on μ-separation models.

PROPOSITION 3.3. *Let $\alpha, \beta \in V$, $C \subseteq V$. If there is a μ-connecting walk from α to β given C, then there is a μ-connecting walk from α to β that furthermore satisfies that every collider is in C.*

DEFINITION 3.4. A *route* from α to β is a walk from α to β such that no node different from β occurs more than once, and β occurs at most twice.

A route is always a path, a cycle or a composition of a path and a cycle that share no edge and only share the vertex β.

PROPOSITION 3.5. *Let $\alpha, \beta \in V$, $C \subseteq V$. If ω is a μ-connecting walk from α to β given C, then there is a μ-connecting route from α to β given C consisting of edges in ω.*

If there is a μ-connecting walk from A to B given C, it does not in general follow that we can also find a μ-connecting path or cycle from A to B given C. As an example of this, consider the following DMG on nodes $\{\alpha, \beta, \gamma\}$: $\alpha \leftarrow \beta \leftarrow \gamma$. There is a μ-connecting walk from α to β given $\emptyset$, and a μ-connecting route, but no μ-connecting path from α to β given $\emptyset$.

3.2. *Marginalization of DMGs.* Given a DG or a DMG, $\mathcal{G}$, we are interested in finding a graph that represents the marginal independence model over a node set $O \subseteq V$, that is, finding a graph $\mathcal{M}$ such that

$$\mathcal{I}(\mathcal{M}) = \big(\mathcal{I}(\mathcal{G})\big)^O. \tag{3.1}$$

It is well known that the class of DAGs with d-separation is not closed under marginalization, that is, for a DAG, $\mathcal{D} = (V, E)$, and $O \subsetneq V$, it is not in general possible to find a DAG with node set O that encodes the same independence model among the variables in O as did the original graph. Richardson and Spirtes [33] gave a concrete counterexample and in Example 3.7 we give a similar example to make the analogous point: DGs read with μ-separation are not closed under marginalization. In this example, we use the following proposition which gives a simple characterization of separability in DGs.

PROPOSITION 3.6. *Consider a DG, $\mathcal{D} = (V, E)$, and let $\alpha, \beta \in V$. Then β is μ-separable (see Definition 2.5) from α in $\mathcal{D}$ if and only if $\alpha \not\to_{\mathcal{D}} \beta$.*

EXAMPLE 3.7. Consider the directed graph, $\mathcal{G}$, in Figure 5. We wish to show that it is not possible to encode the μ-separations among nodes in $O = \{\alpha, \beta, \gamma, \delta\}$ using a DG on these nodes only. To obtain a contradiction, assume $\mathcal{D} = (O, E)$ is a DG such that

$$A \perp_\mu B \mid C\ [\mathcal{D}] \Leftrightarrow A \perp_\mu B \mid C\ [\mathcal{G}] \tag{3.2}$$

for $A, B, C \subseteq O$. There is no $C \subseteq O \setminus \{\alpha\}$ such that $\alpha \perp_\mu \beta \mid C\ [\mathcal{G}]$ and no $C \subseteq O \setminus \{\beta\}$ such that $\beta \perp_\mu \gamma \mid C\ [\mathcal{G}]$. If $\mathcal{D}$ has the property (3.2), then it follows from Proposition 3.6 that $\alpha \to_{\mathcal{D}} \beta$ and $\beta \to_{\mathcal{D}} \gamma$. However, then γ is not μ-separated from α given $\emptyset$ in $\mathcal{D}$. This shows that there exists no DG, $\mathcal{D}$, that satisfies (3.2).

We note that marginalization of a probability model does not only impose conditional independence constraints on the observed variables but also so-called equality and inequality constraints; see, for example, [18] and references therein. In this paper, we will only be

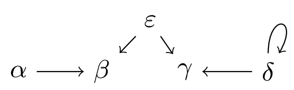


FIG. 5. *The directed graph of Example 3.7 which exemplifies that DGs are not closed under marginalization.*

concerned with the graphical representation of local independence constraints, and not with representing analogous equality or inequality constraints.

In the remainder of this section, we first introduce the *latent projection* of a graph (see also [41] and [34]), and then show that it provides a marginalized DMG in the sense of (3.1). At the end of the section, we give an algorithm for computing the latent projection of a DMG. This algorithm is an adapted version of one described by Sadeghi [36] for a different class of graphs. Koster [25] described a similar algorithm for ADMGs.

DEFINITION 3.8 (Latent projection). Let $\mathcal{G} = (V, E)$ be a DMG, $V = M \dot\cup O$. We define the *latent projection of* $\mathcal{G}$ *on* O to be the DMG (O, D) such that $\alpha \sim \beta \in D$ if and only if there exists an endpoint-identical (and nontrivial) walk between α and β in $\mathcal{G}$ with no colliders and such that every nonendpoint node is in M. Let $m(\mathcal{G}, O)$ denote the latent projection of $\mathcal{G}$ on O.

The definition of latent projection motivates the graphical term *sibling* for DMGs, as one way to obtain an edge $\alpha \leftrightarrow \beta$ is through a latent projection of a larger graph in which α and β share a parent.

To characterize the class of graphs obtainable from a DG via a latent projection, we introduce the *canonical DG of the DMG* $\mathcal{G}$, $\mathcal{C}(\mathcal{G})$, as follows: for each (unordered) pair of nodes $\{\alpha, \beta\} \subseteq V$ such that $\alpha \leftrightarrow_{\mathcal{G}} \beta$, add a distinct auxiliary node, $m_{\{\alpha,\beta\}}$, add edges $m_{\{\alpha,\beta\}} \to \alpha$, $m_{\{\alpha,\beta\}} \to \beta$ to E and then remove all bidirected edges from E. If $\mathcal{D}$ is a DG, then $\mathcal{M} = m(\mathcal{D}, O)$ will satisfy

$$\alpha \leftrightarrow_{\mathcal{M}} \beta \quad \Rightarrow \quad \alpha \leftrightarrow_{\mathcal{M}} \alpha \quad \text{for all } \alpha, \beta \in O \tag{3.3}$$

for all subsets of vertices O. Conversely, if $\mathcal{G} = (V, E)$ is a DMG that satisfies (3.3), then $\mathcal{G}$ is the latent projection of its canonical DG; $m(\mathcal{C}(\mathcal{G}), V) = \mathcal{G}$. The class of DMGs that satisfy (3.3) is closed under marginalization (Proposition 3.9) and has certain regularity properties (see, e.g., Proposition 3.10). These result provide the means for graphically representing marginals of local independence graphs. However, the theory that leads to our main results on Markov equivalence does not require the property (3.3) and, therefore, we develop it for general DMGs.

PROPOSITION 3.9. *Let $O \subseteq V$. The graph $\mathcal{M} = m(\mathcal{G}, O)$ is a DMG. If $\mathcal{G}$ satisfies* (3.3), *then $\mathcal{M}$ does as well.*

PROPOSITION 3.10. *Assume that $\mathcal{G}$ satisfies* (3.3) *and let $\alpha \in V$. Then α has no loops if and only if $\alpha \perp_\mu \alpha \mid V \setminus \{\alpha\}$.*

We also observe directly from the definition that the latent projection operation preserves ancestry and nonancestry in the following sense.

PROPOSITION 3.11. *Let $O \subseteq V$, $\mathcal{M} = m(\mathcal{G}, O)$ and $\alpha, \beta \in O$. Then $\alpha \in \mathrm{An}_{\mathcal{G}}(\beta)$ if and only if $\alpha \in \mathrm{An}_{\mathcal{M}}(\beta)$.*

The main result of this section is the following theorem, which states that the marginalization defined by the latent projection operation preserves the marginal independence model encoded by a DMG.

THEOREM 3.12. *Let $O \subseteq V$, $\mathcal{M} = m(\mathcal{G}, O)$. Assume $A, B, C \subseteq O$. Then*

$$A \perp_\mu B \mid C\ [\mathcal{G}] \quad \Leftrightarrow \quad A \perp_\mu B \mid C\ [\mathcal{M}].$$

```
input    : a DMG, $\mathcal{G} = (V, E)$ a subset $M \subseteq V$ over which to marginalize
output   : a graph $\mathcal{M} = (O, \bar{E})$, $O = V \setminus M$
Initialize $E_0 = E$, $\mathcal{M}_0 = (V, E_0)$, $k = 0$;
while $\Omega_M(\mathcal{M}_k) \neq \varnothing$ do
    Choose $\theta = {}_{\theta}(\alpha, m, \beta) \in \Omega_M(\mathcal{M}_k)$;
    Set $e_{k+1}$ to be the edge $\alpha \sim \beta$ which is endpoint-identical to $\theta$;
    Set $E_{k+1} = E_k \cup \{e_{k+1}\}$;
    Set $\mathcal{M}_{k+1} = (V, E_{k+1})$;
    Update $k = k + 1$
end
return $(\mathcal{M}_k)_O$
```

Algorithm 1: Computing the latent projection of a DMG

3.3. *A marginalization algorithm.* We describe an algorithm to compute the latent projection of a graph on some subset of nodes. For this purpose, we define a *triroute*, θ, to be a walk of length 2, $\langle \alpha, e_1, \gamma, e_2, \beta \rangle$, such that $\gamma \neq \alpha, \beta$. We suppress e_1 and e_2 from the notation and use ${}_{\theta}(\alpha, \gamma, \beta)$ to denote the triroute. We say that a triroute is *colliding* if γ is a collider on θ, and otherwise we say that it is *noncolliding*. This is analogous to the concept of a *tripath* (see, e.g., [26]), but allows for $\alpha = \beta$.

Define $\Omega_M(\mathcal{G})$ to be the set of noncolliding triroutes ${}_{\theta}(\alpha, m, \beta)$ such that $m \in M$ and such that an endpoint-identical edge $\alpha \sim \beta$ is not present in $\mathcal{G}$.

PROPOSITION 3.13. *Algorithm 1 outputs the latent projection of a DMG.*

4. Properties of DMGs.

DEFINITION 4.1 (Markov equivalence). Let $\mathcal{G}_1 = (V, E_1)$ and $\mathcal{G}_2 = (V, E_2)$ be DMGs. We say that $\mathcal{G}_1$ and $\mathcal{G}_2$ are *Markov equivalent* if $\mathcal{I}(\mathcal{G}_1) = \mathcal{I}(\mathcal{G}_2)$. This defines an equivalence relation and we let $[\mathcal{G}_1]$ denote the (Markov) equivalence class of $\mathcal{G}_1$.

EXAMPLE 4.2 (Markov equivalence in DGs). Let $\mathcal{D} = (V, E)$ be a DG. There is a directed edge from α to β if and only if β cannot be separated from α by any set $C \subseteq V \setminus \{\alpha\}$ (Proposition 3.6). This implies that two DGs are Markov equivalent if and only if they are equal. Thus, in the restricted class of DGs, every Markov equivalence class is a singleton and in this sense *identifiable* from its induced independence model. However, when considering Markov equivalence in the more general class of DMGs not every equivalence class of a DG is a singleton as the DG might be Markov equivalent to a DMG. As an example of this, consider the complete DG on a node set V which is Markov equivalent to the complete DMG on V.

DEFINITION 4.3 (Maximality of a DMG). We say that $\mathcal{G}$ is *maximal* if it is complete, or if any added edge changes the induced independence model $\mathcal{I}(\mathcal{G})$.

4.1. *Inducing paths.* Separability of nodes can be studied using the concept of an *inducing path* which has also been used in other classes of graphs [33, 41]. In the context of DMGs and μ-separation, it is natural to define several types of inducing paths due to the asymmetry of μ-separation and the possibility of directed cycles in DMGs.

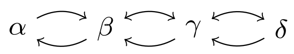


FIG. 6. *Examples of inducing paths in a DMG: the path $\beta \to \alpha$ is a unidirected inducing path from β to α, and also a directed inducing path. The path $\beta \leftrightarrow \gamma$ is a bidirected inducing path. The path $\beta \leftrightarrow \gamma \leftrightarrow \delta$ is a bidirected inducing path from β to δ (and by definition its inverse is a bidirected inducing path from δ to β). The path $\delta \to \gamma \leftrightarrow \beta$ is both a unidirected and a directed inducing path from δ to β, whereas the path $\alpha \to \beta \leftrightarrow \gamma$ is a unidirected inducing path from α to γ, but not a directed inducing path.*

DEFINITION 4.4 (Inducing path). An *inducing path from α to β* is a nontrivial path or cycle, $\pi = \langle \alpha, \dots, \beta \rangle$, which has a head at β and such that there are no noncolliders on π and every node is an ancestor of α or β. The inducing path π is *bidirected* if every edge on π is bidirected. If π is not bidirected, it has one of the forms $\alpha \to \beta$ or

$$\alpha \to \gamma_1 \leftrightarrow \cdots \leftrightarrow \gamma_n \leftrightarrow \beta.$$

and we say that it is *unidirected*. If, furthermore, $\gamma_i \in \mathrm{An}(\beta)$ for all $i = 1, \dots, n$ (or it is on the form $\alpha \to \beta$) then we say that it is *directed*.

Note that an inducing path is by definition either a path or a cycle. An inducing path is either bidirected or unidirected. Some unidirected inducing paths are also directed; see Figure 6 for examples. Propositions 4.7 and 4.8 show how bidirected and directed inducing paths in a certain sense correspond to bidirected and directed edges, respectively.

PROPOSITION 4.5. *Let ν be an inducing path from α to β. The following holds for any $C \subseteq V \setminus \{\alpha\}$. If $\alpha \neq \beta$, then there exists a μ-connecting path from α to β given C. If $\alpha = \beta$, then there exists a μ-connecting cycle from α to β given C. We call such a path or cycle a* ν-induced open path or cycle, *respectively, or simply a ν-induced open walk to cover both the case $\alpha = \beta$ and the case $\alpha \neq \beta$. If the inducing path is bidirected or directed, then the ν-induced open walk is endpoint-identical to the inducing path.*

The following corollary is a direct consequence of Proposition 4.5, showing that β is inseparable from α if there is an inducing path from α to β irrespectively of whether the nodes are adjacent.

COROLLARY 4.6. *Let $\alpha, \beta \in V$. If there exists an inducing path from α to β in $\mathcal{G}$, then β is not μ-separated from α given C for any $C \subseteq V \setminus \{\alpha\}$, that is, $\alpha \in u(\beta, \mathcal{I}(\mathcal{G}))$.*

The following two propositions show that for two of the three types of inducing paths there is a Markov equivalent supergraph in which the nodes are adjacent. This illustrates how one can easily find Markov equivalent DMGs that do not have the same adjacencies. Example 4.12 shows that for a unidirected inducing path it may not be possible to add an edge without changing the independence model.

PROPOSITION 4.7. *If there exists a bidirected inducing path from α to β in $\mathcal{G}$, then adding $\alpha \leftrightarrow \beta$ in $\mathcal{G}$ does not change the independence model.*

PROPOSITION 4.8. *If there exists a directed inducing path from α to β in $\mathcal{G}$, then adding $\alpha \to \beta$ in $\mathcal{G}$ does not change the independence model.*

We say that nodes α and β are *collider-connected* if there exists a nontrivial walk between α and β such that every nonendpoint node is a collider on the walk. We say that α is *directedly collider-connected* to β if α and β are collider-connected by a walk with a head at β.

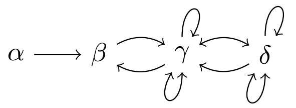


FIG. 7. *A maximal DMG in which δ is inseparable from β, though no edge is between the two. See Example 4.12. We will in general omit the bidirected loops from the visual presentations of DMGs*; *see also the discussion in Section 5.4.*

DEFINITION 4.9. Let $\alpha, \beta \in V$. We define the set

$$D(\alpha, \beta) = \{\gamma \in \mathrm{An}(\alpha, \beta) \mid \gamma \text{ is directedly collider-connected to } \beta\} \setminus \{\alpha\}.$$

Note that if $\alpha \not\to_{\mathcal{G}} \beta$, then $\mathrm{pa}(\beta) \subseteq D(\alpha, \beta)$, and if the graph is furthermore a directed graph then $\mathrm{pa}(\beta) = D(\alpha, \beta)$.

PROPOSITION 4.10. *If there is no inducing path from α to β in $\mathcal{G}$, then β is separated from α by $D(\alpha, \beta)$.*

EXAMPLE 4.11 (Inducing paths). Consider the DMG on nodes $\{\alpha, \gamma\}$ and with a single edge $\gamma \to \alpha$. In this case, there is no inducing path from α to α and α is μ-separated from α by $D(\alpha, \alpha) = \{\gamma\}$. Now add the edge $\alpha \leftrightarrow \gamma$. In this new DMG, there is an inducing path from α to α and therefore α is inseparable from itself.

EXAMPLE 4.12 (Nonadjacency of inseparable nodes in a maximal DMG). Consider the DMG in Figure 7. One can show that this DMG is maximal (Definition 4.3). There is an inducing path from β to δ making δ inseparable from β, yet no arrow can be added between β and δ without changing the independence model. This example illustrates that maximal DMGs do not have the property that inseparable nodes are adjacent. This is contrary to MAGs which form a subclass of ancestral graphs and have this exact property [33].

5. Markov equivalence of DMGs. The main result of this section is that each Markov equivalence class of DMGs has a greatest element, that is, an element which is a supergraph of all other elements. This fact is helpful for understanding and graphically representing such equivalence classes, and potentially also for constructing learning algorithms. We will prove this result by arguing that the independence model of a DMG, $\mathcal{G} = (V, E)$, defines for each node $\alpha \in V$ a set of *potential parents* and a set of *potential siblings*. We then construct the greatest element of $[\mathcal{G}]$ by simply using these sets, and argue that this is in fact a Markov equivalent supergraph. As we only use the independence model to define the sets of potential parents and siblings, the supergraph is identical for all members of $[\mathcal{G}]$, and thus a greatest element. Within the equivalence class, the greatest element is also the only maximal element, and we will refer to it as the maximal element of the equivalence class.

5.1. *Potential siblings.*

DEFINITION 5.1. Let $\mathcal{I}$ be an independence model over V and let $\alpha, \beta \in V$. We say that α and β are *potential siblings* in $\mathcal{I}$ 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 $\beta \in C$,

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

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

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

Potential siblings are defined abstractly above in terms of the independence model only. The following proposition gives a useful characterization for graphical independence models by simply contraposing (s2) and (s3).

PROPOSITION 5.2. *Let $\mathcal{I}(\mathcal{G})$ be the independence model induced by $\mathcal{G}$. Then $\alpha, \beta \in V$ are potential siblings if and only if* (gs1)–(gs3) *hold*:

(gs1) $\beta \in u(\alpha, \mathcal{I}(\mathcal{G}))$ *and* $\alpha \in u(\beta, \mathcal{I}(\mathcal{G}))$,

(gs2) *for all $\gamma \in V$, $C \subseteq V$ such that $\beta \in C$: if there exists a μ-connecting walk from γ to β given C, then there exists a μ-connecting walk from γ to α given C,*

(gs3) *for all $\gamma \in V$, $C \subseteq V$ such that $\alpha \in C$: if there exists a μ-connecting walk from γ to α given C, then there exists a μ-connecting walk from γ to β given C.*

PROPOSITION 5.3. *Assume that $\alpha \leftrightarrow \beta$ is in $\mathcal{G}$. Then α and β are potential siblings in $\mathcal{I}(\mathcal{G})$.*

LEMMA 5.4. *Assume that α and β are potential siblings in $\mathcal{I}(\mathcal{G})$. Let $\mathcal{G}^+$ denote the DMG obtained from $\mathcal{G}$ by adding $\alpha \leftrightarrow \beta$. Then $\mathcal{I}(\mathcal{G}) = \mathcal{I}(\mathcal{G}^+)$.*

The above shows that if α and β are potential siblings in $\mathcal{I}(\mathcal{G})$ then there exists a supergraph, $\mathcal{G}^+$, which is Markov equivalent with $\mathcal{G}$, such that α and β are siblings in $\mathcal{G}^+$. This motivates the term *potential* siblings.

5.2. *Potential parents.* In this section, we will argue that also a set of *potential parents* are determined by the independence model. This case is slightly more involved for two reasons. First, the relation is asymmetric, as for each potential parent edge there is a parent node and a child node. Second, adding directed edges potentially changes the ancestry of the graph.

DEFINITION 5.5. Let $\mathcal{I}$ be an independence model over V and let $\alpha, \beta \in V$. We say that α is a *potential parent* of β in $\mathcal{I}$ if (p1)–(p4) hold:

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

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

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

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

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

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

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

PROPOSITION 5.6. *Let $\mathcal{I}(\mathcal{G})$ be the independence model induced by $\mathcal{G}$. Then $\alpha \in V$ is a potential parent of $\beta \in V$ if and only if* ($gp1$)–($gp4$) *hold*:

(gp1) $\alpha \in u(\beta, \mathcal{I}(\mathcal{G}))$,

(gp2) *for all $\gamma \in V$, $C \subseteq V$ such that $\alpha \notin C$: if there exists a μ-connecting walk from γ to α given C, then there exists a μ-connecting walk from γ to β given C,* (gp3) *for all $\gamma, \delta \in V$, $C \subseteq V$ such that $\alpha \notin C$, $\beta \in C$: if there exists a μ-connecting walk from γ to β given C and a μ-connecting walk from α to δ given C, then there exists a μ-connecting walk from γ to δ given C,*

(gp4) *for all $\gamma \in V$, $C \subseteq V$, such that $\alpha \notin C$: if there exists a μ-connecting walk from β to γ given $C \cup \{\alpha\}$, then there exists a μ-connecting walk from β to γ given C.*

PROPOSITION 5.7. *Assume that $\alpha \to \beta$ is in $\mathcal{G}$. Then α is a potential parent of β in $\mathcal{I}(\mathcal{G})$.*

LEMMA 5.8. *Assume that α is a potential parent of β in $\mathcal{I}(\mathcal{G})$. Let $\mathcal{G}^+$ denote the DMG obtained from $\mathcal{G}$ by adding $\alpha \to \beta$. Then $\mathcal{I}(\mathcal{G}) = \mathcal{I}(\mathcal{G}^+)$.*

5.3. *A Markov equivalent supergraph.* Let $\mathcal{G} = (V, E)$ be a DMG. Define $\mathcal{N}(\mathcal{I}(\mathcal{G})) = (V, E^m)$ to be the DMG with edge set $E^m = E^d \cup E^b$ where E^d is a set of directed edges and E^b a set of bidirected edges such that the directed edge from α to β is in E^d if and only if α is a potential parent of β in $\mathcal{I}(\mathcal{G})$ and the bidirected edge between α and β is in E^b if and only if α and β are potential siblings in $\mathcal{I}(\mathcal{G})$.

THEOREM 5.9. *Let $\mathcal{N} = \mathcal{N}(\mathcal{I}(\mathcal{G}))$. Then $\mathcal{N} \in [\mathcal{G}]$ and $\mathcal{N}$ is a supergraph of all elements of $[\mathcal{G}]$. Furthermore, if we have a finite sequence of DMGs $\mathcal{G}_0, \mathcal{G}_1, \dots, \mathcal{G}_m$, $\mathcal{G}_i = (V, E_i)$, such that $\mathcal{G}_0 = \mathcal{G}$, $\mathcal{G}_m = \mathcal{N}$, and $E_i \subseteq E_{i+1}$ for all $i = 0, \dots, m-1$, then $\mathcal{G}_i$ is Markov equivalent with $\mathcal{N}$ for all $i = 0, \dots, m-1$.*

The graph $\mathcal{N}$ in the above theorem is a supergraph of every Markov equivalent DMG and, therefore, maximal. On the other hand, every maximal DMG is a representative of its equivalence class, and also a supergraph of all Markov equivalent DMGs. This means that we can use the class of maximal DMGs to obtain a unique representative for each DMG equivalence class.

Lemmas 5.4 and 5.8 show that conditions (gs1)–(gs3) and (gp1)–(gp4) are sufficient to Markov equivalently add a bidirected or a directed edge, respectively. The conditions are also necessary in the sense that for each condition one can find example graphs where only a single condition is violated and where the larger graph is not Markov equivalent to the smaller graph.

We can note that α is a potential parent and a potential sibling of α if and only if $\alpha \in u(\alpha, \mathcal{I}(\mathcal{G}))$. This means that in $\mathcal{N}(\mathcal{I}(\mathcal{G}))$ for each node either both a directed and a bidirected loop is present or no loop at all.

5.4. *Directed mixed equivalence graphs.* Theorem 5.9 suggests that one can represent an equivalence class of DMGs by displaying the maximal element and then simply indicate which edges are not present for all members of the equivalence class.

DEFINITION 5.10 (DMEG). Let $\mathcal{N} = (V, F)$ be a maximal DMG. Define $\bar{F} \subseteq F$ such that for $e \in F$ we let $e \in \bar{F}$ if and only if there exists a DMG $\mathcal{G} = (V, \tilde{F})$ such that $\mathcal{G} \in [\mathcal{N}]$ and $e \notin \tilde{F}$. We call $\mathcal{N}' = (V, F, \bar{F})$ a *directed mixed equivalence graph* (DMEG). When visualizing $\mathcal{N}'$, we draw $\mathcal{N}$, but use dashed edges for the set $\bar{F}$; see Figure 8.

Let $\mathcal{N}' = (V, F, \bar{F})$ be a DMEG. The DMG (V, F) is in the equivalence class represented by $\mathcal{N}'$. However, one cannot necessarily remove any subset of $\bar{F}$ and obtain a member of the Markov equivalence class (see Figure 8). Moreover, an equivalence class does not in general contain a least element, that is, an element which is a subgraph of all Markov equivalent graphs.

We will throughout this section let $\mathcal{N} = (V, F)$ be a maximal DMG. For $e \in F$, we will use $\mathcal{N} - e$ to denote the graph $(V, F \setminus \{e\})$. Assume that we have a maximal DMG from which we wish to derive the DMEG. Consider some edge $e \in F$. If $\mathcal{N} - e \in [\mathcal{N}]$, then $e \in \bar{F}$ as there exists a Markov equivalent subgraph of $\mathcal{N}$ in which e is not present. On the other hand, if $\mathcal{N} - e \notin [\mathcal{N}]$ then we note that $\mathcal{N} - e$ is the largest subgraph of $\mathcal{N}$ that does not contain e. Let $\mathcal{K}$ be a subgraph of $\mathcal{N}$ that does not contain e. Then $\mathcal{I}(\mathcal{N}) \subsetneq \mathcal{I}(\mathcal{N} - e) \subseteq \mathcal{I}(\mathcal{K})$. Using Theorem 5.9, we know that all $\mathcal{N}$-Markov equivalent DMGs are in fact subgraphs of $\mathcal{N}$, and using that $\mathcal{K}$ is not Markov equivalent to $\mathcal{N}$ we see that all graphs in $[\mathcal{N}]$ must contain e. This means that when $\mathcal{N} - e \notin [\mathcal{N}]$ then $e \notin \bar{F}$ as e must be present in all Markov equivalent DMGs.

Any loop should in principle be dashed when drawing a DMEG as for each node in a maximal DMG either both the directed and the bidirected loop are present or neither of them. However, we choose to not present them as dashed as if they are present in the maximal DMG, then at least one of them will be present in any Markov equivalent DMG satisfying (3.3), that is, for any DMG which is a marginalization of a DG. In addition, we only draw the directed loop to not overload the visualizations.

5.5. *Constructing a directed mixed equivalence graph.* When constructing a DMEG from $\mathcal{N}$, it suffices to consider the graphs $\mathcal{N} - e$ for each $e \in E$ and determine if they are Markov equivalent to $\mathcal{N}$ or not. A brute-force approach to doing so is to simply check all separation statements in both graphs. However, one can make a considerably more efficient algorithm.

PROPOSITION 5.11. *Assume $\alpha \overset{e}{\to}_{\mathcal{N}} \beta$. It holds that $\mathcal{N} - e \in [\mathcal{N}]$ if and only if $\alpha \in u(\beta, \mathcal{I}(\mathcal{N} - e))$.*

PROPOSITION 5.12. *Assume $\alpha \overset{e}{\leftrightarrow}_{\mathcal{N}} \beta$. Then $\mathcal{N} - e \in [\mathcal{N}]$ if and only if $\alpha \in u(\beta, \mathcal{I}(\mathcal{N} - e))$ and $\beta \in u(\alpha, \mathcal{I}(\mathcal{N} - e))$.*

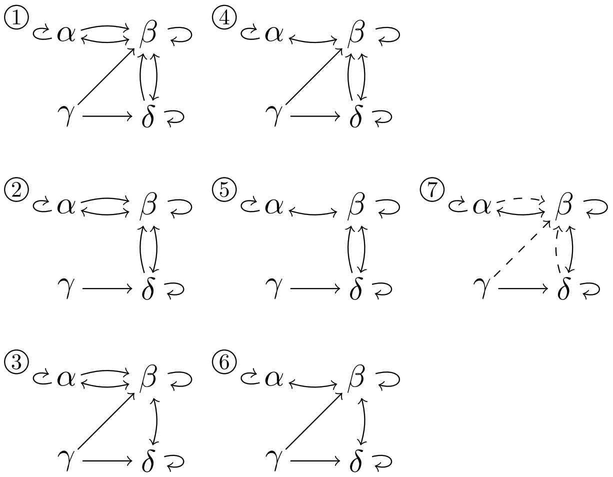


FIG. 8. *The DMG ① is maximal (the bidirected loops at α, β and δ have been omitted from the visual presentation). The DMGs ①–⑥ are the six elements of its Markov equivalence class (when ignoring Markov equivalent removal of loops). The graph ⑦ is the corresponding DMEG. In a DMEG, every solid edge is in every graph in the equivalence class, every absent edge is not in any graph, and every dashed edge is in some, but not in others. Note that every DMG in the above equivalence class contains the edge $\gamma \to \beta$ or the edge $\delta \to \beta$ even though both are dashed in the DMEG. This example shows that not every equivalence class contains a least element.*

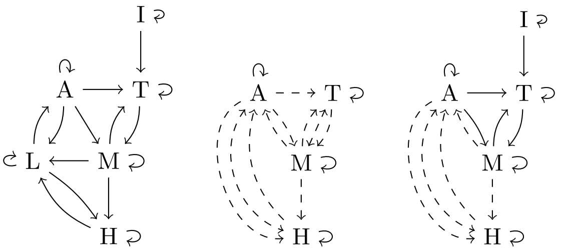


FIG. 9. *Left*: *Local independence graph of Example* 2.3. *Middle*: *DMEG for the marginalization over L and I*. *Right*: *DMEG for the marginalization over L*. *We have omitted the bidirected loops from the DMEGs and presented the directed loops as solid.*

We can now outline a two-step algorithm for constructing the DMEG from an arbitrary DMG, $\mathcal{G}$. We first construct the maximal Markov equivalent graph, $\mathcal{N}$. We know from Theorem 5.9 that one can simply check if each pair of nodes are potential siblings/parents in the independence model induced by $\mathcal{G}$ and construct the maximal Markov equivalent graph directly. This may, however, not be computationally efficient.

The above propositions show that given the maximal DMG, one can efficiently construct the DMEG by evaluating separability once for each directed edge and twice for each bidirected edge. Using Proposition 4.10, one can determine separability by testing a single separation statement, and this means that starting from $\mathcal{N}$, one can construct the corresponding DMEG in a way such that the number of separation statements to test scales linearly in the number of edges in $\mathcal{N}$.

EXAMPLE 5.13 (Gateway drugs, continued). We return to the model in Example 2.3 to consider what happens when it is only partially observed and to give an interpretation of the corresponding local independence model. The local independence graph is assumed to be as depicted on Figure 9, left.

Consider first the situation where L and I are unobserved. In this case, under the faithfulness assumption of the full model (Definition C.5) we can construct the DMEG, which is shown in the center panel of Figure 9, from the local independence model. The DMEG represents the Markov equivalence class which we can infer from the marginal local independence model (L and I are unobserved). Theoretically, the inference requires an oracle to provide us with local independence statements, which will in practice have to be approximated by statistical tests. What is noteworthy is that the DMEG can be inferred from the distribution of the observed variables only, and we do not need to know the local independences of the full model.

If we ignore which edges are dashed and which are not, the graph simply represents the local independence model of the marginal system as the maximal element in the Markov equivalence class. The dashed edges give us additional—and in some sense local—information. As an example, the directed edge from A to H is dashed and we cannot know if there exists a conditioning set that would render H locally independent of A in the full system. On the other hand, the directed edge from T to H is absent, and we can conclude that tobacco use is not directly affecting hard drug use.

Consider instead the situation where I is also observed. I serves as an analogue to an instrumental variable (see, e.g., [31] for an introduction to instrumental variables). The inclusion of this variable identifies some of the structure by removing some dashed edges and making others nondashed.

6. Discussion and conclusion. In this paper, we introduced a class of graphs to represent local independence structures of partially observed multivariate stochastic processes.

Previous work based on directed graphs, that allows for cycles and use the asymmetric δ-separation criterion, was extended to mixed directed graphs to account for latent processes and we introduced μ-separation in mixed directed graphs.

An important task is the characterization of equivalence classes of graphs and this has been studied, for example, in MAGs [5, 45]. In the case of MAGs, a key result is that every element in a Markov equivalence class has the same *skeleton*, that is, the same adjacencies [5]. As shown by Propositions 4.7 and 4.8, this is not the case for DMGs, and Example 4.12 shows that one cannot necessarily within a Markov equivalence class find an element such that two nodes are inseparable if and only if they are adjacent.

We proved instead a central maximality property which allowed us to propose the use of DMEGs to represent a Markov equivalence class of DMGs in a concise way. Given a maximal DMG, we furthermore argued that one can efficiently find the DMEG. Similar results are known for chain graphs, as one can also in a certain sense find a unique, largest graph representing a Markov equivalence class [20], though this graph is not a supergraph of all Markov equivalent graphs as in the case of DMGs. Volf and Studený [42] suggested to use this largest graph as a unique representative of the Markov equivalence class, and they provided an algorithm to construct it.

We emphasize that the characterization given of the maximal element of a Markov equivalence class of DMGs is constructive in the sense that it straightforwardly defines an algorithm for learning a maximal DMG from a local independence oracle. This learning algorithm may not be computationally efficient or even feasible for large graphs, and it is ongoing research to develop efficient learning algorithms and to develop the practical implementations of the tools needed for replacing the oracle by statistical tests.

Acknowledgments. The authors are grateful to Steffen Lauritzen for his helpful comments and suggestions. We also thank two referees and an area editor whose comments have helped improve this manuscript.

This work was supported by Villum Fonden Grant 13358.

SUPPLEMENTARY MATERIAL

Additional results and proofs (DOI: 10.1214/19-AOS1821SUPP; .pdf). The supplementary material consists of Sections A to F. In Sections A and B, we relate μ-separation to Didelez's δ-separation, and also relate our slightly different definitions of local independence. Section C describes how one can unroll a local independence graph and obtain a DAG. We use this to discuss Markov properties and faithfulness in the time series case. In Section D, we provide an augmentation criterion to determine μ-separation using an auxiliary undirected graph. In Section E, we discuss conditions for existence of compensators and elaborate on the definition of local independence. Section F contains the proofs of the main paper.

REFERENCES

[1] AALEN, O. O. (1987). Dynamic modelling and causality. *Scand. Actuar. J.* **3–4** 177–190. MR0943579 https://doi.org/10.1016/j.rser.2011.04.029

[2] AALEN, O. O., BORGAN, Ø., KEIDING, N. and THORMANN, J. (1980). Interaction between life history events. Nonparametric analysis for prospective and retrospective data in the presence of censoring. *Scand. J. Stat.* **7** 161–171. MR0605986

[3] AALEN, O. O., RØYSLAND, K., GRAN, J. M., KOUYOS, R. and LANGE, T. (2016). Can we believe the DAGs? A comment on the relationship between causal DAGs and mechanisms. *Stat. Methods Med. Res.* **25** 2294–2314. MR3553339 https://doi.org/10.1177/0962280213520436

[4] AALEN, O. O., RØYSLAND, K., GRAN, J. M. and LEDERGERBER, B. (2012). Causality, mediation and time: A dynamic viewpoint. *J. Roy. Statist. Soc. Ser. A* **175** 831–861. MR2993496 https://doi.org/10.1111/j.1467-985X.2011.01030.x

[5] Ali, R. A., Richardson, T. S. and Spirtes, P. (2009). Markov equivalence for ancestral graphs. *Ann. Statist.* **37** 2808–2837. MR2541448 https://doi.org/10.1214/08-AOS626

[6] Commenges, D. and Gégout-Petit, A. (2009). A general dynamical statistical model with causal interpretation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 719–736. MR2749916 https://doi.org/10.1111/j.1467-9868.2009.00703.x

[7] Constantinou, P. and Dawid, A. P. (2017). Extended conditional independence and applications in causal inference. *Ann. Statist.* **45** 2618–2653. MR3737904 https://doi.org/10.1214/16-AOS1537

[8] Cox, D. R. and Wermuth, N. (1996). *Multivariate Dependencies: Models, Analysis and Interpretation. Monographs on Statistics and Applied Probability* **67**. CRC Press, London. MR1456990

[9] Danks, D. and Plis, S. (2013). Learning causal structure from undersampled time series. In *JMLR: Workshop and Conference Proceedings (NIPS Workshop on Causality)*.

[10] Dawid, A. P. (2001). Separoids: A mathematical framework for conditional independence and irrelevance. *Ann. Math. Artif. Intell.* **32** 335–372. MR1859870 https://doi.org/10.1023/A:1016734104787

[11] Didelez, V. (2000). *Graphical models for event history analysis based on local independence*. Dissertation, Univ. Dortmund. Logos Verlag Berlin, Berlin. MR1864064

[12] Didelez, V. (2007). Graphical models for composable finite Markov processes. *Scand. J. Stat.* **34** 169–185. MR2325249 https://doi.org/10.1111/j.1467-9469.2006.00528.x

[13] Didelez, V. (2008). Graphical models for marked point processes based on local independence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 245–264. MR2412641 https://doi.org/10.1111/j.1467-9868.2007.00634.x

[14] Eichler, M. (2012). Graphical modelling of multivariate time series. *Probab. Theory Related Fields* **153** 233–268. MR2925574 https://doi.org/10.1007/s00440-011-0345-8

[15] Eichler, M. (2013). Causal inference with multiple time series: Principles and problems. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **371** 20110613, 17. MR3081173 https://doi.org/10.1098/rsta.2011.0613

[16] Eichler, M. and Didelez, V. (2007). Causal reasoning in graphical time series models. In *Proceedings of the* 23*rd Conference on Uncertainty in Artificial Intelligence* 109–116.

[17] Eichler, M. and Didelez, V. (2010). On Granger causality and the effect of interventions in time series. *Lifetime Data Anal.* **16** 3–32. MR2575937 https://doi.org/10.1007/s10985-009-9143-3

[18] Evans, R. J. (2016). Graphs for margins of Bayesian networks. *Scand. J. Stat.* **43** 625–648. MR3543314 https://doi.org/10.1111/sjos.12194

[19] Evans, R. J. and Richardson, T. S. (2014). Markovian acyclic directed mixed graphs for discrete data. *Ann. Statist.* **42** 1452–1482. MR3262457 https://doi.org/10.1214/14-AOS1206

[20] Frydenberg, M. (1990). The chain graph Markov property. *Scand. J. Stat.* **17** 333–353. MR1096723

[21] Gégout-Petit, A. and Commenges, D. (2010). A general definition of influence between stochastic processes. *Lifetime Data Anal.* **16** 33–44. MR2575938 https://doi.org/10.1007/s10985-009-9131-7

[22] Hyttinen, A., Plis, S., Järvisalo, M., Eberhardt, F. and Danks, D. (2016). Causal discovery from subsampled time series data by constraint optimization. In *Proceedings of the Eighth International Conference on Probabilistic Graphical Models* **52** 216–227.

[23] Jensen, A.-M. and Schweder, T. (1986). The engine of fertility—Influenced by interbirth employment? Discussion paper, Central Bureau of Statistics, Oslo.

[24] Kandel, D. (1975). Stages in adolescent involvement in drug use. *Science* **190** 912–914.

[25] Koster, J. T. A. (1999). On the validity of the Markov interpretation of path diagrams of Gaussian structural equations systems with correlated errors. *Scand. J. Stat.* **26** 413–431. MR1712043 https://doi.org/10.1111/1467-9469.00157

[26] Lauritzen, S. and Sadeghi, K. (2018). Unifying Markov properties for graphical models. *Ann. Statist.* **46** 2251–2278. MR3845017 https://doi.org/10.1214/17-AOS1618

[27] Lu, J. (2015). *On Causal Inference for Ordinal Outcomes*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–Harvard Univ. MR3664965

[28] Meek, C. (2014). Toward learning graphical and causal process models. In *Proceedings of the UAI* 2014 *Workshop Causal Inference: Learning and Prediction*.

[29] Mogensen, S. W. and Hansen, N. R. (2020). Supplement to "Markov equivalence of marginalized local independence graphs." https://doi.org/10.1214/19-AOS1821SUPP.

[30] Mogensen, S. W., Malinsky, D. and Hansen, N. R. (2018). Causal learning for partially observed stochastic dynamical systems. In *Proceedings of the* 34*th Conference on Uncertainty in Artificial Intelligence*.

[31] Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2548166 https://doi.org/10.1017/CBO9780511803161

[32] Richardson, T. (2003). Markov properties for acyclic directed mixed graphs. *Scand. J. Stat.* **30** 145–157. MR1963898 https://doi.org/10.1111/1467-9469.00323

[33] Richardson, T. and Spirtes, P. (2002). Ancestral graph Markov models. *Ann. Statist.* **30** 962–1030. MR1926166 https://doi.org/10.1214/aos/1031689015

[34] Richardson, T. S., Evans, R. J., Robins, J. M. and Shpitser, I. (2017). Nested Markov properties for acyclic directed mixed graphs. https://arxiv.org/abs/1701.06686.

[35] Røysland, K. (2012). Counterfactual analyses with graphical models based on local independence. *Ann. Statist.* **40** 2162–2194. MR3059080 https://doi.org/10.1214/12-AOS1031

[36] Sadeghi, K. (2013). Stable mixed graphs. *Bernoulli* **19** 2330–2358. MR3160556 https://doi.org/10.3150/12-BEJ454

[37] Schweder, T. (1970). Composable Markov processes. *J. Appl. Probab.* **7** 400–410. MR0264755 https://doi.org/10.2307/3211973

[38] Sokol, A. and Hansen, N. R. (2014). Causal interpretation of stochastic differential equations. *Electron. J. Probab.* **19** no. 100, 24. MR3275852 https://doi.org/10.1214/ejp.v19-2891

[39] Spirtes, P., Richardson, T. S. and Meek, C. (1997). The dimensionality of mixed ancestral graphs. Technical report. Philosophy Dept, Carnegie Mellon Univ., Pittsburgh, PA.

[40] Vanyukov, M. M., Tarter, R. E., Kirillova, G. P., Kirisci, L., Reynolds, M. D., Kreek, M. J., Conway, K. P., Maher, B. S., Iacono, W. G., Bierut, L., Neale, M. C., Clark, D. B. and Ridenour, T. A. (2012). Common liability to addiction and "gateway hypothesis": Theoretical, empirical and evolutionary perspective. *Drug Alcohol Depend.* **123**.

[41] Verma, T. and Pearl, J. (1991). Equivalence and synthesis of causal models Technical Report Univ. California, Los Angeles.

[42] Volf, M. and Studený, M. (1999). A graphical characterization of the largest chain graphs. *Internat. J. Approx. Reason.* **20** 209–236. MR1685080 https://doi.org/10.1016/S0888-613X(99)00003-1

[43] World Health Organization (2004). Neuroscience of psychoactive substance use and dependence Technical Report World Health Organization.

[44] Xu, H., Farajtabar, M. and Zha, H. (2016). Learning Granger causality for Hawkes processes. In *Proceedings of the* 33*rd International Conference on Machine Learning*.

[45] Zhao, H., Zheng, Z. and Liu, B. (2005). On the Markov equivalence of maximal ancestral graphs. *Sci. China Ser. A* **48** 548–562. MR2157690 https://doi.org/10.1360/04ys0023

SUPPLEMENTARY MATERIAL FOR MARKOV EQUIVALENCE OF MARGINALIZED LOCAL INDEPENDENCE GRAPHS

By Søren Wengel Mogensen and Niels Richard Hansen

University of Copenhagen

In this supplementary material we discuss relations between μ-separation and other asymmetric notions of graphical separation. We also compare our proposed definition of local independence to previous definitions to argue that ours is in fact a generalization. We furthermore relate μ-separation to m-separation. We provide, in particular, a detailed discussion of the local independence model for discrete-time stochastic processes (time series), and we show how to verify μ-separation via separation in an auxiliary undirected graph. We also discuss the existence of the compensators that are used in the definition of local independence for continuous-time stochastic process models. This supplementary material also contains proofs of the results of the main paper. A list of references can be found on the last page.

A. Relation to other asymmetric notions of graphical separation. In this section we relate μ-separation to δ-separation as introduced previously in the literature for directed graphs.

Definition A.1 (Bereaved graph). Let $\mathcal{G} = (V, E_d)$ be a DG, and let $B \subseteq V$. The B-*bereaved graph*, $\mathcal{G}^B$, is constructed from $\mathcal{G}$ by removing every directed edge with a tail at a node in B except loops. More precisely, $\mathcal{G}^B = (V, \bar{E}_d^B)$, where $\bar{E}_d^B = E_d \setminus \left(\bigcup_{\beta \in B} \{(\beta, \delta) \mid \delta \neq \beta\} \right)$.

Didelez [2] considered a DG, and for disjoint sets $A, B, C \subseteq V$ said that B is separated from A by C if there is no μ-connecting walk in $\mathcal{G}^B$, or equivalently, no μ-connecting path. This is called δ-separation. Note that the condition in Definitions 3.1 and 3.2 that a connecting walk be nontrivial makes no difference now due to A and B being disjoint. The condition that a μ-connecting walk ends with a head at $\beta \in B$ is also obsolete as we are evaluating separation in the bereaved graph $\mathcal{G}^B$. Didelez [2] always assumed that a process depended on its own past, and thus did not visualize loops in the DGs as a loop would always be present at every node.

Meek [9] generalized δ-separation to δ^*-separation in a DG (allowing for loops) by considering only nontrivial μ-connecting walks in $\mathcal{G}^B$ for sets

$A, B, C \subseteq V$ such that $A \cap C = \emptyset$ with the motivation that a node can be separated from itself using this notion of separation. However, if we consider the graph $\alpha \to \beta$, and sets $A = \{\alpha\}$, $B = \{\alpha, \beta\}$, $C = \emptyset$, then using δ^*-separation, B is separated from A given C, which runs counter to an intuitive understanding of separation. More importantly, δ^*-separation in the local independence graph will not generally imply local independence.

To establish an exact relationship between δ- and μ-separations and argue that we are indeed proposing a generalization of the former, assume that $\mathcal{G}$ is a DG and that $A, B, C \subseteq V$ are disjoint. We will argue that

$$A \perp_\mu B \mid C \cup B \; [\mathcal{G}] \Leftrightarrow A \perp_\delta B \mid C \; [\mathcal{G}]. \tag{A.1}$$

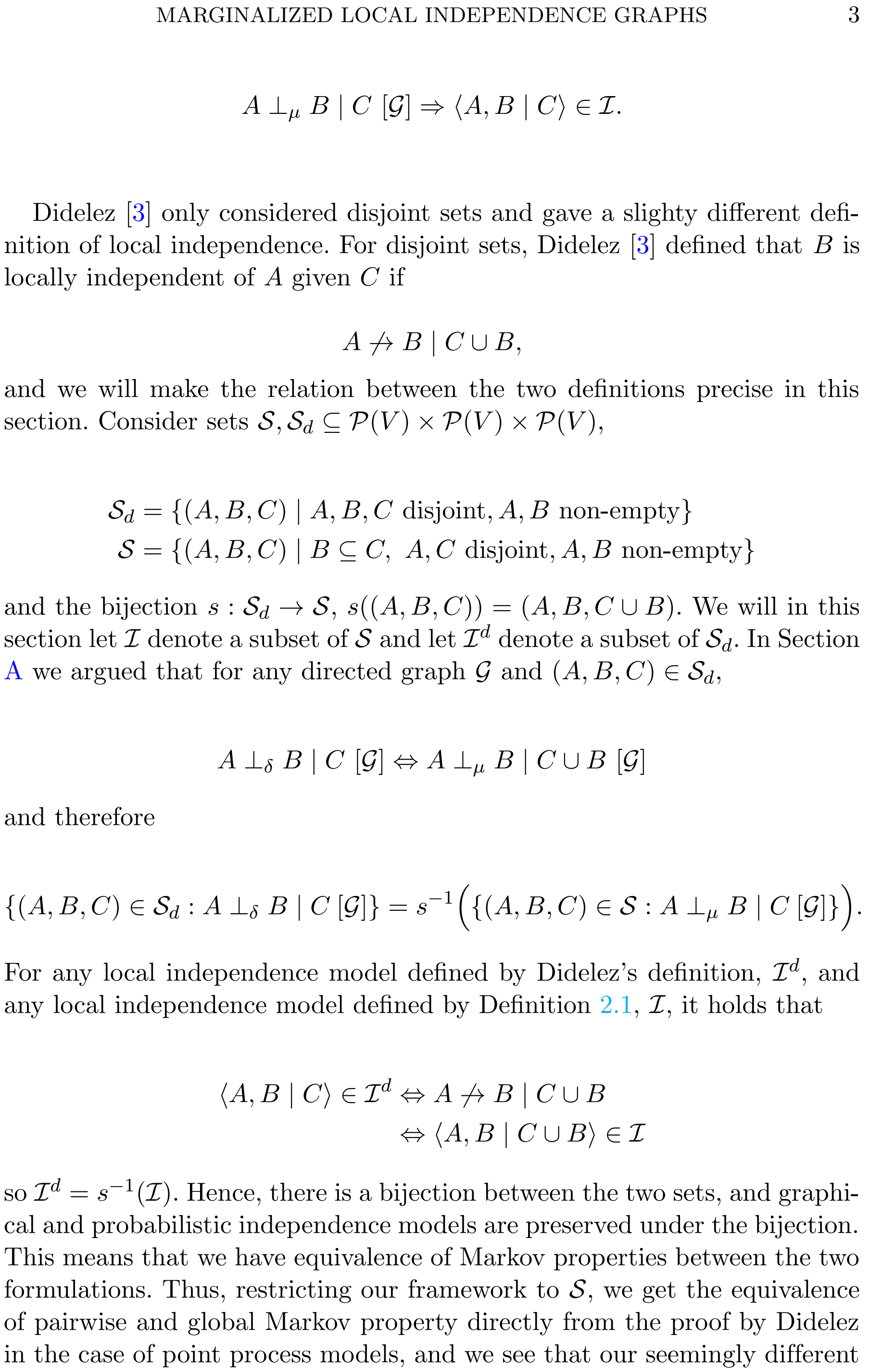
To see that this is the case, consider first a δ-connecting walk from $\alpha \in A$ to $\beta \in B$ given C in $\mathcal{G}^B$, ω. The subwalk from α to the first node on ω which is in B is also present and μ-connecting given $C \cup B$ in $\mathcal{G}$. On the other hand, assume that there exists a μ-connecting sequence, ω, in $\mathcal{G}$. We know that $A \cap B = \emptyset$, and because B is a subset of the conditioning set on the left hand side in (A.1), we must have that the first time the path enters B, it has a head at the node in B, and this implies that a subwalk of ω is δ-connecting, that is, present and connecting in $\mathcal{G}^B$. In Section B we will discuss why B is included in the conditioning set on the left side of (A.1).

B. Markov properties. The equivalence of pairwise and global Markov properties is pivotal in much of graphical modeling. In this section, we will show how our proposed graphical framework fits with known results on Markov properties in the case of point processes and argue that our graphical framework is a generalization of that of Didelez [3] to allow for non-disjoint sets and unobserved processes.

Definition B.1 (The pairwise Markov property). Let $\mathcal{I}$ be an independence model over V. We say that $\mathcal{I}$ satisfies the *pairwise Markov property* with respect to the DG $\mathcal{D}$ if for all $\alpha, \beta \in V$,

$$\alpha \not\to_{\mathcal{D}} \beta \Rightarrow \langle \alpha, \beta \mid V \setminus \{\alpha\} \rangle \in \mathcal{I}.$$

Definition B.2 (The global Markov property). Let $A, B, C \subseteq V$. Let $\mathcal{I}$ be an independence model over V. We say that $\mathcal{I}$ satisfies the *global Markov property* with respect to the DMG $\mathcal{G}$ if $\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}$, i.e., if



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

Didelez [3] only considered disjoint sets and gave a slighty different definition of local independence. For disjoint sets, Didelez [3] defined that B is locally independent of A given C if

$$A \not\rightarrow B \mid C \cup B,$$

and we will make the relation between the two definitions precise in this section. Consider sets $\mathcal{S}, \mathcal{S}_d \subseteq \mathcal{P}(V) \times \mathcal{P}(V) \times \mathcal{P}(V)$,

$$\begin{aligned} \mathcal{S}_d &= \{(A, B, C) \mid A, B, C \text{ disjoint}, A, B \text{ non-empty}\} \\ \mathcal{S} &= \{(A, B, C) \mid B \subseteq C,\; A, C \text{ disjoint}, A, B \text{ non-empty}\} \end{aligned}$$

and the bijection $s : \mathcal{S}_d \to \mathcal{S}$, $s((A, B, C)) = (A, B, C \cup B)$. We will in this section let $\mathcal{I}$ denote a subset of $\mathcal{S}$ and let $\mathcal{I}^d$ denote a subset of $\mathcal{S}_d$. In Section A we argued that for any directed graph $\mathcal{G}$ and $(A, B, C) \in \mathcal{S}_d$,

$$A \perp_\delta B \mid C \; [\mathcal{G}] \Leftrightarrow A \perp_\mu B \mid C \cup B \; [\mathcal{G}]$$

and therefore

$$\{(A, B, C) \in \mathcal{S}_d : A \perp_\delta B \mid C \, [\mathcal{G}]\} = s^{-1}\Big(\{(A, B, C) \in \mathcal{S} : A \perp_\mu B \mid C \, [\mathcal{G}]\}\Big).$$

For any local independence model defined by Didelez's definition, $\mathcal{I}^d$, and any local independence model defined by Definition 2.1, $\mathcal{I}$, it holds that

$$\begin{aligned} \langle A, B \mid C \rangle \in \mathcal{I}^d &\Leftrightarrow A \not\rightarrow B \mid C \cup B \\ &\Leftrightarrow \langle A, B \mid C \cup B \rangle \in \mathcal{I} \end{aligned}$$

so $\mathcal{I}^d = s^{-1}(\mathcal{I})$. Hence, there is a bijection between the two sets, and graphical and probabilistic independence models are preserved under the bijection. This means that we have equivalence of Markov properties between the two formulations. Thus, restricting our framework to $\mathcal{S}$, we get the equivalence of pairwise and global Markov property directly from the proof by Didelez in the case of point process models, and we see that our seemingly different definitions of local independence and graphical separation indeed give an extension of earlier work.

One can show that for two DMGs $\mathcal{G}_1$, $\mathcal{G}_2$, that both have all directed and bidirected loops it holds that

$$\mathcal{I}(\mathcal{G}_1) \cap \mathcal{S} = \mathcal{I}(\mathcal{G}_2) \cap \mathcal{S} \Leftrightarrow \mathcal{I}(\mathcal{G}_1) = \mathcal{I}(\mathcal{G}_2).$$

Let $\mathbb{G}$ denote the class of DMGs such that all directed and bidirected loops are present. Consider now some $\mathcal{G} \in \mathbb{G}$. By the above result we can identify the Markov equivalence class from the independence model restricted to $\mathcal{S}$. This equivalence class has a maximal element which is also in $\mathbb{G}$ and thus one can also in this case represent the Markov equivalence class using a DMEG.

C. Time series and unrolled graphs. In this section we first relate the cyclic DGs and DMGs to acyclic graphs and then use this to discuss Markov properties (see Definition B.2) and faithfulness of local independence models in the time series case.

Definition C.1 (m-separation [10]). Let $\mathcal{G} = (V, E)$ be a DMG and let $\alpha, \beta \in V$. A path between α and β is said to be m-connecting if no noncollider on the path is in C and every collider on the path is in $An(C)$. For disjoint sets $A, B, C \subseteq V$, we say that A and B are m-separated by C if there is no m-connecting path between $\alpha \in A$ and $\beta \in B$. In this case, we write $A \perp_m B \mid C$.

The above m-separation is a generalization of the well-known d-separation in DAGs. In this section we will only consider m-separation for DAGs, and will thus use the d-separation terminology. In Section D we provide a more general relation between μ-separation and m-separation.

We first describe how to obtain a DAG from a DG such that the DAG, if read the right way, will give the same separation model as the DG. This can be useful in time series examples as well as when working with continuous-time models. Sokol and Hansen [13] studied solutions to stochastic differential equations and used a DAG in discrete time to approximate the continuous-time dynamics. Danks and Plis [1] and Hyttinen et al. [5] used similar translations between an *unrolled* graph in which time is discrete and explicit and a *rolled* graph in which time is implicit. Some authors use the term *unfolded* instead of unrolled. In a rolled graph each node represents a stochastic process whereas in an unrolled graph each node represents

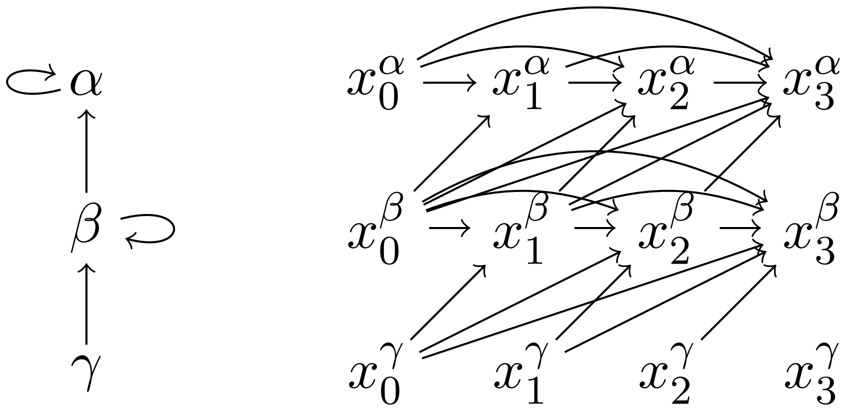


Fig 1. *A directed graph (left) and the corresponding unrolled version with four time points, $\mathcal{D}_3(\mathcal{G})$, (right). x_t^δ denotes the δ-coordinate process at time t for $\delta \in \{\alpha, \beta, \gamma\}$.*

a single random variable. Definition C.2 shows how to unroll a local independence graph and Lemma C.3 establishes a precise relationship between independence models in the rolled and unrolled graphs.

Definition C.2. Let $\mathcal{G} = (V, E)$ be a DG and let $T \in \mathbb{N}$. The unrolled version of $\mathcal{G}$, $\mathcal{D}_T(\mathcal{G}) = (\bar{V}, \bar{E})$, is the DAG on nodes

$$\bar{V} = \{x_t^\alpha \mid (t, \alpha) \in \{0, 1, \dots, T\} \times V\}$$

and with edges

$$\bar{E} = \{x_s^\alpha \to x_t^\beta \mid \alpha \to_{\mathcal{G}} \beta \text{ and } s < t\}.$$

Let $D \subseteq V$ and let $T \in \mathbb{N}$. We define $D_{0:T} = \{x_t^\alpha \in \bar{V} \mid \alpha \in D,\ t \leq T\}$ and $D_T = \{x_t^\alpha \in \bar{V} \mid \alpha \in D,\ t = T\}$.

Lemma C.3. Let $\mathcal{G} = (V, E)$ be a DG. If $A \perp_\mu B \mid C\ [\mathcal{G}]$ then $(A \setminus C)_{0:(T-1)} \perp_d B_T \mid C_{0:(T-1)}\ [\mathcal{D}_T(\mathcal{G})]$. For large enough values of T, the opposite implication holds as well.

Proof. Assume first that $\langle x_{s_0}^{\alpha_0}, e_1, x_{s_1}^{\alpha_1}, \dots, e_l, x_{s_l}^{\alpha_l} \rangle$ is a d-connecting path in $\mathcal{D}_T(\mathcal{G})$. This path has a head at $x_{s_l}^{\alpha_l} \in B_T$. Construct a walk in $\mathcal{G}$ by for each node, $x_{s_k}^{\alpha_k}$, taking the corresponding node, α_k, and for each edge

$x^{\alpha_k}_{s_k} \sim x^{\alpha_{k+1}}_{s_{k+1}}$ taking the corresponding, endpoint-identical edge $\alpha_k \sim \alpha_{k+1}$ in $\mathcal{G}$. On this walk, no noncollider is in C, and every collider is an ancestor of a node in C.

Assume instead that ω is a μ-connecting walk in $\mathcal{G}$ from A to B given C,

$$\alpha_1 \sim \ldots \sim \alpha_{l-1} \to \alpha_l$$

and let $T \geq 3(|E|+1)+1$. Using Proposition 3.5, we can assume that ω has length smaller than or equal to $|E|+1$. We construct a d-connecting walk in $\mathcal{D}_T(\mathcal{G})$ in the following way. Starting from $x^{\alpha_l}_T$, we choose the edge between $x^{\alpha_{l-1}}_{|E|+1}$ and $x^{\alpha_l}_T$. For the remaining edges, $\alpha_k \sim \alpha_{k+1}$, we choose the edge $x^{\alpha_k}_{s_k-1} \to x^{\alpha_{k+1}}_{s_k}$ if $\alpha_k \to \alpha_{k+1}$ in ω, and $x^{\alpha_{k+1}}_{s_k} \to x^{\alpha_k}_{s_k+1}$ if $\alpha_k \leftarrow \alpha_{k+1}$ in ω where s_k is determined by the endpoints of the previous edge. No noncollider on this walk will be in $C_{0:(T-1)}$. Every collider will be in $An_{\mathcal{D}_T(\mathcal{G})}(C_{0:(T-1)})$ as the collider will be in the time slices 0 to $2(|E|+1)$. This d-connecting walk can be trimmed down to a d-connecting path. □

We defined local independence for a class of continuous-time processes in Definition 2.1. In this section we define a similar notion for time series, as also introduced in [4]. Let $V = \{1, \ldots, n\}$. We consider a multivariate time series $(X_t)_{t \in \mathbb{N} \cup \{0\}}$, $X_t = (X^1_t, \ldots, X^n_t)$, of the form

$$X^\alpha_t = f_{\alpha t}(X_{s<t}, \varepsilon^\alpha_t),$$

where $X_{s<t} = \{X^\alpha_u \mid \alpha \in V, u < t\}$. The random variables $\{\varepsilon^\alpha_t\}$ are independent. For $S \subseteq \mathbb{N} \cup \{0\}$ and $D \subseteq V$ we let $X^D_S = \{X^\alpha_s \mid \alpha \in D, s \in S\}$ and $X^D = \{X^\alpha \mid \alpha \in D\}$. In the case of time series, a notable feature of local independence and local independence graphs is that they provide a simple representation in comparison with graphs in which each vertex represents a single time-point variable.

Definition C.4 (Local independence, time series). Let X be a multivariate time series. We say that X^B is *locally independent* of X^A given X^C if for all $t \in \mathbb{N}$, $\beta \in B$, $X^A_{s<t}$ and X^β_t are conditionally independent given $X^C_{s<t}$, that is,

$$X^A_{s<t} \perp\!\!\!\perp X^\beta_t \mid X^C_{s<t}$$

and write $A \not\to B \mid C$.

The above definition induces an independence model over V, which we will also refer to as the local independence model and denote $\mathcal{I}$ in the following. The main question that we address is whether this independence model is graphical. That is, we will construct a DG, consider the Markov and faithfulness properties of $\mathcal{I}$ and this DG, and relate them to Markov and faithfulness properties of the conditional independence model of finite distributions and unrolled versions of the DG.

Definition C.5 (Faithfulness). Let $A, B, C \subseteq V$. Let $\mathcal{I}$ be an independence model on V and let $\mathcal{G}$ be a DMG. We say that $\mathcal{I}$ and $\mathcal{G}$ are *faithful* if $\mathcal{I} = \mathcal{I}(\mathcal{G})$, i.e., if

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

One can give analogous definitions using other notions of graphical separation. Below we also consider faithfulness of a probability distribution and a DAG, implicitly using d-separation instead of μ-separation in the above definition.

Let $\mathcal{D}_T$ for $T \geq 1$ be the DAG on nodes $\{x_s^\alpha \mid s \in \{0, \ldots, T\}, \alpha \in V\}$ such that there is an edge $x_s^\alpha \to x_t^\beta$ if and only if $f_{\beta t}$ depends on the argument X_s^α. Let $D_S = \{x_s^\alpha \mid \alpha \in D, s \in S\}$. Let $\mathcal{G}$ denote the minimal DG such that its unrolled version, $\mathcal{D}_T(\mathcal{G})$, is a supergraph of $\mathcal{D}_T$ for all $T \in \mathbb{N}$.

For all $T \in \mathbb{N}$, the DAG $\mathcal{D}_T(\mathcal{G})$ and the distribution of $X_{s \leq T}$ satisfy

$$x_s^\alpha, x_t^\beta \text{ not adjacent } \Rightarrow X_s^\alpha \perp\!\!\!\perp X_t^\beta \mid (An(X_s^\alpha) \cup An(X_t^\beta)) \setminus \{X_s^\alpha, X_t^\beta\},$$

which is also known as the pairwise Markov property for DAGs. Assume equivalence of the pairwise and global Markov properties for this DAG and the finite-dimensional distribution (see e.g. [7] for necessary and sufficient conditions for this equivalence). Assume that B is μ-separated from A by C in the DG $\mathcal{G}$, $A \perp_\mu B \mid C \; [\mathcal{G}]$. By Lemma C.3, $(A \setminus C)_{s<T} \perp_m B_T \mid C_{s<T} \; [\mathcal{D}_T(\mathcal{G})]$, and by the global Markov property in this DAG, $X_{s<T}^{A \setminus C} \perp\!\!\!\perp X_T^B \mid X_{s<T}^C$. This holds for any T, and therefore $A \setminus C \not\to B \mid C$. It follows that $A \not\to B \mid C$. This means that $\mathcal{I}$ satisfies the global Markov property with respect to $\mathcal{G}$.

Assume furthermore that the distribution of X_T and the DAG $\mathcal{D}_T(\mathcal{G})$ for some $T \in \mathbb{N}$ are faithful and that $T \geq 3(|E| + 1) + 1$. Meek [8] studied faithfulness of DAGs and argued that faithful distributions exist for any DAG. If $A \not\to B \mid C$, then $A \setminus C \not\to B \mid C$ and $X_{s<T}^{A \setminus C} \perp\!\!\!\perp X_T^B \mid X_{s<T}^C$. By faithfulness of the distribution of X_T and the DAG $\mathcal{D}_T(\mathcal{G})$, we have $(A \setminus C)_{s<T} \perp_m B_T \mid C_{s<T} \; [\mathcal{D}_T(\mathcal{G})]$ and using Lemma C.3 this implies that $A \perp_\mu B \mid C \; [\mathcal{G}]$, giving us faithfulness of $\mathcal{I}$ and $\mathcal{G}$.

In summary, for every DG there exists a time series such that the local independence model induced by its distribution and the DG are faithful.

D. An augmentation criterion. In this section we present results that allow us to determine μ-separation from graphical separation in an undirected graph. An *undirected graph* is a graph, (V, E), with an edge set that consists of unordered pairs of nodes such that every edge is of the type $-$. Let A, B, and C be disjoint subsets of V. We say that A and B are separated by C if every path between $\alpha \in A$ and $\beta \in B$ contains a node in C.

When working with d-separation in DAGs, it is possible to give an equivalent separation criterion using a derived undirected graph, the *moral* graph [6]. Didelez [2] also gives both pathwise and so-called moral graph criteria for δ-separation. The augmented graph below is a generalization of the moral graph [10, 11] which allows one to give a criterion for m-separation based on an augmented graph. We use the similarity of μ-separation and m-separation to give an augmentation graph criterion for μ-separation. The first step in making a connection to m-separation is to explicate that each node of a DMG represents an entire stochastic process, and notably, both the past and the present of that process. We do that using graphs of the below type.

Definition D.1. Let $\mathcal{G} = (V, E)$ and let $B = \{\beta_1, \ldots, \beta_k\} \subseteq V$. The B-history version of $\mathcal{G}$, denoted by $\mathcal{G}(B)$, is the DMG with node set $V \dot\cup \{\beta_1^p, \ldots, \beta_k^p\}$ such that $\mathcal{G}(B)_V = \mathcal{G}$ and

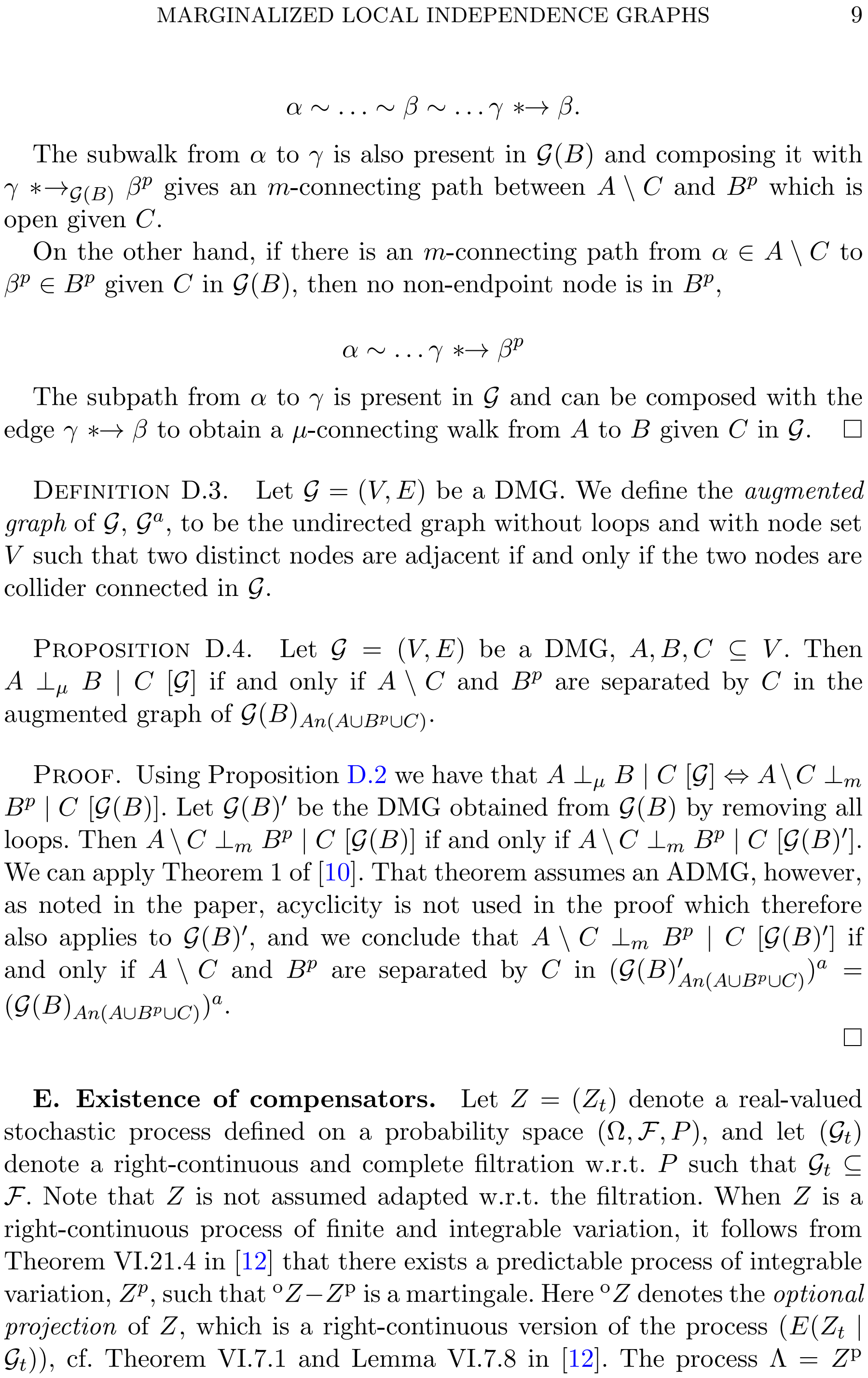
- $\alpha \leftrightarrow_{\mathcal{G}(B)} \beta_i^p$ if $\alpha \leftrightarrow_{\mathcal{G}} \beta_i$ and $\alpha \in V, \beta_i \in B$,
- $\alpha \rightarrow_{\mathcal{G}(B)} \beta_i^p$ if $\alpha \rightarrow_{\mathcal{G}} \beta_i$ and $\alpha \in V, \beta_i \in B$.

$\mathcal{G}(B)$ is a graph such that every node $b \in B$ is simply split in two: one that represents the present and one that represents the past. We define $B^p = \{\beta_1^p, \ldots, \beta_k^p\}$.

Proposition D.2. Let $\mathcal{G} = (V, E)$ be a DMG, and let $A, B, C \subseteq V$. Then

$$A \perp_\mu B \mid C \; [\mathcal{G}] \Leftrightarrow A \setminus C \perp_m B^p \mid C \; [\mathcal{G}(B)].$$

Proof. Assume first that there is a μ-connecting walk from $\alpha \in A$ to $\beta \in B$ given C in $\mathcal{G}$. By definition $\alpha \in A \setminus C$. By Proposition 3.5 there is a μ-connecting route,



MARGINALIZED LOCAL INDEPENDENCE GRAPHS 9

$$\alpha \sim \ldots \sim \beta \sim \ldots \gamma \ast\!\!\rightarrow \beta.$$

The subwalk from α to γ is also present in $\mathcal{G}(B)$ and composing it with $\gamma \ast\!\!\rightarrow_{\mathcal{G}(B)} \beta^p$ gives an m-connecting path between $A \setminus C$ and B^p which is open given C.

On the other hand, if there is an m-connecting path from $\alpha \in A \setminus C$ to $\beta^p \in B^p$ given C in $\mathcal{G}(B)$, then no non-endpoint node is in B^p,

$$\alpha \sim \ldots \gamma \ast\!\!\rightarrow \beta^p$$

The subpath from α to γ is present in $\mathcal{G}$ and can be composed with the edge $\gamma \ast\!\!\rightarrow \beta$ to obtain a μ-connecting walk from A to B given C in $\mathcal{G}$. □

Definition D.3. Let $\mathcal{G} = (V, E)$ be a DMG. We define the *augmented graph* of $\mathcal{G}$, $\mathcal{G}^a$, to be the undirected graph without loops and with node set V such that two distinct nodes are adjacent if and only if the two nodes are collider connected in $\mathcal{G}$.

Proposition D.4. Let $\mathcal{G} = (V, E)$ be a DMG, $A, B, C \subseteq V$. Then $A \perp_\mu B \mid C$ $[\mathcal{G}]$ if and only if $A \setminus C$ and B^p are separated by C in the augmented graph of $\mathcal{G}(B)_{An(A \cup B^p \cup C)}$.

Proof. Using Proposition D.2 we have that $A \perp_\mu B \mid C$ $[\mathcal{G}] \Leftrightarrow A \setminus C \perp_m B^p \mid C$ $[\mathcal{G}(B)]$. Let $\mathcal{G}(B)'$ be the DMG obtained from $\mathcal{G}(B)$ by removing all loops. Then $A \setminus C \perp_m B^p \mid C$ $[\mathcal{G}(B)]$ if and only if $A \setminus C \perp_m B^p \mid C$ $[\mathcal{G}(B)']$. We can apply Theorem 1 of [10]. That theorem assumes an ADMG, however, as noted in the paper, acyclicity is not used in the proof which therefore also applies to $\mathcal{G}(B)'$, and we conclude that $A \setminus C \perp_m B^p \mid C$ $[\mathcal{G}(B)']$ if and only if $A \setminus C$ and B^p are separated by C in $(\mathcal{G}(B)'_{An(A \cup B^p \cup C)})^a = (\mathcal{G}(B)_{An(A \cup B^p \cup C)})^a$.

□

E. Existence of compensators. Let $Z = (Z_t)$ denote a real-valued stochastic process defined on a probability space $(\Omega, \mathcal{F}, P)$, and let $(\mathcal{G}_t)$ denote a right-continuous and complete filtration w.r.t. P such that $\mathcal{G}_t \subseteq \mathcal{F}$. Note that Z is not assumed adapted w.r.t. the filtration. When Z is a right-continuous process of finite and integrable variation, it follows from Theorem VI.21.4 in [12] that there exists a predictable process of integrable variation, Z^p, such that ${}^oZ - Z^p$ is a martingale. Here ${}^oZ$ denotes the *optional projection* of Z, which is a right-continuous version of the process $(E(Z_t \mid \mathcal{G}_t))$, cf. Theorem VI.7.1 and Lemma VI.7.8 in [12]. The process $\Lambda = Z^p$

is called the dual predictable projection or compensator of the optional projection ${}^{o}Z$ as well as of the process Z itself. It depends on the filtration $(\mathcal{G}_t)$.

If Z is adapted w.r.t. a (right-continuous and complete) filtration $(\mathcal{F}_t)$, it has a compensator $\tilde{\Lambda} = Z^{\mathrm{p}}$ such that $Z - \tilde{\Lambda}$ is an $\mathcal{F}_t$ martingale. When $\mathcal{G}_t \subseteq \mathcal{F}_t$ it may be of interest to understand the relation between Λ, as defined above w.r.t. $(\mathcal{G}_t)$, and $\tilde{\Lambda}$. If $\tilde{\Lambda}$ is continuous with $\tilde{\Lambda}_0 = 0$, say, we may ask if Λ equals the *predictable projection*, $E(\tilde{\Lambda}_t \mid \mathcal{G}_{t-})$. As $\tilde{\Lambda}$ is assumed continuous and is of finite variation,

$$\tilde{\Lambda}_t = \int_0^t \tilde{\lambda}_s \mathrm{d}s.$$

If $(\tilde{\lambda}_t)$ itself is an integrable right-continuous process, then its optional projection, $(E(\tilde{\lambda}_t \mid \mathcal{G}_t))$, is an integrable right-continuous process, and

$$E(\tilde{\Lambda}_t \mid \mathcal{G}_{t-}) = \int_0^t E(\tilde{\lambda}_s \mid \mathcal{G}_s)\mathrm{d}s$$

is a finite-variation, continuous version of the predictable projection of $\tilde{\Lambda}$. It is clear that

$$E(Z_t \mid \mathcal{G}_t) - \int_0^t E(\tilde{\lambda}_s \mid \mathcal{G}_s)\mathrm{d}s$$

is a $\mathcal{G}_t$ martingale, thus

$$\Lambda_t = \int_0^t E(\tilde{\lambda}_s \mid \mathcal{G}_s)\mathrm{d}s$$

is a compensator of Z w.r.t. the filtration $(\mathcal{G}_t)$.

We formulate the consequences of the discussion as a criterion for determining local independence via the computation of conditional expectations. The setup is as in Definition 2.1 in Section 2.1.

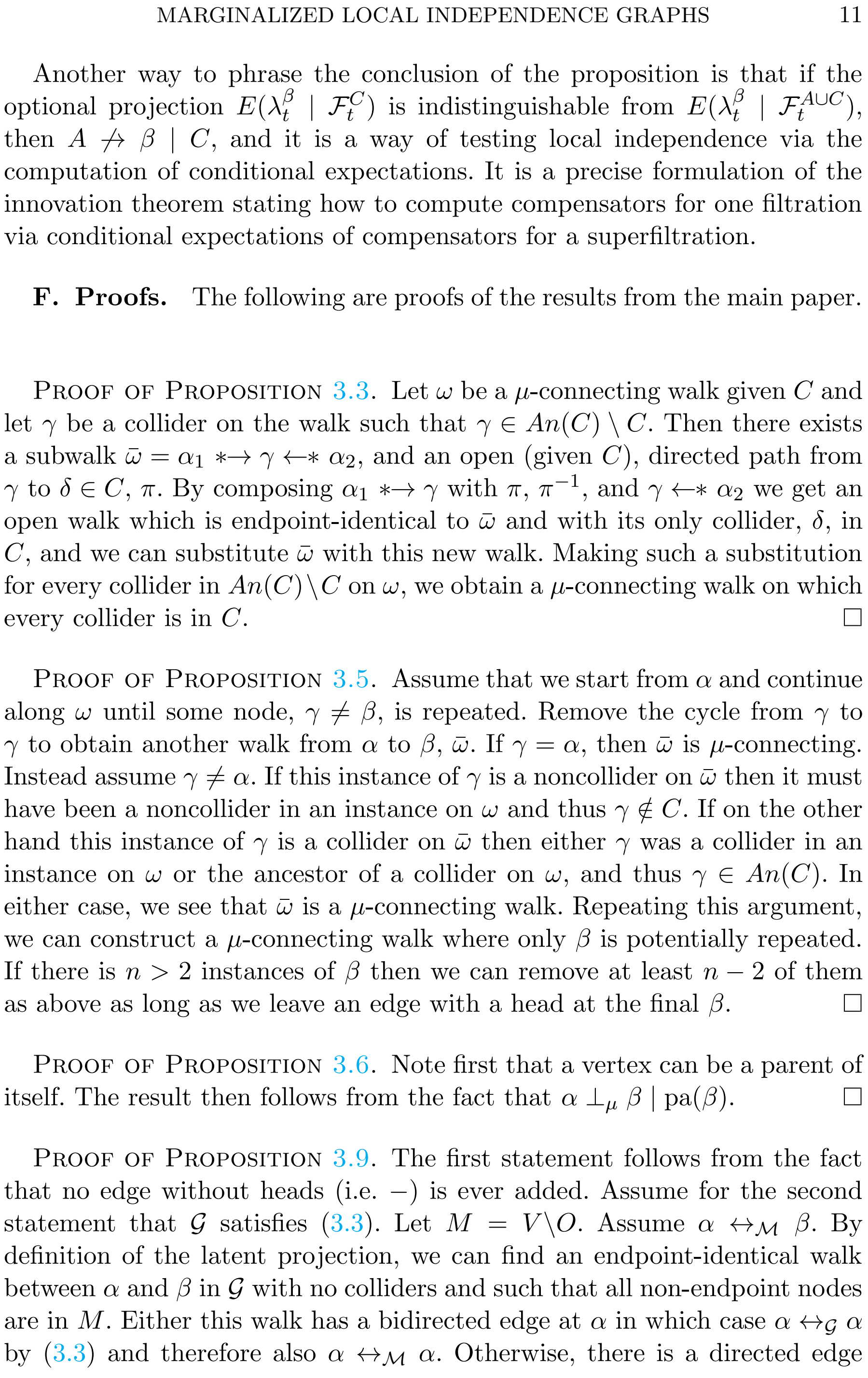
Proposition E.1. Assume that the process X^β for all $\beta \in V$ has a compensator w.r.t. the filtration $(\mathcal{F}_t^V)$ of the form

$$\Lambda_t^{V,\beta} = \Lambda_0^{V,\beta} + \int_0^t \lambda_s^\beta \mathrm{d}s$$

for an integrable right-continuous process (λ_t^β) and a deterministic constant $\Lambda_0^{V,\beta}$. Then X^β is locally independent of X^A given X^C for $A, C \subseteq V$ if the optional projection

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

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



Another way to phrase the conclusion of the proposition is that if the optional projection $E(\lambda_t^\beta \mid \mathcal{F}_t^C)$ is indistinguishable from $E(\lambda_t^\beta \mid \mathcal{F}_t^{A\cup C})$, then $A \not\rightarrow \beta \mid C$, and it is a way of testing local independence via the computation of conditional expectations. It is a precise formulation of the innovation theorem stating how to compute compensators for one filtration via conditional expectations of compensators for a superfiltration.

F. Proofs. The following are proofs of the results from the main paper.

Proof of Proposition 3.3. Let ω be a μ-connecting walk given C and let γ be a collider on the walk such that $\gamma \in An(C) \setminus C$. Then there exists a subwalk $\bar{\omega} = \alpha_1 \ast\!\!\rightarrow \gamma \leftarrow\!\!\ast \alpha_2$, and an open (given C), directed path from γ to $\delta \in C$, π. By composing $\alpha_1 \ast\!\!\rightarrow \gamma$ with π, π^{-1}, and $\gamma \leftarrow\!\!\ast \alpha_2$ we get an open walk which is endpoint-identical to $\bar{\omega}$ and with its only collider, δ, in C, and we can substitute $\bar{\omega}$ with this new walk. Making such a substitution for every collider in $An(C)\setminus C$ on ω, we obtain a μ-connecting walk on which every collider is in C. □

Proof of Proposition 3.5. Assume that we start from α and continue along ω until some node, $\gamma \neq \beta$, is repeated. Remove the cycle from γ to γ to obtain another walk from α to β, $\bar{\omega}$. If $\gamma = \alpha$, then $\bar{\omega}$ is μ-connecting. Instead assume $\gamma \neq \alpha$. If this instance of γ is a noncollider on $\bar{\omega}$ then it must have been a noncollider in an instance on ω and thus $\gamma \notin C$. If on the other hand this instance of γ is a collider on $\bar{\omega}$ then either γ was a collider in an instance on ω or the ancestor of a collider on ω, and thus $\gamma \in An(C)$. In either case, we see that $\bar{\omega}$ is a μ-connecting walk. Repeating this argument, we can construct a μ-connecting walk where only β is potentially repeated. If there is $n > 2$ instances of β then we can remove at least $n-2$ of them as above as long as we leave an edge with a head at the final β. □

Proof of Proposition 3.6. Note first that a vertex can be a parent of itself. The result then follows from the fact that $\alpha \perp_\mu \beta \mid \mathrm{pa}(\beta)$. □

Proof of Proposition 3.9. The first statement follows from the fact that no edge without heads (i.e. $-$) is ever added. Assume for the second statement that $\mathcal{G}$ satisfies (3.3). Let $M = V \setminus O$. Assume $\alpha \leftrightarrow_{\mathcal{M}} \beta$. By definition of the latent projection, we can find an endpoint-identical walk between α and β in $\mathcal{G}$ with no colliders and such that all non-endpoint nodes are in M. Either this walk has a bidirected edge at α in which case $\alpha \leftrightarrow_{\mathcal{G}} \alpha$ by (3.3) and therefore also $\alpha \leftrightarrow_{\mathcal{M}} \alpha$. Otherwise, there is a directed edge from some node $\gamma \in M$ such that $\gamma \to_{\mathcal{G}} \alpha$. Then the walk $\alpha \leftarrow \gamma \to \alpha$ is present in $\mathcal{G}$ and therefore $\alpha \leftrightarrow_{\mathcal{M}} \alpha$ because $\mathcal{M}$ is a latent projection. □

Proof of Proposition 3.10. Assume first that α has no loops. In this case, there are no bidirected edges between α and any node, and therefore the edges that have a head at α have a tail at the previous node. Any nontrivial walk between α and α is therefore blocked by $V \setminus \{\alpha\}$. Conversely, if α has a loop, then $\alpha \ast\!\!\to \alpha$ is a μ-connecting walk given $V \setminus \{\alpha\}$. □

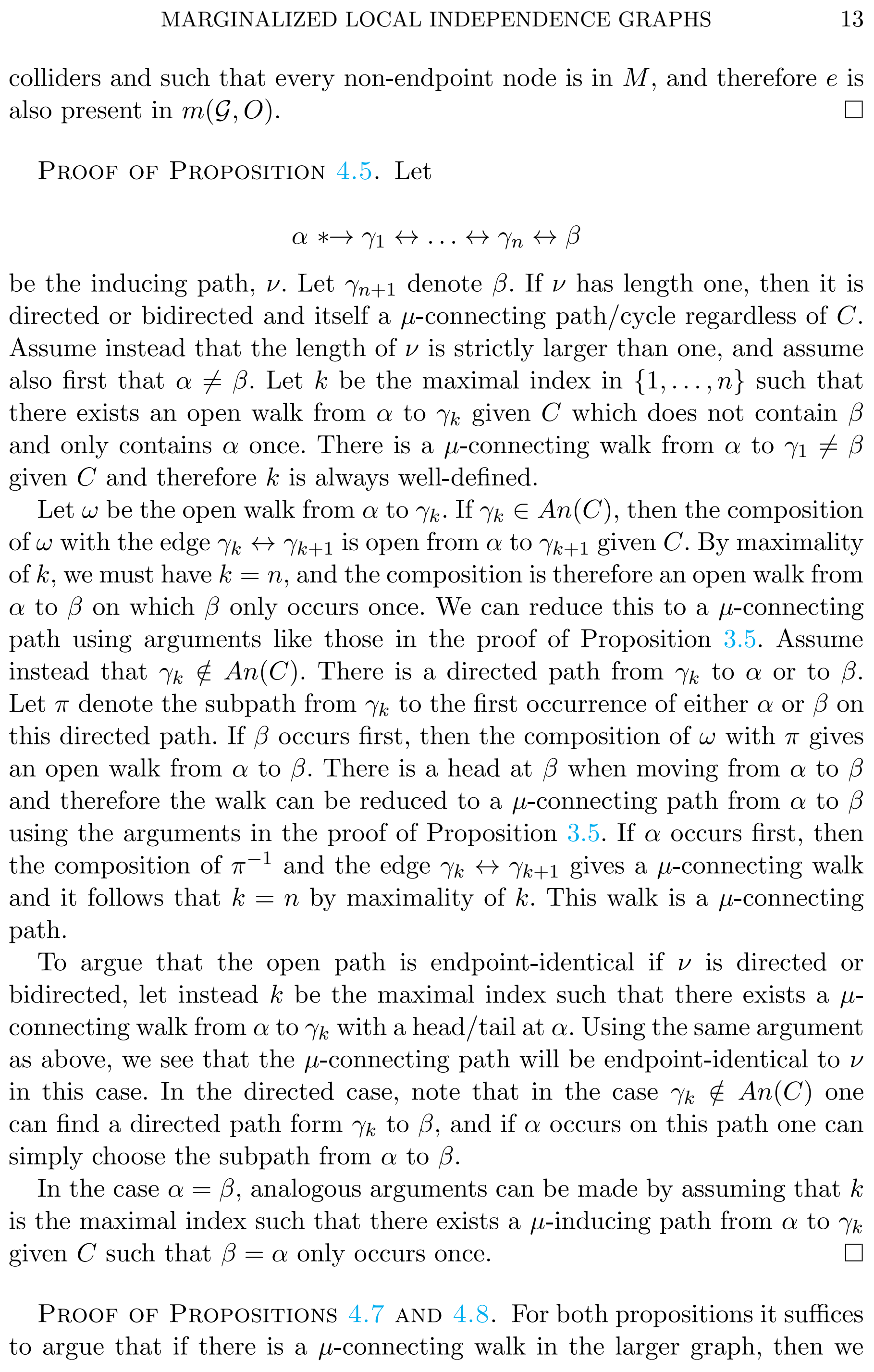
Proof of Theorem 3.12. Let $M = V \backslash O$. Let first ω be a μ-connecting walk from $\alpha \in A$ to $\beta \in B$ given C in $\mathcal{G}$. Using Proposition 3.3, we can find a μ-connecting walk from $\alpha \in A$ to $\beta \in B$ given C in $\mathcal{G}$ such that all colliders are in C. Denote this walk by $\bar{\omega}$. Every node, m, on $\bar{\omega}$ which is in M is on a subwalk of $\bar{\omega}$, $\delta_1 \sim \ldots \sim m \sim \ldots \sim \delta_2$, such that $\delta_1, \delta_2 \in O$ and all other nodes on the subwalk are in M. There are no colliders on this subwalk and therefore there is an endpoint-identical edge $\delta_1 \sim \delta_2$ in $\mathcal{M}$. Substituting all such subwalks with their corresponding endpoint-identical edges gives a μ-connecting walk in $\mathcal{M}$.

On the other hand, let ω be a μ-connecting walk from A to B given C in $\mathcal{M}$. Consider some edge in ω which is not in $\mathcal{G}$. In $\mathcal{G}$ there is an endpoint-identical walk with no colliders and no non-endpoint nodes in C. Substituting each of these edges with such an endpoint-identical walk gives a μ-connecting walk in $\mathcal{G}$ using Proposition 3.11. □

Proof of Proposition 3.13. We first note that in Algorithm 1 adding an edge will never remove any triroutes. Therefore, Algorithm 1 returns the same output regardless of the order in which the algorithm adds edges.

Let $\mathcal{M}$ denote the output of Algorithm 1 which is clearly a DMG. The graphs $\mathcal{M}$ and $m(\mathcal{G}, O)$ have the same node set, thus it suffices to show that also the edge sets are equal. Assume first $\alpha \overset{e}{\sim}_{m(\mathcal{G},O)} \beta$. Then there exist an endpoint-identical walk in $\mathcal{G}$ that contains no colliders and such that all the non-endpoint nodes are in $M = V \setminus O$, $\alpha \sim \gamma_1 \sim \ldots \sim \gamma_n \sim \gamma_{n+1} = \beta$. Let e_l be the edge between α and γ_l which is endpoint-identical to the subwalk from α to γ_l. If e_l is present in $\mathcal{M}_k$ at some point during Algorithm 1, then edge e_{l+1} will also be added before the algorithm terminates, $l = 1, \ldots, n$. We see that e_1 is in $\mathcal{G}$, and this means that e is also present in $\mathcal{M}$.

On the other hand, assume that some edge e is in $\mathcal{M}$. If e is not in $\mathcal{G}$, then we can find a noncolliding, endpoint-identical triroute in the graph $\mathcal{M}_k$ (k has the value that it takes when the algorithm terminates) such that the noncollider is in M. By repeatedly using this argument, we can from any edge, e, in $\mathcal{M}$ construct an endpoint-identical walk in $\mathcal{G}$ that contains no



colliders and such that every non-endpoint node is in M, and therefore e is also present in $m(\mathcal{G}, O)$. □

PROOF OF PROPOSITION 4.5. Let

$$\alpha \ast\!\!\rightarrow \gamma_1 \leftrightarrow \ldots \leftrightarrow \gamma_n \leftrightarrow \beta$$

be the inducing path, ν. Let γ_{n+1} denote β. If ν has length one, then it is directed or bidirected and itself a μ-connecting path/cycle regardless of C. Assume instead that the length of ν is strictly larger than one, and assume also first that $\alpha \neq \beta$. Let k be the maximal index in $\{1, \ldots, n\}$ such that there exists an open walk from α to γ_k given C which does not contain β and only contains α once. There is a μ-connecting walk from α to $\gamma_1 \neq \beta$ given C and therefore k is always well-defined.

Let ω be the open walk from α to γ_k. If $\gamma_k \in An(C)$, then the composition of ω with the edge $\gamma_k \leftrightarrow \gamma_{k+1}$ is open from α to γ_{k+1} given C. By maximality of k, we must have $k = n$, and the composition is therefore an open walk from α to β on which β only occurs once. We can reduce this to a μ-connecting path using arguments like those in the proof of Proposition 3.5. Assume instead that $\gamma_k \notin An(C)$. There is a directed path from γ_k to α or to β. Let π denote the subpath from γ_k to the first occurrence of either α or β on this directed path. If β occurs first, then the composition of ω with π gives an open walk from α to β. There is a head at β when moving from α to β and therefore the walk can be reduced to a μ-connecting path from α to β using the arguments in the proof of Proposition 3.5. If α occurs first, then the composition of π^{-1} and the edge $\gamma_k \leftrightarrow \gamma_{k+1}$ gives a μ-connecting walk and it follows that $k = n$ by maximality of k. This walk is a μ-connecting path.

To argue that the open path is endpoint-identical if ν is directed or bidirected, let instead k be the maximal index such that there exists a μ-connecting walk from α to γ_k with a head/tail at α. Using the same argument as above, we see that the μ-connecting path will be endpoint-identical to ν in this case. In the directed case, note that in the case $\gamma_k \notin An(C)$ one can find a directed path form γ_k to β, and if α occurs on this path one can simply choose the subpath from α to β.

In the case $\alpha = \beta$, analogous arguments can be made by assuming that k is the maximal index such that there exists a μ-inducing path from α to γ_k given C such that $\beta = \alpha$ only occurs once. □

PROOF OF PROPOSITIONS 4.7 AND 4.8. For both propositions it suffices to argue that if there is a μ-connecting walk in the larger graph, then we can also find a μ-connecting walk in the smaller graph. Using Proposition 4.5 we can find endpoint-identical walks that are open given $C \setminus \{\alpha\}$ and replacing $\alpha \ast\!\!\rightarrow \beta$ with such a walk will give a walk which is open given C. For Proposition 4.8 one should note that adding the edge respects the ancestry of the nodes due to transitivity. □

Proof of Proposition 4.10. Assume there is no inducing path from α to β and let ω be some walk from α to β with a head at β. Note that ω must have length at least 2.

$$\alpha = \gamma_0 \overset{e_0}{\sim} \gamma_1 \overset{e_1}{\sim} \dots \overset{e_{m-1}}{\sim} \gamma_m \overset{e_m}{\ast\!\!\rightarrow} \beta.$$

There must exist an $i \in \{0, 1, \dots, m\}$ such that γ_i is not directedly collider-connected to β along ω or such that $\gamma_i \notin An(\alpha, \beta)$. Let j be the largest such index. Note first that γ_m is always directedly collider-connected to β along ω and γ_0 is always in $An(\alpha, \beta)$. If $j \neq m$ and γ_j is not directedly collider-connected to β along ω, then γ_{j+1} is a noncollider and ω is closed in $\gamma_{j+1} \in D(\alpha, \beta)$ (note that $\alpha = \gamma_{j+1}$ is impossible as there would then be an inducing path from α to β). If $j \neq 0$ and $\gamma_j \notin An(\alpha, \beta)$ then there is some $k \in \{1, \dots, j\}$ such that γ_k is a collider and $\gamma_k \notin An(\alpha, \beta)$ and ω is therefore closed in this collider. □

Proof of Proposition 5.3. We verify that (gs1)–(gs3) hold.
(gs1) The edge $\alpha \leftrightarrow \beta$ constitutes an inducing path in both directions.
(gs2-3) Let $\gamma \in V, C \subseteq V$ such that $\beta \in C$, and assume that there is a μ-connecting walk from γ to β given C in $\mathcal{G}$. This walk has a head at β and composing the walk with $\alpha \leftrightarrow \beta$ creates an μ-connecting walk from γ to α given C. □

Proof of Lemma 5.4. Any μ-connecting walk in $\mathcal{G}$ is also present and μ-connecting in $\mathcal{G}^+$, hence $\mathcal{I}(\mathcal{G}^+) \subseteq \mathcal{I}(\mathcal{G})$.

Assume $\gamma, \delta \in V, C \subseteq V$ and assume that ρ is a μ-connecting route from γ to δ given C in $\mathcal{G}^+$. Let e denote the edge $\alpha \leftrightarrow \beta$. Using (gs1), there exist an inducing path from α to β in $\mathcal{G}$ and one from β to α. Denote these by ν_1 and ν_2. If e is not in ρ, then ρ is also in $\mathcal{G}$ and μ-connecting as the addition of the bidirected edge does not change the ancestry of $\mathcal{G}$.

If e occurs twice in ρ then it contains a subroute $\alpha \overset{e}{\leftrightarrow} \beta \overset{e}{\leftrightarrow} \alpha$ and $\alpha = \delta$ (or with the roles interchanged). Either one can find a μ-connecting subroute of ρ with no occurrences of e or $\alpha \notin C$. If $\beta \in C$, then compose the subroute of ρ from γ to the first occurrence of α (which is either trivial or can be assumed to have a tail at α) with the ν_1-induced open walk from α to β

using Proposition 4.5. This is a μ-connecting walk in $\mathcal{G}$ from γ to β and using (gs2) the result follows. If $\beta \notin C$, then the result follows from composing the subroute from γ to α with the ν_1-induced open walk from α to β and the ν_2-inducing open walk from β to α.

If e only occurs once on ρ, consider first a ρ of the form

$$\underbrace{\gamma \sim \ldots \sim \alpha}_{\rho_1} \overset{e}{\leftrightarrow} \underbrace{\beta \sim \ldots \mathrel{*\!\!\rightarrow} \delta}_{\rho_2}.$$

Assume first that $\alpha \notin C$. Let π denote the ν_1-induced open walk from α to β and note that π has a head at β. If $\gamma = \alpha$ then π composed with ρ_2 is a μ-connecting walk from γ to δ in $\mathcal{G}$. If $\gamma \neq \alpha$ we can just replace e with π, and the resulting composition of the walks ρ_1, π and ρ_2 is a μ-connecting walk from γ to δ in $\mathcal{G}$. If instead $\alpha \in C$, then $\gamma \neq \alpha$ and α is a collider on ρ, and ρ_1 thus has a head at α and is μ-connecting from γ to α given C in $\mathcal{G}$. Using (gs3) we can find a μ-connecting walk from γ to β given C in $\mathcal{G}$. Composing this with ρ_2 gives a μ-connecting walk from γ to δ given C in $\mathcal{G}$.

If ρ instead has the form

$$\gamma \sim \ldots \sim \beta \overset{e}{\leftrightarrow} \alpha \sim \ldots \mathrel{*\!\!\rightarrow} \delta,$$

a similar argument using (gs2) applies. In conclusion, $\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(\mathcal{G}^+)$. □

Proof of Proposition 5.7. We verify that (gp1)–(gp4) hold.
(gp1) $\alpha \to \beta$ constitutes an inducing path from α to β.
(gp2) Let ω be a μ-connecting walk from γ to α given C, $\alpha \notin C$. Then ω composed with $\alpha \to \beta$ is μ-connecting from γ to β given C.
(gp3) Let ω_1 be a μ-connecting walk from γ to β given C, $\alpha \notin C, \beta \in C$, and let ω_2 be a μ-connecting walk from α to δ given C. The composition of ω_1, $\alpha \to \beta$, and ω_2 is μ-connecting.
(gp4) Let ω be a μ-connecting walk from β to γ given $C \cup \{\alpha\}$, $\alpha \notin C$. If this walk is closed given C, then there exists a collider on ω, which is an ancestor of α and not in $An(C)$. Let δ be the collider on ω with this property which is the closest to γ. Then we can find a directed and open path from δ to β and composing the inverse of this with the subwalk of ω from δ to γ gives us a connecting walk. □

Proof of Lemma 5.8. As $An_{\mathcal{G}}(C) \subseteq An_{\mathcal{G}^+}(C)$ for all $C \subseteq V$, any μ-connecting path in $\mathcal{G}$ is also μ-connecting in $\mathcal{G}^+$, and it therefore follows that $\mathcal{I}(\mathcal{G}^+) \subseteq \mathcal{I}(\mathcal{G})$.

We will prove the other inclusion by considering a μ-connecting walk from γ to δ given C in $\mathcal{G}^+$ and argue that we can find another μ-connecting walk in $\mathcal{G}^+$ that fits into cases (a) or (b) below. In both cases, we will use the potential parents properties to argue that there is also a μ-connecting walk from γ to δ given C in $\mathcal{G}$. Let e denote the edge $\alpha \to \beta$.

Let ν denote the inducing path from α to β in $\mathcal{G}$ which we know to exist by (gp1) and Proposition 4.10. Say we have a μ-connecting walk in $\mathcal{G}^+$, ω, from γ to δ given C. There can be two reasons why ω is not μ-connecting in $\mathcal{G}$: 1) e is in ω, 2) there exist colliders, $c_1, \ldots, c_k$, on ω, which are in $An_{\mathcal{G}^+}(C)$ but not in $An_{\mathcal{G}}(C)$. We will in this proof call such colliders *newly closed*. If there exists a newly closed collider on ω, c_i, then there exists in $\mathcal{G}$ a directed path from c_i to α on which no node is in C, and furthermore $\alpha \notin C$. Note that this path does not contain β, and the existence of a newly closed collider implies that $\beta \in An_{\mathcal{G}}(C)$.

Using Proposition 3.5, we can find a route, ρ, in $\mathcal{G}^+$ from γ to δ, which is μ-connecting in $\mathcal{G}^+$. Assume first that e occurs at most once on ρ. If there are newly closed colliders on ρ, we will argue that we can find a μ-connecting walk in $\mathcal{G}^+$ with no newly closed colliders and such that e occurs at most once. Assume that $c_1, \ldots, c_k$ are newly closed colliders, ordered by their occurrences on the route ρ. We allow for $k = 1$, in which case $c_1 = c_k$. We will divide the argument into three cases, and we use in all three cases that a μ-connecting walk in $\mathcal{G}$ is also present in $\mathcal{G}^+$ and has no newly closed colliders nor occurrences of e. We also use that $\alpha \notin C$ when applying (gp2).

(i) e is between γ and c_1 on ρ.
Consider the subwalk of ρ from γ to the first occurrence of α. If this subwalk has a tail at α (or is trivial) then we can compose it with the inverse of the path from c_k to α and the subwalk from c_k to δ. This walk is open. If there is a head at α, then using (gp2) we can find a μ-connecting walk from γ to β in $\mathcal{G}$, compose it with e, the inverse of the path from c_k to α and the subwalk from c_k to δ. This is open as $\beta \in An_{\mathcal{G}}(C)$ and $\alpha \notin C$ whenever there exist newly closed colliders.

(ii) e is between c_k and δ on ρ.
Consider the subwalk of ρ from γ to c_1, and compose it with the directed path from c_1 to α. This is μ-connecting in $\mathcal{G}$ and using (gp2) we can find a μ-connecting walk in $\mathcal{G}$ from γ to β. Composing this walk with the subwalk of ρ from β to δ gives a μ-connecting walk from γ to δ, noting that $\beta \in An_{\mathcal{G}}(C)$.

(iii) e is between c_1 and c_k on ρ or not on ρ at all.
Composing the subwalk from γ to c_1 with the directed path from c_1 to α gives a μ-connecting walk from γ to α given C in $\mathcal{G}$, and by (gp2) we can find a μ-connecting walk from γ to β in $\mathcal{G}$, thus there are no newly closed colliders on this walk and it does not contain e. Composing it

with e, the directed path from c_k to α and the subwalk from c_k to δ gives a μ-connecting walk in $\mathcal{G}^+$.

In all cases (i), (ii), and (iii) we have argued that there exists a μ-connecting walk from γ to δ in $\mathcal{G}^+$ that contains no newly closed colliders and that contains e at most once. Denote this walk by $\tilde{\omega}$. If $\tilde{\omega}$ does not contain e at all, then we are done. Otherwise, two cases remain, depending on the orientation of e in the μ-connecting walk $\tilde{\omega}$:

(a) Assume first we have a walk of the form

$$\gamma \sim \ldots \overset{e_\alpha}{\sim} \alpha \to \beta \sim \ldots \ast\!\!\to \delta,$$

If there is a tail on e_α at α, or if $\gamma = \alpha$, then we can substitute e with the open path between α and β induced by ν and obtain an open walk. Otherwise, assume a head on e_α at α. $\tilde{\omega}$ is μ-connecting in $\mathcal{G}^+$ and therefore $\alpha \notin C$. Using (gp2), there exists a μ-connecting walk from γ to β, and composing this walk with the (potentially trivial) subwalk from β to δ gives a μ-connecting walk from γ to δ given C in $\mathcal{G}$.

(b) Consider instead a walk of the form

$$\gamma \sim \ldots \overset{e_\beta}{\sim} \beta \leftarrow \alpha \sim \ldots \ast\!\!\to \delta.$$

If there is a head on e_β at β, β is a collider. If $\beta \in C$, then (gp3) directly gives a μ-connecting walk from γ to δ given C in $\mathcal{G}$. If instead $\beta \in An_{\mathcal{G}^+}(C) \setminus C$ then we can find a directed path, π, in $\mathcal{G}^+$ from β to $\varepsilon \in C$. The edge e is not present on π and therefore we can compose the subwalk from γ to β with π, π^{-1}, and the subwalk from β to δ to obtain an open walk from γ to δ without any newly closed colliders, only one occurrence of e, and such that there is a tail at β just before the occurence of e.

We have reduced this case to walks, $\tilde{\omega}$, of the form

$$\underbrace{\gamma \sim \ldots \leftarrow \beta}_{\tilde{\omega}_1} \leftarrow \underbrace{\alpha \sim \ldots \ast\!\!\to \delta}_{\tilde{\omega}_2},$$

where $\tilde{\omega}_1$ is potentially trivial. Let $\bar{\pi}$ denote the ν-induced open path or cycle from α to β in $\mathcal{G}$. Using Proposition 3.5 there is a μ-connecting route, $\bar{\rho}$, from α to δ given C in $\mathcal{G}$. If there is a tail at α on $\bar{\rho}$ or on $\bar{\pi}$ then the composition of $\tilde{\omega}_1$, $\bar{\pi}$ and $\bar{\rho}$ is μ-connecting. Otherwise, if $\alpha \neq \beta$, the composition of $\bar{\pi}$ and $\bar{\rho}$ is a μ-connecting walk from β to δ given $C \cup \{\alpha\}$ in $\mathcal{G}$ as α does not occur as a noncollider on this composition. Using (gp4) there is also one given C. As there is a tail at β on $\tilde{\omega}$ we can compose $\tilde{\omega}_1$ with this walk to obtain an open walk from γ to δ given C in $\mathcal{G}$. If $\alpha = \beta$ the composition of $\tilde{\omega}_1$ with $\tilde{\omega}_2$ is an open walk from γ to δ given C in $\mathcal{G}$.

Assume finally that e occurs twice on ρ. In this case ρ contains a subroute $\beta \overset{e}{\leftarrow} \alpha \overset{e}{\rightarrow} \beta$ and $\beta = \delta$. In this case $\alpha \notin C$. If there are any newly closed colliders, consider the one closest to γ, c. The subroute of ρ from γ to c composed with the directed path from c to α gives a μ-connecting path and (gp2) gives the result. Else if there is a head at α on the ν-induced open walk then (gp2) again gives the result. Otherwise, compose the subroute from γ to the first β, the inverse of the ν-induced open walk, and the ν-induced open walk to obtain an open walk in $\mathcal{G}$ from γ to $\beta = \delta$. □

Proof of Theorem 5.9. Propositions 5.3 and 5.7 show that $\mathcal{N}$ is in fact a supergraph of $\mathcal{G}$, and as E^m only depends on the independence model, it also shows that $\mathcal{N}$ is a supergraph of any element in $[\mathcal{G}]$. We can sequentially add the edges that are in $\mathcal{N}$ but not in $\mathcal{G}$, and Lemmas 5.4 and 5.8 show that this is done Markov equivalently, meaning that $\mathcal{N} \in [\mathcal{G}]$. □

Lemma F.1. *Let $\alpha, \beta \in V$. If there is a directed edge, e, from α to β, and a unidirected inducing path from α to β of length at least two in $\mathcal{N}$, then there is a directed inducing path from α to β in $\mathcal{N} - e$.*

Proof of Lemma F.1. Let ν denote the unidirected inducing path and $\gamma_1, \ldots, \gamma_n$ the non-endpoint nodes of ν. Then $\gamma_i \in An_{\mathcal{N}}(\{\alpha, \beta\})$ and also $\gamma_i \in An_{\mathcal{N}}(\beta)$ due to the directed edge from α to β. It follows that either $\gamma_i \in An_{\mathcal{N}}(\alpha)$ or $\gamma_i \in An_{(\mathcal{N}-e)}(\beta)$. If $\gamma_i \in An_{\mathcal{N}}(\alpha)$, let e_i denote the directed edge from γ_i to β, and let $\mathcal{N}^+ = (V, F \cup \{e_i\})$. We will argue that $\mathcal{N} = \mathcal{N}^+$ using the maximality of $\mathcal{N}$. Note first that the edge does not change the ancestry of the graph in the sense that $An_{\mathcal{N}}(\gamma) = An_{\mathcal{N}^+}(\gamma)$ for all $\gamma \in V$. Note also that there is a bidirected inducing path between γ_i and β in $\mathcal{N}$, and therefore $\gamma_i \leftrightarrow_{\mathcal{N}} \beta$. Assume that e_i is in a μ-connecting path in $\mathcal{N}^+$. There is a directed path from γ_i to α in $\mathcal{N}$ and therefore e_i can either be substituted with $\gamma_i \rightarrow \alpha_i \rightarrow \ldots \rightarrow \alpha_k \rightarrow \alpha \rightarrow \beta$ (if $\alpha_1, \ldots, \alpha_k, \alpha \notin C$), or with $\gamma_i \leftrightarrow \beta$ (otherwise), and we see that $\mathcal{I}(\mathcal{N}) = \mathcal{I}(\mathcal{N}^+)$. By maximality of $\mathcal{N}$ we have that $\mathcal{N} = \mathcal{N}^+$ which implies that $e_i \in F$. Thus $\gamma_i \in An_{(\mathcal{N}-e)}(\beta)$. This shows that ν is also a directed inducing path in $\mathcal{N} - e$. □

Lemma F.2. *Let edges $\alpha \rightarrow \beta$, $\beta \rightarrow \alpha$ and $\alpha \leftrightarrow \beta$ be denoted by e_1, e_2, e_3, respectively. If $e_1, e_3 \in F$, then $\mathcal{N} - e_1 \in [\mathcal{N}]$. If $e_1, e_2, e_3 \in F$, then $\mathcal{N} - e_3 \in [\mathcal{N}]$.*

Proof of Lemma F.2. Note that if edges $\gamma \ast\!\!\to \alpha$, $\alpha \leftrightarrow \beta$, and $\alpha \to \beta$ are present in a maximal DMG, then so is $\gamma \ast\!\!\to \beta$ by Propositions 4.7 and 4.8. Assume $e_1, e_3 \in E$. Using the above observation, note that every vertex that is a parent of α in $\mathcal{N}$ is also a parent of β, thus $An_{\mathcal{N}}(\delta) \setminus \{\alpha\} = An_{(\mathcal{N}-e_1)}(\delta) \setminus \{\alpha\}$ for all $\delta \in V$. Consider a μ-connecting walk, ω, in $\mathcal{N}$ given C. Any collider different from α on this walk is in $An_{(\mathcal{N}-e_1)}(C)$. If $\alpha \notin An_{(\mathcal{N}-e_1)}(C)$ is a collider, then we can substitute the subwalk $\gamma_1 \ast\!\!\to \alpha \leftarrow\!\!\ast \gamma_2$ with $\gamma_1 \ast\!\!\to \beta \leftarrow\!\!\ast \gamma_2$. If e_1 is the first edge on ω and α the first node, then just substitute e_1 with e_3. Else, we need to consider two cases: in the first case there is a subwalk $\gamma \ast\!\!\to \alpha \to \beta$ (or $\beta \leftarrow \alpha \leftarrow\!\!\ast \gamma$) and therefore an edge $\gamma \ast\!\!\to \beta$ in $\mathcal{N} - e_1$ if $\gamma \neq \alpha$. If $\gamma = \alpha$, we can simply remove the loop, replacing e_1 with e_3 if γ was the final node on ω. In the second case, there is a subwalk $\gamma \leftarrow \alpha \to \beta$ (or $\beta \leftarrow \alpha \to \gamma$), and we can substitute e_1 with e_3 if $\beta \neq \gamma$. If $\beta = \gamma$, then we can substitute $\beta \leftarrow \alpha \to \beta$ with $\beta \leftrightarrow \beta$.

The proof of the other statement is similar. □

Proof of Proposition 5.11. One implication is immediate by contraposition: if $\alpha \notin u(\beta, \mathcal{I}(\mathcal{N} - e))$, then $\mathcal{N} - e \notin [\mathcal{N}]$.

Assume $\alpha \in u(\beta, \mathcal{I}(\mathcal{N} - e))$. There exists an inducing path, ν, from α to β in $\mathcal{N} - e$. If ν is directed, then the conclusion follows from Proposition 4.8. If ν is unidirected and of length one, then it is also directed. If it is unidirected and has length at least two, it follows from Lemma F.1 that there also exists a directed inducing path in $\mathcal{N} - e$. Proposition 4.8 finishes the argument. Assume that ν is bidirected. Then $\alpha \leftrightarrow_{\mathcal{N}} \beta$ due to maximality and Proposition 4.7. Lemma F.2 gives the result. □

Proof of Proposition 5.12. One implication follows by contraposition. Assume instead that $\alpha \in u(\beta, \mathcal{I}(\mathcal{N} - e))$ and $\beta \in u(\alpha, \mathcal{I}(\mathcal{N} - e))$. Then there is an inducing path from α to β and one from β to α in $\mathcal{N} - e$. Denote these by ν_1 and ν_2. If one of them is bidirected, then the conclusion follows. Assume instead that none of them are bidirected and assume first that both are a single edge. The conclusion then follows using Lemma F.2.

Assume now that ν_1 or ν_2 is an inducing path of length at least 2. Say that $\beta \to \gamma_1 \leftrightarrow \ldots \leftrightarrow \gamma_m \leftrightarrow \alpha$ is an inducing path. If ν_1 is the inducing path $\alpha \to_{\mathcal{N}} \beta$ of length one, then there is also a bidirected inducing path between γ_1 and β in $\mathcal{N}$, and there will also be a bidirected inducing path in $\mathcal{N} - e$ between α and β. If instead ν_1 is the inducing path $\alpha \to \phi_1 \leftrightarrow \ldots \leftrightarrow \phi_k \leftrightarrow \beta$ then $\gamma_1 \leftrightarrow_{\mathcal{N}} \phi_1$. In this case $\alpha \leftrightarrow \gamma_m \ldots \gamma_1 \leftrightarrow \phi_1 \ldots \phi_k \leftrightarrow \beta$ can be trimmed down to a bidirected inducing path in $\mathcal{N} - e$. □

REFERENCES

[1] Danks, D. and Plis, S. (2013). Learning causal structure from undersampled time series. In *JMLR: Workshop and Conference Proceedings (NIPS Workshop on Causality)*.

[2] Didelez, V. (2000). Graphical models for event history analysis based on local independence, PhD thesis, Universität Dortmund.

[3] Didelez, V. (2008). Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society, Series B* **70** 245-264.

[4] Eichler, M. and Didelez, V. (2010). On Granger causality and the effect of interventions in time series. *Lifetime Data Analysis* **16** 3-32.

[5] Hyttinen, A., Plis, S., Järvisalo, M., Eberhardt, F. and Danks, D. (2016). Causal discovery from subsampled time series data by constraint optimization. In *Proceedings of the Eighth International Conference on Probabilistic Graphical Models* **52** 216-227.

[6] Lauritzen, S. (1996). *Graphical models.* Oxford: Clarendon.

[7] Lauritzen, S. and Sadeghi, K. (2018). Unifying Markov properties for graphical models. *Annals of Statistics* **46** 2251-2278.

[8] Meek, C. (1995). Strong completeness and faithfulness in Bayesian networks. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI1995).*

[9] Meek, C. (2014). Toward learning graphical and causal process models. In *Proceedings of the UAI 2014 Workshop Causal Inference: Learning and Prediction.*

[10] Richardson, T. (2003). Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics* **30** 145-157.

[11] Richardson, T. and Spirtes, P. (2002). Ancestral graph Markov models. *The Annals of Statistics* **30** 962-1030.

[12] Rogers, L. C. G. and Williams, D. (2000). *Diffusions, Markov processes, and martingales. Cambridge Mathematical Library* **2**. Cambridge University Press, Cambridge Itô calculus, Reprint of the second (1994) edition.

[13] Sokol, A. and Hansen, N. R. (2014). Causal interpretation of stochastic differential equations. *Electronic Journal of Probability* **19** 1-24.

Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
2100 Copenhagen
Denmark
E-mail: swengel@math.ku.dk

Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
2100 Copenhagen
Denmark
E-mail: niels.r.hansen@math.ku.dk

Algorithms for DMGs

In this section, we discuss algorithms to solve problems related to directed mixed graphs. Several of these problems are related in the sense that a solution to one will help us find solutions to others. If we can find maximal Markov equivalent graphs of $\mathcal{G}_1$ and $\mathcal{G}_2$, respectively, we can also decide if $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent by simply checking if their maximal Markov equivalent graphs are equal. If we can learn a maximal Markov equivalent graph from an independence oracle, then we can also find a maximal Markov equivalent graph of $\mathcal{G}$ by simply querying $\mathcal{I}(\mathcal{G})$. We will see that these problems are computationally hard and that there are most likely no polynomial-time algorithms for their solution. This means that we should look either for approximation algorithms or for algorithms that do well on average, even though they may do poor in the worst cases.

The input of the algorithms are graphs, DMGs specifically, and we may think of them as encoded by pairs of adjacency matrices, thus their input size is a polynomial in the number of nodes. Therefore, we can think of the input size as simply the number of nodes in the graphs (Sipser, 2013).

Deciding μ-separation Proposition D.4 in the supplementary material of Paper **A** shows how to decide μ-separation using an augmented graph. One can find such a graph in polynomial time and this way one can show that deciding μ-separation can be done in polynomial time.

Marking a DMG From the results on Markov equivalence in DMGs in Paper **A**, we know that every equivalence class has a unique, maximal member and that every other member is a subgraph of this maximal member. Therefore, one can represent the Markov equivalence class by drawing the maximal graph and simply showing for each edge if it is in every graph in the equivalence class, or if there exists some graph in which the edge is not present. We call this new graphical object a *directed mixed equivalence graph* (DMEG), and we say that we *mark* a DMG to obtain a DMEG. As noted in Paper **A**, it is straightforward to find an algorithm for marking DMGs such that the number of required separation tests scales linearly in the number of edges. Combining this with the above observation leads to a polynomial-time algorithm for marking a DMG.

Approximate maximalization In the next section, we will argue that given a DMG, $\mathcal{G}$, it is computationally hard to find the maximal Markov equivalent DMG, $\mathcal{N}$. For this reason, we describe an approximate algorithm which is only guaranteed to return a DMG $\mathcal{G}^+$ such that $\mathcal{G} \subseteq \mathcal{G}^+ \subseteq \mathcal{N}$. We say that two nodes are *connected* if there exists a walk between them. We let $\mathcal{G}^{\alpha,d}$ denote the graph on nodes V obtained from $\mathcal{G}$ by removing all directed edges that are not out of α and all bidirected edges that are incident with α. That is, for all $\gamma, \delta \in V$,

$$\begin{aligned} \gamma \rightarrow_{\mathcal{G}^{\alpha,d}} \delta \quad & \text{if and only if } \gamma \rightarrow_{\mathcal{G}} \delta \text{ and } \gamma = \alpha, \\ \gamma \leftrightarrow_{\mathcal{G}^{\alpha,d}} \delta \quad & \text{if and only if } \gamma \leftrightarrow_{\mathcal{G}} \delta, \gamma \neq \alpha \text{ and } \delta \neq \alpha. \end{aligned}$$

Proposition 3.4. Let $\mathcal{G} = (V, E)$ be a DMG, and let $\alpha, \beta \in V$ such that $\alpha \neq \beta$. There is a directed inducing path from α to β in $\mathcal{G}$ if and only if $\alpha \in \text{an}(\beta)$ and α and β are connected in $(\mathcal{G}^{\alpha,d})_{\text{an}_{\mathcal{G}}(\beta)}$.

Proof. This follows directly from the definition of a directed inducing path. □

We define $\mathcal{G}^b$ to be the graph obtained from $\mathcal{G}$ after removing every directed edge.

Proposition 3.5. Let $\mathcal{G} = (V, E)$ be a DMG, and let $\alpha, \beta \in V$ such that $\alpha \neq \beta$. There is a bidirected inducing path from α to β in $\mathcal{G}$ if and only if α and β are connected in $(\mathcal{G}^b)_{\text{an}_{\mathcal{G}}(\alpha,\beta)}$.

Proof. This also follows directly, this time by using the definition of a bidirected inducing path. □

These propositions use the ancestral relations of the DMG. We will see in Chapter 4 that we can find these ancestral relations in polynomial time, e.g., using the so-called *condensation* of the directed part of the DMG. From the above propositions, it follows that we can decide in polynomial time if there is a directed or bidirected inducing path between a pair of nodes. By simply looping through all (ordered) pairs of nodes, we can add directed and bidirected edges Markov equivalently whenever there is an inducing path of the appropriate type. This gives a polynomial-time algorithm for approximate maximalization of a DMG.

Hardness of DMGs

Several inference problems in graphical models are known to be computationally hard, often NP-hard (Meek, 2001; Chickering et al., 2004; Chandrasekaran et al., 2008; Koller and Friedman, 2009). Deciding morality of undirected graphs is also known to be NP-hard (Verma and Pearl, 1993). On the other hand, constraint-based learning of equivalence classes of sparse maximal ancestral graphs can be done with a polynomial number of tests (Claassen et al., 2013). Deciding Markov equivalence of DAGs (under d-separation) can be done in polynomial time, and the same is the case for maximal ancestral graphs (m-separation) (Ali et al., 2009). This is also true in DGs (without loops) under d-separation (Richardson, 1997).

If we consider DGs under μ-separation, we know from Example **A**.4.2 that two DGs are Markov equivalent if and only if they are equal and therefore deciding Markov equivalence can also be done in polynomial time. In the case of DMGs (under μ-separation), the potential sibling and potential parent criteria allow us to find maximal DMGs and we can decide Markov equivalence of two DMGs by simply comparing their maximal Markov equivalent graphs. However, there are too many conditions in those criteria for this approach to give a polynomial-time algorithm. In this section, we argue that such an algorithm is unlikely to exist by showing that the problem of deciding Markov equivalence for DMGs is coNP-complete. The hardness of this problem is purely graphical, not based on inference, and the nature of the result is therefore different from most of the hardness results mentioned above. We are in trouble even before looking at data as the DMGs themselves give rise to computational hardship.

Theorem 3.6. It is coNP-complete to decide Markov equivalence of two DMGs.

Deciding Markov equivalence of DMGs is linked to finding maximal DMGs: if we can find maximal DMGs then we can also decide Markov equivalence of $\mathcal{G}_1$ and $\mathcal{G}_2$ by simply comparing the maximal Markov equivalent graphs of $\mathcal{G}_1$ and $\mathcal{G}_2$. This is done in polynomial time, and therefore finding maximal DMGs is (Turing) coNP-hard. If we assume that we have a constraint-based algorithm for learning maximal DMGs from an independence oracle, then we could maximalize a DMG, $\mathcal{G}$, by querying $\mathcal{I}(\mathcal{G})$, and this shows that learning maximal DMGs is also (Turing) coNP-hard (since each query is done in polynomial time).

Proof. If two graphs are not Markov equivalent, then there is a (polynomially sized) certificate consisting of sets A, B, and C, such that B is μ-separated from A given C in one graph, but not in the other. Deciding μ-separation can be done in polynomial time, and this shows that the problem is in **coNP**.

We do a reduction from 3DNF tautology to argue that the problem is coNP-hard. Assume we have a Boolean formula in 3DNF form, on variables $x_1, \dots, x_n$,

$$(z_1^1 \wedge z_2^1 \wedge z_3^1) \vee \dots \vee (z_1^N \wedge z_2^N \wedge z_3^N),$$

such that z_i^j is a *literal*, i.e., a variable x_l or its negation $\neg x_l$. We say that x_l is a *positive* literal, and that $\neg x_l$ is a *negative* literal. We construct a graph that has (among other nodes) two nodes for each literal, one for each variable, and one for the negation of each variable. We let ζ_i^j and $\tilde{\zeta}_i^j$ be nodes corresponding to the literal z_i^j, χ_l corresponds to the variable x_l, and λ_l corresponds to its negation, $\neg x_l$. Now define

$$V^- = \{\zeta_i^j, \tilde{\zeta}_i^j\} \cup \{\chi_l, \lambda_l\},$$

and

$$V = \{\alpha, \beta, \delta\} \cup V^- \cup \{\gamma^v, \varepsilon^v\}_{v \in V^-}.$$

We construct a graph on nodes V, containing (among others) the edges in Figure 3.2. We also include all directed loops. For each $v \in V^-$, we also add the edges in Figure 3.3. Finally, for each variable, x_l, we take all positive literals that correspond to a variable x_l and join them (including χ_l) by a directed cycle (any cycle works), and all the negative literals and join them (including λ_l) by a directed cycle. This graph is denoted by $\mathcal{G}$. We also construct $\mathcal{G}^+$ from $\mathcal{G}$ by simply adding $\delta \leftrightarrow \beta$, and we denote this edge by e. The reduction from the 3DNF formula to these two graphs is done in polynomial time in the number of terms. We say that $\{\zeta_i^j\}_{i=1,2,3} \subseteq V$ is a ζ^j*-component* of the graph and that $\{\tilde{\zeta}_i^j\}_{i=1,2,3} \subseteq V$ is a $\tilde{\zeta}^j$*-component* (similarly for terms that have fewer than three literals). We say that $\{\chi_l, \lambda_l\}_{l=1,\dots,n}$ is the $\chi - \lambda$*-component* of the graph.

We now argue that the formula is a tautology if and only if $\mathcal{G}$ and $\mathcal{G}^+$ are Markov equivalent. Assume first that the formula is a tautology. In this part of the proof we give two arguments that both show that if the formula is a tautology, then $\mathcal{G}$ and $\mathcal{G}^+$ are Markov equivalent. One is a direct proof which follows below, and the other uses the potential siblings criteria of paper **A** and is stated as Lemma 3.7 below. For the direct proof, consider any μ-connecting route in $\mathcal{G}^+$, from ϕ_1 to ϕ_2 given C,

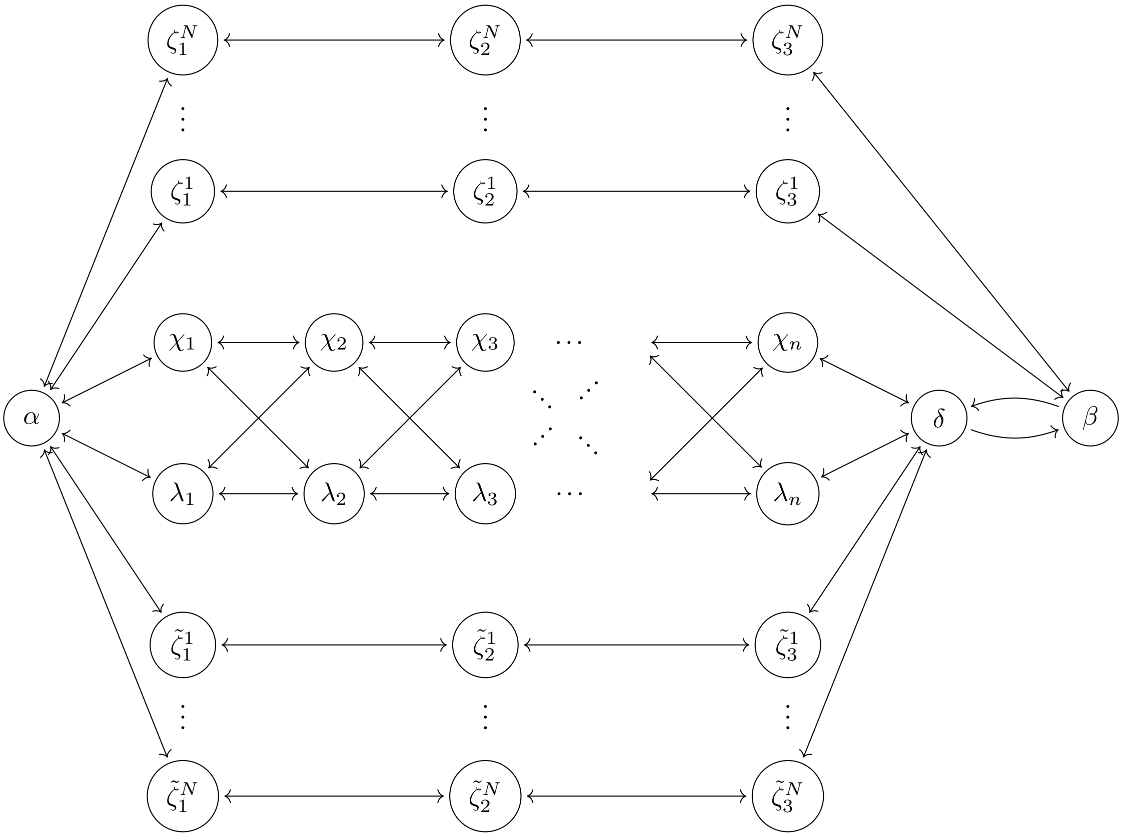


Figure 3.2: A subgraph of $\mathcal{G}$ in the proof of Theorem 3.6.

$$\phi_1 \sim \ldots \sim \phi_2.$$

Seeing that all directed loops are included at all nodes, it suffices to consider routes such that e occurs at most once. Let ω denote such a route. If e is on this route, we split into cases based on ϕ_1. If $\phi_1 \neq \alpha, \beta, \delta$, then there exists an inducing path from ϕ_1 to β and one to δ. Consider the route,

$$\phi_1 \sim \ldots \sim \rho_1 \overset{e}{\leftrightarrow} \rho_2 \sim \phi_2.$$

If $\rho_2 = \delta$, then we can take the open walk from ϕ_1 to δ in $\mathcal{G}$ (its existence follows from the existence of the inducing path) and compose it with the subwalk from the occurrence of e to ϕ_2 and this gives a μ-connecting walk in $\mathcal{G}$. The analogous argument holds if $\rho_2 = \beta$. If instead $\phi_1 = \delta$ or $\phi_1 = \beta$, we can find a subroute of ω which is connecting in $\mathcal{G}$ or compose a subroute of ω with $\delta \to \beta$, $\beta \to \delta$, $\delta \to \delta$, or $\beta \to \beta$. If instead $\phi_1 = \alpha$, we divide into cases based on which node, θ, is found before e on the route: $\theta \sim \rho_1 \overset{e}{\leftrightarrow} \rho_2$ where $\{\rho_1, \rho_2\} = \{\beta, \delta\}$. Note that if $\theta \neq \beta, \delta$, then the value of θ identifies if $\rho = \beta$ or $\rho_1 = \delta$ as no node except β and δ are adjacent to both β and δ. If $\theta = \varepsilon^v$ for some $v \in V^-$ (analogously, if $\theta = \gamma^v$ for some v), note that either

$$v \sim \varepsilon^v \leftrightarrow \delta \overset{e}{\leftrightarrow} \beta$$

or

$$\gamma^v \sim \varepsilon^v \leftrightarrow \delta \overset{e}{\leftrightarrow} \beta$$

In the first case, we can substitute this for

$$v \sim \gamma^v \leftrightarrow \beta \text{ or } v \sim \varepsilon^v \sim \gamma^v \leftrightarrow \beta$$

and in the second case for

$$\gamma^v \leftrightarrow \beta$$

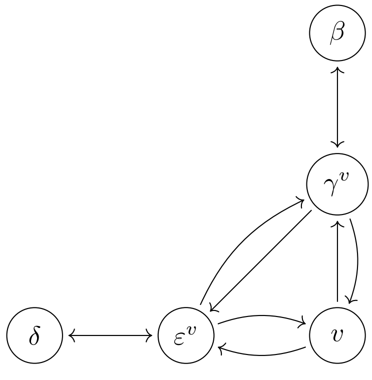


Figure 3.3: A subgraph of $\mathcal{G}$ in the proof of Theorem 3.6.

to obtain a μ-connecting walk in $\mathcal{G}$. We can argue analogously if $\theta = \gamma^v$. If $\theta = \zeta_3^j$ and the walk exits the ζ^j-component (before reaching α), then there is a tail at a node from which there is an inducing path to δ, and we can find a μ-connecting walk. If it travels along the ζ^j component to reach α, then the correspoding $\tilde{\zeta}^j$ component is open. Analogously, if $\theta = \tilde{\zeta}_3^j$. Assume $\theta = \chi_n$ or $\theta = \lambda_n$. If the route exits the $\chi - \lambda$ component (before reaching α), then there is a tail at a node from which there is an inducing path to β. If $\theta = \beta$, then $\phi_2 = \beta$ as well and there is either a μ-connecting subroute, or we can compose a subroute with the loop $\beta \to \beta$ to obtain one. Similarly, if $\theta = \delta$. Finally, if there is a walk completely within the $\chi - \lambda$ component, then we can use the fact that the formula is a tautology to see that we can find an open walk from α to β. Consider the following assignment of truth values,

$$x_l = 1 \text{ if and only if } \chi_l \text{ is on } \omega.$$

The formula is a tautology, and this means that there is a term which evaluates to true under this assignment, say, term j,

$$z_1^j \wedge z_2^j \wedge z_3^j.$$

If z_i^j is a positive literal, then the corresponding variable equals one under the assignment and $\chi_l \in \mathrm{an}(C)$. If z_i^j is a negative literal, then the corresponding variable was assigned zero, and therefore χ_l is not on ω and we must have that λ_l is on ω and therefore in $\mathrm{an}(C)$. We therefore see that there is a μ-connecting walk between α and β through the ζ^j-component of the graph.

Assume instead that the formula is not a tautology, and assume that H is some assignment of the variables $x_1, \ldots, x_n$ such that the formula is false. Define

$$C^- = \{\chi_l : x_l = 1 \text{ in } H\} \cup \{\lambda_l : x_l = 0 \text{ in } H\}$$

and let $C = \mathrm{an}(C^-) \cup \{\beta, \delta\}$. We see that β is not μ-separated from α by C in $\mathcal{G}^+$. We will now are argue that the opposite is the case in $\mathcal{G}$. It is clear that a connecting walk cannot go through δ. It cannot also not be contained in a ζ- or $\tilde{\zeta}$-component because of the choice of C and the fact that the assignment evaluates to false. It also cannot cross between any pair of components as every cyclic component is either fully in C or not at all. Finally, it cannot pass through γ^v and ε^v for any v, for the same reason. □

Lemma 3.7. Nodes β and δ are potential siblings (Definition **A**.5.1) in the DMG $\mathcal{G}$ as defined in the proof of Theorem 3.6 if the 3DNF formula is a tautology.

Proof. Some of the terminology and the graphs are introduced in the proof of Theorem 3.6. We show conditions (gs1), (gs2), and (gs3) of Proposition **A**.5.2 which proves that α and β are potential siblings in $\mathcal{G}$. (gs1) follows directly as $\beta \to_{\mathcal{G}} \delta$ and $\delta \to_{\mathcal{G}} \beta$. Assume there is μ-connecting walk from $\gamma \in V$ to δ. If $\gamma \neq \alpha$ then it is clear that there is also a μ-connecting walk from γ to β. If a connecting walk from α to δ visits β, then either there is a head at β or $\beta \notin C$, so assume that the walk does not visit β. If it traverses an entire $\tilde{\zeta}^j$-component then the corresponding ζ^j-component is open. If it stays within the $\chi - \lambda$ component, then the result follows from the fact that the formula is a tautology.

Otherwise, there must at some point exist a node on the walk, ε, such that there is a tail at ε, $\varepsilon \notin C$. There is an inducing path from ε to β and we compose the subwalk from α to ε with the open path (existence follows from the inducing path) from ε to β. This is μ-connecting walk: either there is a tail at ε on the subwalk from α to ε or $\varepsilon \in \mathrm{an}(C)$ (otherwise the μ-connecting walk from α to δ would not be able to leave the cyclic component).

If there is a μ-connecting walk from γ to β given C, then similar arguments give the result. This shows that (gs1), (gs2), and (gs3) all hold and that δ and β are potential siblings. □

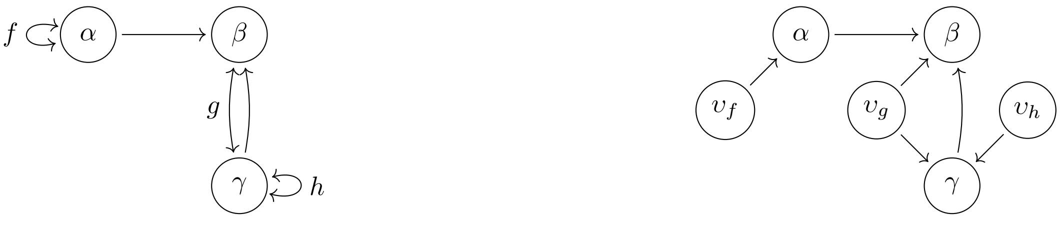


Figure 3.4: Left: A DMG on nodes $\{\alpha, \beta, \gamma\}$. The bidirected edges have been named for ease of reference. This DMG is not reachable as β is incident with a bidirected edge, yet there is no bidirected loop at β. Right: Canonical DG of the DMG. For each bidirected edge, e, a node, v_e, has been added. The latent projection of this DG onto $\{\alpha, \beta, \gamma\}$ does not equal the original DMG.

Subclasses of DMGs

In this section, we study two proper subclasses of directed mixed graphs. First, we describe the class of DMGs that can be obtained as latent projections of DGs. Second, we look into a proper subclass of DMGs which can represent any independence model of a DMG.

Reachable DMGs

We have described the class of DMGs and a marginalization algorithm. The DMGs are (graphical) marginals of DGs and it is natural to ask if any DMG can be obtained as the marginalization of a DG. The answer is in the negative and in this section we will describe the class of DMGs that can be obtained as a marginalization of a DG on a larger set of nodes. This is mentioned in Paper **A** (see **A**.(3.3)), but no argument is included, and we elaborate instead in this section. We say that a DMG $\mathcal{G} = (V, E)$ is *reachable* if it holds for all $\alpha, \beta \in V$ that $\alpha \leftrightarrow_{\mathcal{G}} \beta$ implies $\alpha \leftrightarrow_{\mathcal{G}} \alpha$. When f is an edge in $\mathcal{G}$, we let $\mathrm{e}_{\mathcal{G}}(f)$ denote the set of nodes that f is between.

Definition 3.8 (Canonical DG)**.** Let $\mathcal{G} = (V, E)$ be a DMG, and let $E_b \subset E$ be the set of bidirected edges of $\mathcal{G}$. Define $U = \{v_f : f \in E_b\}$ such that $U \cap V = \emptyset$. Its *canonical DG* is the graph $\mathcal{D}(\mathcal{G}) = (V \cup U, F)$ such that

- for all $\alpha, \beta \in V$, $\alpha \rightarrow_{\mathcal{D}(\mathcal{G})} \beta$ if and only if $\alpha \rightarrow_{\mathcal{G}} \beta$,
- for all $v_f \in U, \beta \in V$, $v_f \rightarrow_{\mathcal{D}(\mathcal{G})} \beta$ if and only if $\beta \in \mathrm{e}_{\mathcal{G}}(f)$.

Note that if $v_f \in U$, then f is a bidirected edge in $\mathcal{G}$ and $\mathrm{e}_{\mathcal{G}}(f)$ is well-defined. We can now show that a DMG can be obtained as the latent projection of a DG if and only if it is reachable.

Proposition 3.9. Let $\mathcal{G} = (O, E)$ be a DMG. There exists a DG, $\mathcal{D} = (V, E_D)$, such that $O \subseteq V$ and $m(\mathcal{D}, O) = \mathcal{G}$ if and only if $\mathcal{G}$ is reachable.

Proof. If $\mathcal{G}$ is not reachable, then there exists $\alpha, \beta \in V$ such that $\alpha \leftrightarrow_{\mathcal{G}} \beta$ and $\alpha \not\leftrightarrow_{\mathcal{G}} \alpha$. If $\mathcal{G}$ is a latent projection of a DG, $\mathcal{D}$, then there exists a walk $\alpha \leftarrow \gamma \ldots \rightarrow \beta$ and $\gamma \notin O$. In that case, $\mathcal{D}$ also contains the walk $\alpha \leftarrow \gamma \rightarrow \alpha$, and $\mathcal{G}$ contains the edge $\alpha \leftrightarrow \alpha$ which is a contradiction.

If $\mathcal{G}$ is reachable, we will argue that it is the latent projection of its canonical DG (Definition 3.8), i.e. we wish to argue that $\mathcal{G} = m(\mathcal{D}(\mathcal{G}), O)$. We denote $m(\mathcal{D}(\mathcal{G}), O)$ by $\mathcal{M}$. The two graphs have the same node set, so it suffices to argue that the edge sets are also identical. Assume first that e is in $\mathcal{G}$, and between $\alpha \in O$ and $\beta \in O$. If e is directed, then it is also in $\mathcal{D}(\mathcal{G})$, and therefore also in $\mathcal{M}$. If e is bidirected, then there is node $v_e \in U$ in the canonical DG which is a parent of α and β (these may be equal), and therefore $\alpha \leftrightarrow_{\mathcal{M}} \beta$. On the other hand, assume that e is an edge in $\mathcal{M}$, again between $\alpha \in O$ and $\beta \in O$. If it is directed, then it is also in the canonical DG, and therefore in $\mathcal{G}$. If it is bidirected, then there exists a node $v_f \in U$ such that v_f is a parent of α and β in the canonical DG. If $\alpha \neq \beta$, then $\alpha \leftrightarrow_{\mathcal{G}} \beta$. If $\alpha = \beta$, then α is adjacent with a bidirected edge in $\mathcal{G}$, and using that $\mathcal{G}$ is reachable, it follows that $\alpha \leftrightarrow_{\mathcal{G}} \alpha$. □

We can interpret the class of reachable DMGs as graphs representing systems with the property that if a process depends on an unobserved process, then it must also depend on itself.

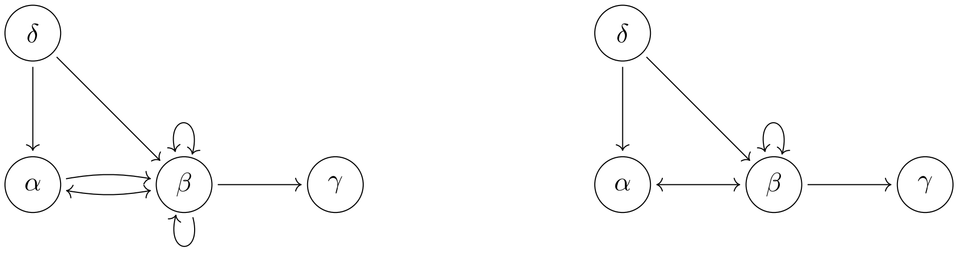


Figure 3.5: Illustration of Example 3.14. Left: maximal DMG, $\mathcal{G}$. Right: its confounding projection, $\mathcal{P}(\mathcal{G})$.

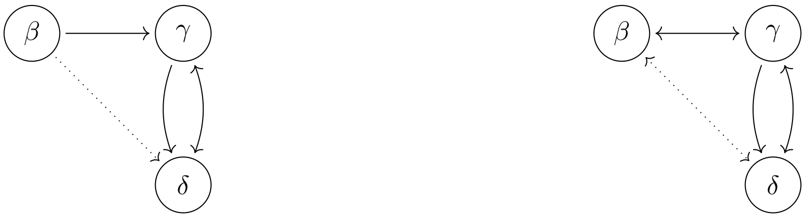


Figure 3.6: Consider a DMG such that $\beta \ast\!\!\rightarrow \gamma \leftrightarrow \delta$. If $\gamma \rightarrow \delta$ (or if just $\gamma \in \mathrm{an}(\delta)$) and the DMG is maximal, then there is an endpoint-identical edge between β and δ, $\beta \ast\!\!\rightarrow \delta$ (Propositions **A**.4.7 and **A**.4.8). The proof of Theorem 3.15 uses this observation.

Reduced DMGs

The purpose of this section is to find a class of graphs which is a strict subclass of the DMGs, though can express the μ-separation model of any DMG. We will see that a class using at most two edges between any pair of nodes can represent any Markov equivalence class.

Definition 3.10 (Confounded directed edges)**.** Let $\mathcal{G} = (V, E)$ be a DMG, and assume $\alpha \overset{e}{\rightarrow}_{\mathcal{G}} \beta$, $\alpha, \beta \in V$. If $\alpha \leftrightarrow_{\mathcal{G}} \beta$, then we say that e is *confounded*, and otherwise we say that e is *unconfounded*.

Definition 3.11 (Reduced DMG)**.** We say that a DMG is *reduced* if it has no confounded directed edges.

Definition 3.12 (Confounding projection)**.** Let $\mathcal{G} = (V, E)$ be a DMG. The *confounding projection* of $\mathcal{G}$ is the DMG $\mathcal{P}(\mathcal{G}) = (V, F)$ obtained from $\mathcal{G}$ by removing all confounded directed edges.

It follows immediately from the definition of the confounding projection that the output is a reduced DMG and that the reduced DMGs are a proper subclass of the DMGs. We call this the *confounding* projection as it retains the confounding and ignores parts of the ancestry of the original DMG.

Proposition 3.13. Let $\mathcal{G}$ be a DMG. It holds that $\alpha \in \mathrm{an}_{\mathcal{P}(\mathcal{G})}(\beta)$ if and only if there exists a directed path from α to β in $\mathcal{G}$ on which every edge is unconfounded.

Proof. If $\alpha \in \mathrm{an}_{\mathcal{P}(\mathcal{G})}(\beta)$, then there exists a directed path from α to β in $\mathcal{P}(\mathcal{G})$, and therefore also in $\mathcal{G}$. All edges on this path must be unconfounded. On the other hand, if there exists a directed path in $\mathcal{G}$ consisting of unconfounded edges, then this path exists in $\mathcal{P}(\mathcal{G})$ as well. □

Example 3.14. As an example of the above proposition consider the DMG, $\mathcal{G}$, in Figure 3.5 (left). α is an ancestor of γ in $\mathcal{G}$, however there is no directed path from α to γ such that every edge is unconfounded and therefore α is not an ancestor of γ in $\mathcal{P}(\mathcal{G})$ (Figure 3.5, right). On the other hand, δ is also an ancestor of γ in $\mathcal{G}$, and the directed path $\delta \rightarrow \beta \rightarrow \gamma$ is such that every edge is unconfounded and δ is therefore also an ancestor of γ in $\mathcal{P}(\mathcal{G})$.

Theorem 3.15. Let $\mathcal{G} = (V, E)$ be a maximal DMG. Then $\mathcal{G}$ and $\mathcal{P}(\mathcal{G})$ are Markov equivalent.

Proof. If there is a μ-connecting walk in $\mathcal{G}$ from α to β given C then there is also a μ-connecting walk from α to β such that all colliders on the walk are in C. Assume ω is such a walk and of minimal length (i.e. there is no strictly shorter μ-connecting walk from α to β such that all colliders are in C). Every bidirected edge on the walk is also in $\mathcal{P}(\mathcal{G})$. Consider a directed edge, $\gamma \rightarrow \delta$. If it is not in $\mathcal{P}(\mathcal{G})$, then $\gamma \leftrightarrow \delta$ in $\mathcal{P}(\mathcal{G})$ and for each such edge we just substitute $\gamma \leftrightarrow \delta$ for the directed edge to obtain $\tilde{\omega}$, a walk in $\mathcal{P}(\mathcal{G})$ with the same (ordered) node set as ω. Note that this walk has a head at β.

Assume that γ is a node that has different collider status in ω than in $\tilde{\omega}$. In this case, γ (or, more precisely, this instance of γ) must be a noncollider on ω and a collider on $\tilde{\omega}$. Assume there is a head at γ on ω, say $\varepsilon \ast\!\!\rightarrow \gamma \rightarrow \delta$. As $\gamma \leftrightarrow_{\mathcal{G}} \delta$ it follows that $\varepsilon \ast\!\!\rightarrow_{\mathcal{G}} \delta$ which is a contradiction as this would create as strictly shorter, μ-connecting walk with all its colliders in C. If there is a tail at γ on ω, $\varepsilon \leftarrow \gamma \rightarrow \delta$, then either γ is also a noncollider on $\tilde{\omega}$, or $\varepsilon \leftrightarrow \gamma$ in $\mathcal{G}$. However, then $\varepsilon \leftrightarrow_{\mathcal{G}} \delta$ which is again a contradiction. This means that in $\mathcal{P}(\mathcal{G})$ there exists a walk with the same (ordered) node set as ω such that each node has the same collider status. Every collider is in C and therefore this walk is μ-connecting in $\mathcal{P}(\mathcal{G})$.

The other direction is immediate as the projection is a subgraph of $\mathcal{G}$. □

Example 3.16. If $\mathcal{G}$ is not maximal, then it need not be Markov equivalent with its confounding projection. As an example, consider $\alpha \rightarrow \gamma \rightarrow \beta$ such that $\gamma \rightarrow \beta$ is confounded. This graph is not maximal, and it is also not Markov equivalent with its confounding projection in which β is separated from α by the empty set.

Chapter 4

Directed correlation graphs

Stochastic models of dynamical systems are often explicitly *driven* by error processes. The system develops dynamically and the error process is the only new information that is put into the system. The error process is often thought of as erratic, 'white noise'. In some models, the error process is assumed to consist of independent processes, while in others the error processes may exhibit correlated behavior. This means that between time points, there is independence, but within time points the different error variables may depend on each other.

Example 4.1. As an example, we consider the following autoregressive model of order 2 with correlated errors, where $t \in \mathbb{Z}$ and $X_t = (X_t^\alpha, X_t^\beta, X_t^\gamma)$,

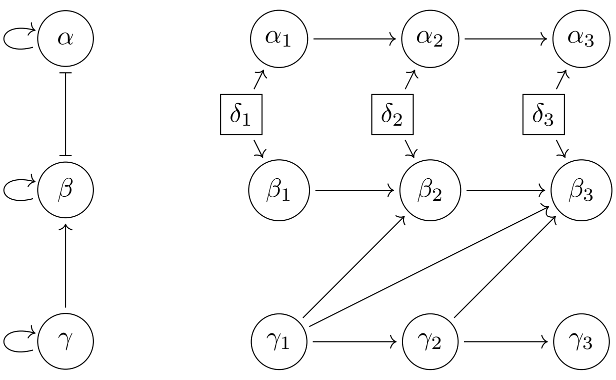


Figure 4.1: A similar figure is found in Paper **B**.

$$X_t = \begin{pmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{pmatrix} X_{t-1} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & b_{23} \\ 0 & 0 & 0 \end{pmatrix} X_{t-2} + \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} E_t + \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} D_t$$

where $E_t = (E_t^\alpha, E_t^\beta, E_t^\gamma)$ is a vector of noise variables and D_t is a noise variable such that $\{E_s^\phi, D_s\}_{s\in\mathbb{Z},\phi\in\{\alpha,\beta,\delta\}}$ is a family of independent random variables. We will think of the two first terms as autoregressive terms, and the two last terms as error terms. In the figure (left), the directed edges correspond to nonzero autoregressive coefficients while the blunt edge corresponds to correlation of error terms. The δ-nodes represent the D-process. The DAG on the right represents the independence structure in the time series (at lag 2) and we can think of the cDG as a 'rolled' version of the DAG.

One could write such models in greater generality, e.g., by allowing the unobserved D-process to enter in other ways. However, our main point is the observation that the one-dimensional D-process does not function as a general confounding process as its variables are independent across time points. In the figure, this means that the directed edges forwards in time, $\delta_i \to \delta_{i+1}$, are missing. Loosely speaking, the correlated errors create fewer dependences than a general confounding process does due to the independence of the error process variables between time points. If we wish a graphical representation of local independence in such a time series, or in another stochastic process, then we could use DMGs and obtain a Markov property. However, we can obtain a stronger Markov property by accounting for the possibility of correlated error processes. This is the motivation for studying *directed correlation graphs* as they allow a more fine-grained description of the local independence model of a dynamical system which is driven by correlated error processes.

In this chapter, Paper **B** describes the cDGs and proves a global Markov property in a specific model class. It studies Markov equivalence of cDGs and proves that determining Markov equivalence is computationally hard. In the following sections, we show that finding minimal graphical representations of cDGs is also computationally hard. Finally, we discuss some possible remedies for these computational issues.

Paper **B**

Søren Wengel Mogensen and Niels Richard Hansen. Graphical modeling of stochastic processes driven by correlated errors. 2020b

Graphical modeling of stochastic processes driven by correlated errors

Søren Wengel Mogensen and Niels Richard Hansen
University of Copenhagen

Abstract

We study a class of graphs that represent local independence structures in stochastic processes allowing for correlated error processes. Several graphs may encode the same local independences and we characterize such equivalence classes of graphs. The number of conditions in our characterizations grows superpolynomially as a function of the size of the node set in the graph. We show that deciding Markov equivalence is coNP-complete which suggests that our characterizations cannot be improved upon substantially. We prove a global Markov property in the case of a multivariate Ornstein-Uhlenbeck process which is driven by correlated Brownian motions.

1 Introduction

Graphical modeling studies how to relate graphs to properties of probability distributions (Lauritzen, 1996). There is a rich literature on graphical modeling of distributions of multivariate random variables (Maathuis et al., 2018), in particular on graphs as representations of conditional independences. In stochastic processes, local independence can be used as a concept analogous to conditional independence and several papers use graphs to encode local independences (Didelez, 2006, 2008; Aalen et al., 2012; Røysland, 2012; Mogensen et al., 2018; Mogensen and Hansen, 2020). Didelez (2000, 2008) studies graphical modeling of local independence of multivariate point processes. Mogensen et al. (2018) also consider diffusions. This previous work only models direct influence between coordinate processes in a multivariate stochastic process. We consider the case of correlated error processes which was also considered by Eichler and Didelez (2007); Eichler (2007, 2012b, 2013) in the time series case (i.e., stochastic processes indexed by discrete time). A specific local independence structure can be represented by several different graphs, and the characterization of such Markov equivalence classes is an important question in graphical modeling. We study these equivalence classes and characterize them. Our characterization is computationally demanding as it may involve exponentially many conditions (as a function of the number of nodes in the graphs). We give a complexity result, proving that deciding Markov equivalence in this class of graphs is coNP-hard, and therefore one would not except to find a characterization which is verified more easily.

The graphical results apply to various models of stochastic processes. As an example, we study systems of linear stochastic differential equations (SDEs), and in particular Ornstein-Uhlenbeck processes. Such models have been used in numerous fields such as psychology (Heath, 2000), neuroscience (Ricciardi and Sacerdote, 1979; Shimokawa et al., 2000; Ditlevsen and Lansky, 2005), finance (Stein and Stein, 1991; Schöbel and Zhu, 1999; Bormetti et al., 2010), biology (Bartoszek et al., 2017), and survival analysis (Aalen and Gjessing, 2004; Lee and Whitmore, 2006). In this paper, we show that Ornstein-Uhlenbeck processes with correlated driving Brownian motions satisfy a global Markov property with respect to a certain graph. Previous work in continous-time models considers independent error processes and the present work extends this framework to cases where the driving processes are correlated. To our knowledge, our result is the first such result in continuous-time models. In discrete-time models, i.e., time series, this is analogous to Eichler and Didelez (2007); Eichler (2007, 2012b, 2013). These papers consider graphical modeling of time series in discrete time with correlated errors and the graphical and algorithmic results we present also apply to these model classes.

Section 2 introduces local independence for Itô processes. Section 3 defines *directed correlation graphs* (cDGs) – the class of graphs that we will use throughout the paper to represent local independences in a stochastic process. In Section 3 we state a global Markov property for Ornstein-Uhlenbeck processes. Section 4 gives a characterization of the cDGs that encode the same independences. This directly leads to an algorithm for checking equivalence of cDGs. This algorithm runs in exponential time (in the number of nodes in the graphs). We prove in Section 5 that deciding Markov equivalence is coNP-complete.

2 Local independence

Before diving into a formal introduction, we will consider a motivating example.

Example 1. Consider the three-dimensional Ornstein-Uhlenbeck process, which solves the following stochastic differential equation

$$\mathrm{d}\begin{pmatrix} X_t^\alpha \\ X_t^\beta \\ X_t^\gamma \end{pmatrix} = \underbrace{\begin{pmatrix} M_{\alpha\alpha} & 0 & 0 \\ M_{\beta\alpha} & M_{\beta\beta} & 0 \\ 0 & 0 & M_{\gamma\gamma} \end{pmatrix}}_{=M} \begin{pmatrix} X_t^\alpha \\ X_t^\beta \\ X_t^\gamma \end{pmatrix} \mathrm{d}t + \underbrace{\begin{pmatrix} \sigma_\alpha & 0 & 0 & 0 \\ 0 & \sigma_\beta & 0 & \rho_\beta \\ 0 & 0 & \sigma_\gamma & \rho_\gamma \end{pmatrix}}_{=\sigma_0} \mathrm{d}\begin{pmatrix} W_t^1 \\ W_t^2 \\ W_t^3 \\ W_t^4 \end{pmatrix}$$

where $(W_t^1, W_t^2, W_t^3, W_t^4)^T$ is a standard four-dimensional Brownian motion. In this example, all entries in the matrix M above that are not explicitly 0 are assumed nonzero and likewise for σ_0.

The interpretation of the stochastic differential equation via the Euler-Maru-

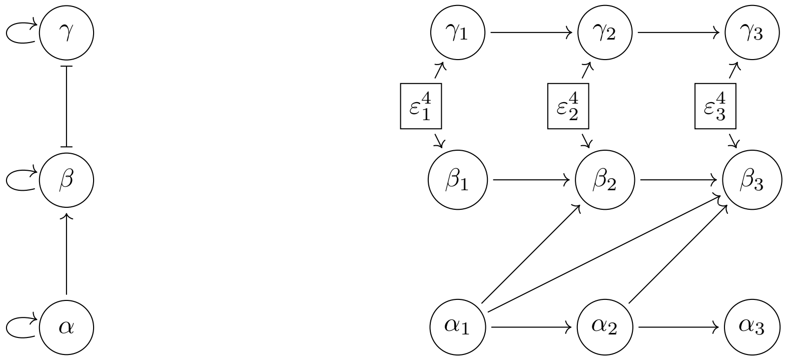


Figure 1: A local independence graph (left) and a 'rolled out' graph (right) where time is made explicit. The two graphs represent the same local independence structure. A node δ for $\delta \in \{\alpha, \beta, \gamma\}$ represents the increments of the X_t^δ-process at time t. On the right, the ε^4-process is a 'white noise' process that creates dependence between X_t^β and X_t^γ. In the 'rolled' version of the graph (left) this is represented by a *blunt* edge, $\beta \mapsto\!\!\!\!\!\dashv \gamma$.

yama scheme yields the update equation

$$\begin{aligned}
\Delta \tilde{X}_t^\alpha &= \tilde{X}_{t+\Delta}^\alpha - \tilde{X}_t^\alpha = M_{\alpha\alpha}\tilde{X}_t^\alpha + \sqrt{\Delta}\sigma_\alpha \varepsilon_t^1 \\
\Delta \tilde{X}_t^\beta &= \tilde{X}_{t+\Delta}^\beta - \tilde{X}_t^\beta = M_{\beta\alpha}\tilde{X}_t^\alpha + M_{\beta\beta}\tilde{X}_t^\beta + \sqrt{\Delta}\left(\sigma_\beta \varepsilon_t^2 + \rho_\beta \varepsilon_t^4\right) \\
\Delta \tilde{X}_t^\gamma &= \tilde{X}_{t+\Delta}^\gamma - \tilde{X}_t^\gamma = M_{\gamma\gamma}\tilde{X}_t^\gamma + \sqrt{\Delta}\left(\sigma_\gamma \varepsilon_t^3 + \rho_\gamma \varepsilon_t^4\right)
\end{aligned}$$

where $\varepsilon_t \sim \mathcal{N}(0, I)$. The Euler-Maruyama scheme evaluated in $t = n\Delta$ for $n \in \mathbb{N}_0$ gives a process, $(\tilde{X}_{n\Delta})_{n \geq 0}$, which, as $\Delta \to 0$, converges to the Ornstein-Uhlenbeck process, $(X_t)_{t \geq 0}$, solving the stochastic differential equation. From the update equations we see that the infinitesimal increment of each coordinate depends on that coordinate's own value, and coordinate β depends, in addition, on coordinate α (because $M_{\beta\alpha} \neq 0$). Moreover, the increments for coordinates β and γ are correlated as they share the error variable ε_t^4. Figure 1 provides a graphical representation, with arrows readily read off from the drift matrix M and the diffusion matrix $\sigma\sigma^T$. The 'rolled out' graph in the figure is the DAG that corresponds to the Euler-Maruyama scheme.

The main purpose of this paper is to clarify the mathematical interpretation of the local independence graph in Figure 1, and our results include a characterization of all graphs with equivalent mathematical content. The novelty is that we allow for $\sigma_0\sigma_0^T$ to be nondiagonal as in the example above.

2.1 Itô processes and local independence graphs

We will for the purpose of this paper focus on vector-valued, continuous-time stochastic processes with continuous sample paths. Thus let $X = (X_t)_{t \in \mathcal{T}}$

denote such an n-dimensional process with time index $t \in \mathcal{T} \subseteq \mathbb{R}$ and with $X_t = (X_t^\alpha)_{\alpha \in [n]} \in \mathbb{R}^n$ being a real-valued vector indexed by $[n] = \{1, \ldots, n\}$. The time index set $\mathcal{T}$ will in practice be of the forms $[0, T]$, $[0, \infty)$, or $\mathbb{R}$, however, we will in general just assume that $\mathcal{T}$ is an interval containing 0.

The purpose of *local independence* is to give a mathematically precise definition of what it means for the historical evolution of one coordinate, α, to *not* be predictive of the infinitesimal increment of another coordinate, β, given the historical evolution of a set, $C \subseteq [n]$, of coordinates. As such, it is a continuous-time version of Granger causality (see, e.g., Granger and Newbold, 1986; Aalen, 1987; Didelez, 2008; Commenges and Gégout-Petit, 2009), and its formulation is directly related to filtration problems for stochastic processes. In a statistical context, local independence allows us to express simplifying structural constraints that are directly useful for forecasting and such constraints are also useful for causal structure learning.

The process X is defined on the probability space $(\Omega, \mathcal{F}, P)$ and we let $\sigma(X_s^\delta; s \leq t, \delta \in D) \subseteq \mathcal{F}$ denote the σ-algebra on Ω generated by X_s^δ for all $s \in \mathcal{T}$ up to time t and all $\delta \in D$. For technical reasons, we define $\mathcal{F}_t^D$ to be the P-completion of the σ-algebra

$$\bigcap_{t' > t} \sigma(X_s^\delta; s \leq t', \delta \in D),$$

so that $(\mathcal{F}_t^D)_{t \in \mathcal{T}}$ is a complete, right-continuous filtration for all $D \subseteq [n]$. We will let $\mathcal{F}_t = \mathcal{F}_t^{[n]}$ denote the filtration generated by all coordinates of the process. Within this setup we will restrict attention to Itô processes with continuous drift and constant diffusion coefficient.

Definition 2 (Regular Itô processes)**.** We say that X is a regular Itô process if there exists a continuous, $\mathcal{F}_t$-adapted process, λ, with values in $\mathbb{R}^n$, and an $n \times n$ invertible matrix σ such that

$$W_t = \sigma^{-1} \left(X_t - X_0 - \int_0^t \lambda_s \mathrm{d}s \right)$$

is a standard $\mathcal{F}_t$-adapted Brownian motion.

A regular Itô process is sometimes written in differential form as

$$\mathrm{d}X_t = \lambda_t \mathrm{d}t + \sigma \ \mathrm{d}W_t. \tag{1}$$

Here λ_t is known as the drift of the process and σ as the (constant) diffusion coefficient. We define the *diffusion matrix* for a regular Itô process as the positive definite matrix

$$\Sigma = \sigma \sigma^T. \tag{2}$$

Observe that the process X_t may, as in Example 1, be defined as the solution of the stochastic differential equation

$$\mathrm{d}X_t = \lambda_t \mathrm{d}t + \sigma_0 \ \mathrm{d}W_t \tag{3}$$

for an m-dimensional standard Brownian motion W and with the diffusion coefficient σ_0 an $n \times m$ matrix. If σ_0 has rank n, such a solution is also a regular Itô process with diffusion matrix $\Sigma = \sigma_0\sigma_0^T$. Indeed, we can take $\sigma = (\sigma_0\sigma_0^T)^{1/2}$ in Definition 2. Observe also that for any regular Itô process,

$$X_t - X_0 - \int_0^t \lambda_s \mathrm{d}s = \sigma W_t$$

is an $\mathcal{F}_t$-martingale and $\int_0^t \lambda_s \mathrm{d}s$ is the compensator of X_t in its Doob-Meyer decomposition.

Definition 3. Let X be a regular Itô process with drift λ, let $\alpha, \beta \in [n]$ and let $C \subseteq [n]$. We say that β is locally independent of α given C, and write $\alpha \not\to \beta \mid C$, if the process

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

is a version of $t \mapsto E(\lambda_t^\beta \mid \mathcal{F}_t^{C \cup \{\alpha\}})$.

It follows immediately from the definition that $\alpha \not\to \beta \mid [n] \setminus \{\alpha\}$ if λ_t^β is $\mathcal{F}_t^{[n]\setminus\{\alpha\}}$-measurable. That is, if λ_t^β does not depend on the sample path of coordinate α.

The definition below of a local independence graph generalizes the definitions of Didelez (2008) and Mogensen and Hansen (2020) for continuous time stochastic processes to allow for a nondiagonal Σ. Eichler (2007) gives the analogous definition in the case of time series with correlated errors.

Definition 4 (Local independence graph)**.** Consider a regular Itô diffusion with diffusion matrix Σ. A local independence graph is a graph, $\mathcal{D}$, with nodes $[n]$ such that

$$\alpha \not\to_{\mathcal{D}} \beta \quad \Rightarrow \quad \alpha \not\to \beta \mid [n] \setminus \{\alpha\}$$

and such that for $\alpha \neq \beta$

$$\alpha \not\mapsto_{\mathcal{D}} \beta \quad \Rightarrow \quad \Sigma_{\alpha\beta} = 0$$

where $\to_{\mathcal{D}}$ denotes a directed edge in $\mathcal{D}$ and $\alpha \mapsto_{\mathcal{D}} \beta$ denotes a blunt edge.

It follows from the definitions that we can read off a local independence graph for a regular Itô diffusion directly from λ and Σ by including the edge $\alpha \to_{\mathcal{D}} \beta$ whenever λ_t^β depends upon coordinate α and the edge $\alpha \mapsto_{\mathcal{D}} \beta$ whenever $\Sigma_{\alpha\beta} \neq 0$. However, it is possible that the functional form of λ_t^β appears to depend on the coordinate α, while actually $\alpha \not\to \beta \mid [n] \setminus \{\alpha\}$. In such a case, the resulting local independence graph will not be minimal.

Using μ-separation to define a separation model for directed graphs, a main result in Mogensen et al. (2018) is the fact that for regular Itô diffusions with a diagonal $\sigma\sigma^T$, the local independence graph satisfies a global Markov property – if certain integrability constraints are satisfied. A local independence graph satisfying the global Markov property combined with graph algorithms for determining μ-separation allows us to answer the filtration question: for $D \subseteq [n]$ and $\beta \in [n]$, which coordinates in D does $E(\lambda_t^\beta \mid \mathcal{F}_t^D)$ depend upon?

2.2 Itô diffusions

Itô diffusions with a constant diffusion coefficient are particularly interesting examples of Itô processes. They are Markov processes, but they are not closed under marginalization. One reason for the interest in the general class of Itô processes is that they are closed under marginalization, and marginalizing an Itô diffusion gives, in particular, an Itô process.

A regular Itô diffusion is a regular Itô process such that the drift is of the form

$$\lambda_t = \lambda(X_t)$$

for a continuous function $\lambda : \mathbb{R}^n \to \mathbb{R}^n$. In differential form

$$\mathrm{d}X_t = \lambda(X_t)\ \mathrm{d}t + \sigma\ \mathrm{d}W_t.$$

Proposition 5. Let X be a regular Itô diffusion with a continuously differentiable drift λ. If $\partial_\alpha \lambda_\beta = 0$ then $\alpha \not\rightarrow \beta \mid [n] \setminus \{\alpha\}$.

Proof. If $\partial_\alpha \lambda_\beta = 0$, then

$$\lambda_t^\beta = \lambda_\beta((X_t^\delta)_{\delta \in [n] \setminus \{\alpha\}})$$

is $\mathcal{F}_t^{[n] \setminus \{\alpha\}}$-measurable. □

While Proposition 5 is straightforward from the definitions, it gives a simple operational procedure for determining that $\alpha \not\rightarrow \beta \mid [n] \setminus \{\alpha\}$ and thus a local independence graph according to Definition 4.

Example 6 (Smoluchowski diffusion)**.** The purpose of this example is to link the notion of local independence and the local independence graph to classical undirected graphical models for a special class of diffusions that are widely studied in equilibrium statistical physics. A Smoluchowski diffusion is a regular Itô diffusion with

$$\lambda(x) = -\nabla V(x)$$

for a continuously differentiable function $V : \mathbb{R}^n \to \mathbb{R}$ and $\sigma = \sqrt{2\tau I}$ for a constant $\tau > 0$. Thus the diffusion matrix $\Sigma = 2\tau I$ is diagonal. The function V is called the potential and τ is called a temperature parameter. Since the drift is a gradient, the dynamics of a Smoluchowski diffusion is a gradient flow perturbed by white noise. If $V(x) \to \infty$ for $\|x\| \to \infty$ and

$$Z = \int e^{-\frac{1}{\tau}V(x)} \mathrm{d}x < \infty,$$

the diffusion has the Gibbs measure with density

$$\pi(x) = \frac{1}{Z} e^{-\frac{1}{\tau}V(x)}$$

as equilibrium distribution, see Proposition 4.2 in Pavliotis (2014). When V is twice differentiable, Proposition 5 gives a local independence graph $\mathcal{D}$ with arrows $\alpha \to_{\mathcal{D}} \beta$ whenever $\partial_\alpha \lambda_\beta = \partial_\alpha \partial_\beta V \neq 0$. Since

$$\partial_\alpha \lambda_\beta = \partial_\alpha \partial_\beta V = \partial_\beta \partial_\alpha V = \partial_\beta \lambda_\alpha$$

the graph $\mathcal{D}$ enjoys the symmetry property that $\alpha \to_{\mathcal{D}} \beta$ if and only if $\beta \to_{\mathcal{D}} \alpha$. We denote by $\mathcal{G}$ the undirected version of $\mathcal{D}$. For any $\alpha, \beta \in [n]$ with $\alpha \not-_{\mathcal{G}} \beta$ it follows from $\partial_\alpha \partial_\beta V = \partial_\beta \partial_\alpha V = 0$ that

$$V(x) = V_1(x_\alpha, x_{-\{\alpha,\beta\}}) + V_2(x_\beta, x_{-\{\alpha,\beta\}})$$

where $x_{-\{\alpha,\beta\}}$ denotes the vector x with coordinates x_α and x_β removed. From this decomposition of V we see that π has the pairwise Markov property with respect to $\mathcal{G}$, and it follows from the Hammersley-Clifford theorem that π factorizes according to $\mathcal{G}$. That is, the potential has the following additive decomposition

$$V(x) = \sum_{c \in \mathcal{C}(\mathcal{G})} V_c(x_c)$$

where $\mathcal{C}(\mathcal{G})$ denotes the cliques of $\mathcal{G}$. This establishes a correspondence between local independences for a Smoluchowski diffusion and Markov properties of its equilibrium distribution.

For Smoluchowski diffusions we have demonstrated a strong link between local independences representing structural constraints of the dynamics on the one side and Markov properties of an equilibrium distribution on the other side. We emphasize that this link is a consequence of the symmetry of the drift of Smoluchowski diffusions combined with the diffusion matrix being a scalar multiple of the identity matrix. For diffusions with a non-gradient drift or with a more complicated diffusion matrix the equilibrium distribution may have no conditional independences even though there are strong structural constraints on the dynamics of the process that can be expressed in terms of a sparse local independence graph. The simplest process which can illustrate this is the Ornstein-Uhlenbeck process.

Example 7 (Ornstein-Uhlenbeck processes)**.** A regular Itô diffusion with drift

$$\lambda(x) = M(x - \mu)$$

for an $n \times n$ matrix M and a n-dimensional vector μ is called a regular Ornstein-Uhlenbeck process. It follows from Proposition 5 that $\alpha \not\to \beta \mid [n] \setminus \{\alpha\}$ if $M_{\beta\alpha} = 0$. If M is a stable matrix, the Ornstein-Uhlenbeck process has an invariant Gaussian distribution $\mathcal{N}(\mu, \Gamma_\infty)$, where Γ_∞ solves the Lyapunov equation

$$M\Gamma_\infty + \Gamma_\infty M^T + \Sigma = 0,$$

see Proposition 3.5 in Pavliotis (2014) or Theorem 2.12 in Jacobsen (1993).

If M is also symmetric, then λ is a gradient, and if $\Sigma = 2\tau I$ we see that the solution of the Lyapunov equation is $\Gamma_\infty = -\tau M^{-1}$, and λ is the negative gradient of the quadratic potential

$$V(x) = -\frac{1}{2\tau}(x-\mu)^T M(x-\mu) = \frac{1}{2}(x-\mu)^T \Gamma_\infty^{-1}(x-\mu).$$

Thus the equilibrium distribution is a Gaussian graphical model whose graph $\mathcal{G}$ has edges determined by the non-zero entries of $\Gamma_\infty^{-1} = -\frac{1}{\tau}M$. For this Smoluchowski diffusion we see very explicitly that the edge $\alpha - \beta$ is in $\mathcal{G}$ if and only if both $\alpha \to \beta$ and $\beta \to \alpha$ are in the local independence graph $\mathcal{D}$. However, it is not difficult to find an asymmetric but stable matrix M such that Γ_∞^{-1} is a dense matrix, even if $\Sigma = I$, and the local independence graph cannot in general be determined from Markov properties of the invariant distribution.

For a general M and general Σ, and with $D \subseteq [n]$, it follows that

$$\begin{aligned} E(\lambda_t^\beta \mid \mathcal{F}_t^D) &= \sum_{\delta \in V} M_{\beta\delta}(E(X_t^\delta \mid \mathcal{F}_t^D) - \mu_\delta) \\ &= \sum_{\delta \in \mathrm{pa}(\beta)} M_{\beta\delta}(E(X_t^\delta \mid \mathcal{F}_t^D) - \mu_\delta), \end{aligned}$$

where $\mathrm{pa}(\beta) = \{\delta \mid M_{\beta\delta} \neq 0\}$ denotes the set of parents of β in the local independence graph determined by M and Σ. Thus determining by Definition 3 if $\alpha \not\to \beta \mid C$ amounts to determining if $E(X_t^\delta \mid \mathcal{F}_t^C)$ are versions of $E(X_t^\delta \mid \mathcal{F}_t^{C\cup\{\alpha\}})$ for $\delta \in \mathrm{pa}(\beta)$. In words, this means that if we can predict the values of all the processes, X_t^δ for $\delta \in \mathrm{pa}(\beta)$, that enter into the drift of coordinate β just as well from the C-histories as we can from the $C \cup \{\alpha\}$-histories then β is locally independent of α given C.

The following sections of this paper will develop the graph theory needed to answer questions about local independence via graphical properties of the local independence graph. This theory can be applied as long as the processes considered have the global Markov property with respect to the local independence graph, and we show that this is the case for regular Ornstein-Uhlenbeck processes.

3 Graphs and Markov properties

3.1 Directed correlation graphs

A *graph* is a pair (V, E) where V is a set of nodes and E is a set of edges. Every edge is between a pair of nodes. Edges can be of different types. In this paper, we will consider *directed* edges, $\to$, *bidirected* edges, $\leftrightarrow$, and *blunt* edges, $\mapsto$. Let $\alpha, \beta \in V$. Note that $\alpha \to \beta$ and $\beta \to \alpha$ are different edges. We do not distinguish between $\alpha \leftrightarrow \beta$ and $\beta \leftrightarrow \alpha$, nor between $\alpha \mapsto \beta$ and $\beta \mapsto \alpha$. We allow directed, and bidirected loops (self-edges), $\alpha \to \alpha$ and $\alpha \leftrightarrow \alpha$, but not blunt loops, $\alpha \mapsto \alpha$. If α and β are joined by a blunt edge, $\alpha \mapsto \beta$, then we say that they are *spouses*.

We use $\alpha \sim \beta$ to symbolize a generic edge between $\alpha \in V$ and $\beta \in V$ of any of these three types. We use $\alpha \ast\!\!\rightarrow \beta$ to symbolize that either $\alpha \rightarrow \beta$ or $\alpha \leftrightarrow \beta$.

Definition 8. Let $\mathcal{G} = (V, E)$ be a graph. We say that $\mathcal{G}$ is a *directed correlation graph* (cDG) if every edge is directed or blunt. We say that $\mathcal{G}$ is a *directed mixed graphs* (DMG) if every $e \in E$ is either directed or bidirected.

The class of DMGs was studied by Mogensen et al. (2018); Mogensen and Hansen (2020). Eichler (2007, 2012b) studied classes of graphs similar to cDGs as well as a class of graphs which contains both the DMGs and the cDGs as subclasses.

A *walk* is an ordered, alternating sequence of nodes (γ_i) and edges $(\sim_i)$ such that each edge, $\sim_i$, is between γ_i and γ_{i+1},

$$\gamma_1 \sim_1 \gamma_2 \sim_2 \cdots \sim_k \gamma_{k+1}$$

We say that γ_1 and γ_{k+1} are *endpoint nodes.* We say that a walk is *trivial* if it has no edges and therefore only a single node. We say that a walk is *nontrivial* if it contains at least one edge. We say that an nonendpoint node, γ_i, is a *collider* if one of the following holds

$$\begin{aligned} \gamma_{i-1} \ast\!\!\rightarrow \gamma_i &\leftarrow\!\!\ast\ \gamma_{i+1}, \\ \gamma_{i-1} \ast\!\!\rightarrow \gamma_i &\vdash\!\dashv \gamma_{i+1}, \\ \gamma_{i-1} \vdash\!\dashv \gamma_i &\leftarrow\!\!\ast\ \gamma_{i+1}, \\ \gamma_{i-1} \vdash\!\dashv \gamma_i &\vdash\!\dashv \gamma_{i+1}, \end{aligned}$$

and otherwise we say that it is a *noncollider.* We say that $\alpha \ast\!\!\rightarrow \beta$ has a *head* at β, and that $\alpha \rightarrow \beta$ has a *tail* at α. We say that $\alpha \vdash\!\dashv \beta$ has a *stump* at α. We say that edges $\alpha \vdash\!\dashv \beta$ and $\alpha \ast\!\!\rightarrow \beta$ have a *neck* at β. It follows that γ_i above is a collider if and only if both adjacent edges have a neck at γ_i. A *path* is a walk such that every node occurs at most once. We say that a path from α to β is *directed* if every edge on the path is directed and pointing towards β. If there is a directed path from α to β, then we say that α is an *ancestor* of β. We let $\mathrm{an}(\beta)$ denote the set of ancestors of β, and for $C \subseteq V$, we define $\mathrm{an}(C) = \cup_{\gamma \in C} \mathrm{an}(\gamma)$. Note that $C \subseteq \mathrm{an}(C)$. We will use *$\mu$-connecting walks* and *$\mu$-separation* to encode independence structures in cDGs.

Definition 9 (μ-connecting walk, Mogensen and Hansen (2020))**.** Consider a nontrivial walk, ω,

$$\alpha \sim_1 \gamma_2 \sim_2 \cdots \sim_{k-1} \gamma_k \sim_k \beta$$

and a set $C \subseteq V$. We say that ω is *μ-connecting* from α to β given C if $\alpha \notin C$, every collider on ω is in $\mathrm{an}(C)$, no noncollider is in C, and $\sim_k$ has a head at β.

It is essential that the above definition uses walks, and not only paths. As an example consider $\alpha \vdash\!\dashv \beta \leftarrow \gamma$. In this graph, there is no μ-connecting path from α to β given β, but there is a μ-connecting walk.

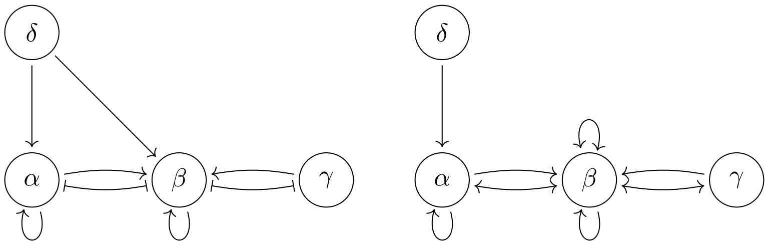


Figure 2: Example cDG (left) and example DMG (right). The blunt edges in a cDG correspond to correlated driving processes which is different from the bidirected edges of a DMG as those correspond to marginalization, i.e., unobserved processes. The notion of μ-separation can be applied to both classes of graphs, however, the separations models are different. Left: cDG on nodes $V = \{\alpha, \beta, \gamma, \delta\}$. γ is μ-separated from δ by α as $\beta \notin \text{an}(\alpha)$ is a collider on any walk from δ to γ. On the other hand, α is not μ-separated from β given $\emptyset$ as e.g. $\beta \mapsto\!\!\!\!\!\dashv \alpha \rightarrow \alpha$ is μ-connecting given $\emptyset$. That walk is not μ-connecting from β to α given α, however, $\beta \leftarrow \delta \rightarrow \alpha$ is μ-connecting from α to β given α. We see that α is μ-separated from β given $\{\alpha, \delta\}$. Right: bidirected edges have heads at both ends and this means that $\beta \leftrightarrow \alpha$ is μ-connecting from β to α given any subset of $V \setminus \{\beta\}$. This is not true in the cDG (left).

Definition 10 (μ-separation, Mogensen and Hansen (2020))**.** Let $\mathcal{G} = (V, E)$ be a cDG or a DMG and let $A, B, C \subseteq V$. We say that B is *μ-separated* from A given C in $\mathcal{G}$ if there is no μ-connecting walk from any $\alpha \in A$ to any $\beta \in B$ given C.

Mogensen and Hansen (2020) introduced μ-separation as a generalization of δ-separation (Didelez, 2000, 2008), however, only in DMGs, and not in cDGs.

Remark 11. Eichler (2007); Eichler and Didelez (2010); Eichler (2012a) describe graphs that represent local independence (or Granger non-causality) in time series. The cDGs are a subclass of the graphs in that line of work, however, we use a different representation. In the aforementioned papers, blunt edges are represented by $-$ while we use $\mapsto\!\!\!\!\!\dashv$. The former notation could suggest that a blunt edge acts like an edge with tails in both ends, however, this is not the case. It also does not act like the bidirected edges in a DMG, and this warrants the usage of an edge with a third kind of mark.

Also note that while this is not paramount in the case of cDGs, notational clarity and simplicity become more important when considering graphical marginalizations of these graphs. In this case, one needs to consider also edges that, when composed with other edges, act like a blunt edge in one end and like a directed edge in the other and this is described by Eichler (2012b). Using our notation, this is naturally visualized by the edge $\mapsto$. We will not consider this larger class of graphs in this paper, however, we choose this notation as it extends naturally to that case.

3.2 A global Markov property

In this section, we state a result showing that an Ornstein-Uhlenbeck process satisfies a global Markov property with respect to its local independence graph and we use V to denote both the node set of a graph and the set indexing the coordinate processes of a multivariate process. In the case of a diagonal Σ the global Markov property was shown in Mogensen et al. (2018), and we extend this to the case of nondiagonal Σ, i.e., allowing for correlated driving Brownian motions. The proof is found in Appendix A and it uses a set of equations describing the conditional mean processes, $t \mapsto \mathrm{E}\big[X_t^U \mid \mathcal{F}_t^W\big]$, $V = U \dot\cup W$ (Liptser and Shiryayev, 1977). From this representation, we can reason about the measurability of the conditional mean processes. We first give a more general definition of local independence in Itô processes to allow non-singleton sets A and B.

Definition 12. Let X be a regular Itô process with drift λ, and let $A, B, C \subseteq V$. We say that B is locally independent of A given C, and write $A \not\rightarrow B \mid C$, if for all $\beta \in B$ the process

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

is a version of $t \mapsto E(\lambda_t^\beta \mid \mathcal{F}_t^{C \cup A})$.

Theorem 13. Let $X = (X_t)_{t \geq 0}$ be a regular Ornstein-Uhlenbeck process with local independence graph $\mathcal{D} = (V, E)$ (Definition 4), and let $A, B, C \subseteq V$. Assume that X_0 is a vector of independent and non-degenerate random variables. If B is μ-separated from A given C in $\mathcal{D}$, then B is locally independent of A given C.

4 Markov equivalence

Different cDGs can encode the same separation models, and in this section we will describe the so-called *Markov equivalence classes* of cDGs. When $\mathcal{D} = (V, E)$ is a cDG, we define its independence model, $\mathcal{I}(\mathcal{D})$, as the collection of μ-separations that hold, i.e.,

$$\mathcal{I}(\mathcal{D}) = \{(A, B, C) : A, B, C \subseteq V,\ A \perp_\mu B \mid C\ [\mathcal{D}]\}.$$

Definition 14 (Markov equivalence)**.** Let $\mathcal{D}_1 = (V, E_1)$, $\mathcal{D}_2 = (V, E_2)$ be cDGs. We say that $\mathcal{D}_1$ and $\mathcal{D}_2$ are *Markov equivalent* if $\mathcal{I}(\mathcal{D}_1) = \mathcal{I}(\mathcal{D}_2)$.

For any finite set V, Markov equivalence is an equivalence relation on the set of cDGs with node set V. We let $[\mathcal{G}]$ denote the Markov equivalence class of a graph $\mathcal{G}$. For a cDG, $\mathcal{D} = (V, E)$, and a directed or blunt edge e between $\alpha, \beta \in V$, we use $\mathcal{D} + e$ to denote the cDG $(V, E \cup \{e\})$.

Definition 15 (Maximality)**.** Let $\mathcal{D}$ be a cDG. We say that $\mathcal{D}$ is *maximal* if no edge can be added Markov equivalently, i.e., if for every edge, e, which is not in $\mathcal{D}$, $\mathcal{D}$ and $\mathcal{D} + e$ are not Markov equivalent.

The following proposition can be found in Mogensen and Hansen (2020) in the case of DGs.

Proposition 16. Let $\mathcal{D} = (V, E)$ be a cDG. Then $\alpha \rightarrow_{\mathcal{D}} \beta$ if and only if $\alpha \perp_\mu \beta \mid V \setminus \{\alpha\}$ does not hold.

Proof. If the edge is in the graph, it is μ-connecting given any subset of V that does not contain α, in particular given $V \setminus \{\alpha\}$. On the other hand, assume $\alpha \rightarrow \beta$ is not in the graph. Any μ-connecting walk from α to β must have a head at β,

$$\alpha \sim \ldots \sim \gamma \rightarrow \beta.$$

We must have that $\gamma \neq \alpha$, and it follows that γ is in the conditioning set, i.e., the walk is closed. □

We can decide μ-separation by considering separation in a certain undirected graph, an *augmented graph*, which is a generalization of a *moral graph* (Cowell et al., 1999). Richardson and Spirtes (2002); Richardson (2003) used a similar approach to decide m-separation in *ancestral graphs* and *acyclic direced mixed graphs.* Didelez (2000) used a moral graph to decide δ-separation in DGs. When $\mathcal{D} = (V, E)$ is a cDG and $\bar{V} \subseteq V$, we let $\mathcal{D}_{\bar{V}}$ denote the induced graph on nodes $\bar{V}$, i.e., $\mathcal{D}_{\bar{V}} = (\bar{V}, \bar{E})$,

$$\bar{E} = \{e \in E : \ e \text{ is between } \alpha, \beta \in \bar{V}\}.$$

We say that α and β are *collider connected* if there exists a walk from α to β such that every nonendpoint node is a collider. The augmented graph of a cDG is the undirected graph where all collider connected pairs of nodes are adjacent (omitting loops). Given an undirected graph and three disjoint subsets of nodes A, B, and C, we say that A and B are *separated* by C if every path between $\alpha \in A$ and $\beta \in B$ intersects C.

Proposition 17 (Augmentation criterion for μ-separation)**.** Let $\mathcal{D} = (V, E)$ be a cDG such that $\gamma \rightarrow \gamma$ for all $\gamma \in V$. Let $A, B, C \subseteq V$, and assume that $B = \{\beta_1, \ldots, \beta_j\}$. Let $B^p = \{\beta_1^p, \ldots, \beta_j^p\}$ and define the graph $\mathcal{D}(B)$ with node set $V \,\dot{\cup}\, B^p$ such that $\mathcal{D}_V = \mathcal{D}$ and

$$\alpha \rightarrow_{\mathcal{D}(B)} \beta_i^p \text{ if } \alpha \rightarrow_{\mathcal{D}} \beta_i \text{ and } \alpha \in V, \beta_i \in B.$$

Then $A \perp_\mu B \mid C \ [\mathcal{D}]$ if and only if $A \setminus C$ and B^p are separated by C in the augmented graph of $\mathcal{D}(B)_{\mathrm{an}(A \cup B^p \cup C)}$.

Proof. The proofs of Propositions D.2 and D.4 by Mogensen and Hansen (2020) give the result. First one shows that $A \perp_\mu B \mid C \ [\mathcal{D}]$ if and only if $A \setminus C \perp_m B^p \mid C \ [\mathcal{D}(B)]$. The second statement is then shown to be equivalent to separation in the relevant augmented graph using Theorem 1 in Richardson (2003). Richardson (2003) studies acyclic graphs, however, the proof also applies to cyclic graphs as noted in the paper. □

In graphs that represent conditional independence in multivariate distributions, such as ancestral graphs and acyclic directed mixed graphs, one can use *inducing paths* to characterize which nodes cannot be separated by any conditioning set (Verma and Pearl, 1991; Richardson and Spirtes, 2002). In DMGs, inducing paths can be defined similarly (Mogensen and Hansen, 2020). In cDGs, we define both inducing paths and *weak inducing paths*. We say that a path is a *collider path* if every nonendpoint node on the path is a collider.

Definition 18 (Inducing path (strong))**.** A (nontrivial) collider path from α to β is a *(strong) inducing path* if the final edge has a head at β and every nonendpoint node is an ancestor of α or of β.

A *cycle* is a path $\alpha \sim \ldots \sim \beta$ composed with an edge $\beta \sim \alpha$. Mogensen and Hansen (2020) also allow cycles in the definition of inducing paths. In the following, we assume that $\alpha \to \alpha$ for all $\alpha \in V$ and therefore this would be an unnecessary complication. We see immediately that in a cDG, the only inducing path is a directed edge. However, we include this definition to conform with the terminology in DMGs where more elaborate inducing paths exist. If we drop one of the conditions from Definition 18, then we obtain a graphical structure which is more interesting in cDGs, a *weak inducing path.*

Definition 19 (Weak inducing path)**.** A (nontrivial) collider path between α and β is a *weak inducing path* if every nonendpoint node is an ancestor of α or β.

We note that a strong inducing path is also a weak inducing path. Furthermore, if there is a weak inducing path from α to β, there is also one from β to α. Also note that a *weak* inducing path is most often called an inducing path in the literature on acyclic graphs.

Proposition 20. Let $\mathcal{D} = (V, E)$ be a cDG such that $\alpha \to \alpha$ for all $\alpha \in V$. There is a weak inducing path between α and β if and only if there is no $C \subseteq V \setminus \{\alpha, \beta\}$ such that $\alpha \perp_\mu \beta \mid C$.

Mogensen and Hansen (2020) showed a similar result in the case of strong inducing paths in DMGs.

Proof. Assume first that there is no weak inducing path from α to β in $\mathcal{D}$, and define

$$D(\alpha, \beta) = \{\gamma \in \text{an}(\alpha, \beta) \mid \gamma \text{ and } \beta \text{ are collider connected } \} \setminus \{\alpha, \beta\}.$$

We will show that β is μ-separated from α by $D(\alpha, \beta)$. We can assume that $\alpha \neq \beta$ as we have assumed that all nodes have loops. If there is a μ-connecting walk from α to β given $C \subseteq V \setminus \{\alpha, \beta\}$, then there is also a μ-connecting walk which is a path composed with a directed edge, $\gamma \to \beta$. We must have that $\gamma \neq \alpha$, and if $\gamma \neq \beta$ then the walk is closed by $D(\alpha, \beta)$. Assume instead that

$\gamma = \beta$. Let π denote some path between α and β. Blunt and directed edges are weak inducing paths (in either direction) so π must be of length 2 or more,

$$\alpha = \gamma_0 \overset{e_0}{\sim} \gamma_1 \overset{e_1}{\sim} \dots \overset{e_{j-1}}{\sim} \gamma_j \overset{e_j}{\sim} \beta.$$

There must exist $i \in \{0, 1, \dots, j\}$ such that either γ_i is not collider connected to β along π or $\gamma_i \notin \mathrm{an}(\alpha, \beta)$. Let i_+ denote the largest such number in $\{0, 1, \dots, j\}$. Assume first that γ_{i_+} is not collider connected to β along π. In this case, $i_+ \neq j$. Then $\gamma_{i_+ + 1}$ is a noncollider on π and it is in $D(\alpha, \beta)$, and it follows that π is not μ-connecting. Note that necessarily $\gamma_{i_+ + 1} \neq \alpha, \beta$. On the other hand, assume $\gamma_{i_+} \notin \mathrm{an}(\alpha, \beta)$. Then $i_+ \neq 0$, and there is some collider, γ_k, on π, $k \in \{1, ..., i_+\}$. We have that $\gamma_k \notin \mathrm{an}(\alpha, \beta)$ and π is closed in this collider.

On the other hand, assume that there is a weak inducing path between α and β. If $\alpha = \beta$, then $\alpha \to \beta$ which is connecting given $C \subseteq V \setminus \{\alpha, \beta\}$. Assume $\alpha \neq \beta$. If α and β are adjacent, then $\alpha \sim \beta \to \beta$ is μ-connecting given $C \subseteq V \setminus \{\alpha, \beta\}$. Consider the weak inducing path,

$$\alpha \sim \gamma_1 \sim \dots \gamma_j \sim \beta = \gamma_{j+1}.$$

Let k be the maximal number in the set $\{1, \dots, j\}$ such that there is a walk between α and γ_k with all colliders in $\mathrm{an}(C)$, no noncolliders in C, and which has a neck at γ_k. We see that $\gamma_1 \neq \beta$ fits this description, i.e., k is well-defined. Let ω be the walk from α to γ_k. If $\gamma_k \in \mathrm{an}(C)$, then the composition of ω with $\gamma_k \sim \gamma_{k+1}$ gives either a new such walk (if the edge is blunt) and by maximality $\gamma_{k+1} = \beta$, or if the edge is directed then also $\gamma_{k+1} = \beta$ (the weak inducing path is a collider path), and composing either walk with $\beta \to \beta$ gives a connecting walk. Assume instead that $\gamma_k \notin \mathrm{an}(C)$, and consider again ω. There is a directed path from γ_k to α or to β. Let $\bar{\pi}$ denote the subpath from γ_k to the first instance of either α or β. If α occurs first, we compose $\bar{\pi}^{-1}$ with $\gamma_k \sim \gamma_{k+1}$ and argue as in the case of $\gamma_j \in \mathrm{an}(C)$ above. In β occurs first, ω composed with $\bar{\pi}$ is connecting. □

We say that β is *inseparable* from α if there is no $C \subseteq V \setminus \{\alpha\}$ such that β is μ-separated from α by C.

Example 21. This example is meant to illustrate that the separation models encoded by cDGs are a strict superset of those encoded by DGs. Consider the cDG in Figure 3. We can ask if there is a DG on the same node set that encodes the same separations. Using Proposition 16, we see that any such DG must include edges $\alpha \to \beta$ and $\beta \to \gamma$, and that it cannot include $\alpha \to \gamma$ nor $\gamma \to \alpha$. Then it must include the edge $\gamma \to \beta$ as otherwise γ would be μ-separated from α given $\{\beta\}$. However, this is a contradiction as β would then be inseparable from γ.

Example 22. Mogensen and Hansen (2020) use μ-separation in *directed mixed graphs* (DMGs) to represent local independence models. Between every pair of nodes, α and β, in a DMG there is a subset of the edges $\{\alpha \to \beta, \alpha \leftarrow \beta, \alpha \leftrightarrow \beta\}$. We can also ask if the separation model represented by the cDG in Figure 3 can

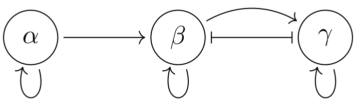


Figure 3: A cDG, $\mathcal{D}$, on nodes $V = \{\alpha, \beta, \gamma\}$ such that the separation model $\mathcal{I}(\mathcal{D})$ cannot be represented by a DMG on nodes V. See Example 22.

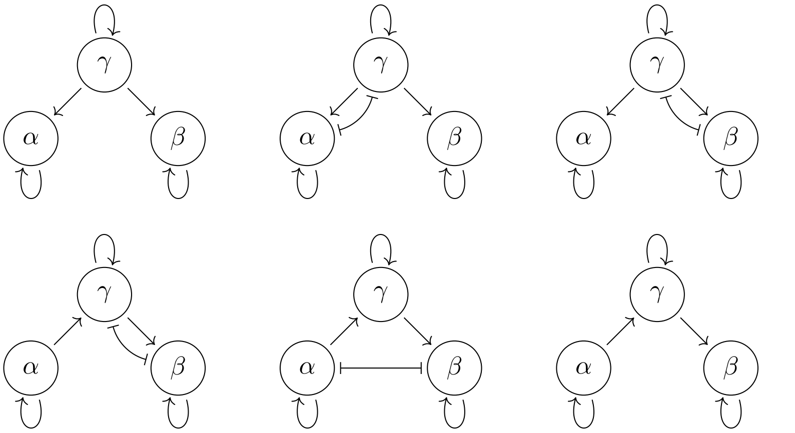


Figure 4: First row: an equivalence class illustrating that a greatest element need not exist. Second row: the left and middle graphs are Markov equivalent. The graph on the right is the largest graph which is a subgraph of both of them, and this graph is not Markov equivalent, i.e., the Markov equivalence class of the left (and middle) graph does not have a least element. Theorem 32 gives a characterization of Markov equivalence of cDGs.

be described by a DMG, i.e., allowing directed and bidirected edges. The node γ is separable from α and vice versa, i.e., there can be no edge between the two in the DMG. The node γ is not separated from α given $\{\beta\}$, and therefore β must be a collider on a path between the two. However, then there is a head at β on an edge from γ and therefore β is inseparable from γ which is a contradiction.

DGs constitute a subclass of cDGs and within the class of DGs every Markov equivalence class is a singleton, i.e., two DGs are Markov equivalent if and only if they are equal.

Proposition 23 (Mogensen and Hansen (2020)). Let $\mathcal{D}_1 = (V, E_1)$ and $\mathcal{D}_2 = (V, E_2)$ be DGs. Then $\mathcal{D}_1 \in [\mathcal{D}_2]$ if and only if $\mathcal{D}_1 = \mathcal{D}_2$.

Proposition 23 does not hold in general when $\mathcal{D}_1$ and $\mathcal{D}_2$ are cDGs. As an example, consider a graph on nodes $\{\alpha, \beta\}$ such that $\alpha \to \beta$ and $\beta \to \alpha$. This graph is Markov equivalent with the graph where $\alpha \mapsto \beta$ is added. The next result is an immediate consequence of Proposition 16 and shows that Markov equivalent cDGs always have the same directed edges.

Corollary 24. Let $\mathcal{D}_1 = (V, E_1)$ and $\mathcal{D}_1 = (V, E_1)$ be cDGs. If they are Markov equivalent, then for all $\alpha, \beta \in V$ it holds that $\alpha \to_{\mathcal{D}_1} \beta$ if and only if $\alpha \to_{\mathcal{D}_2} \beta$.

While the local independence graph (a cDG) is not in general identifiable from its independence model, Proposition 23 shows that DGs are identifiable from their induced independence models (within the class of DGs).

Mogensen and Hansen (2020) use DMGs to represent marginalized local independence models and show the below result on Markov equivalence. The class of cDGs represent local independences allowing for correlation in the error process and it is natural to ask if the same result on Markov equivalence holds in this class of graphs. The answer is in the negative as illustrated by Example 26. For graphs $\mathcal{D}_1 = (V, E_1)$ and $\mathcal{D}_2 = (V, E_2)$, we write $\mathcal{D}_1 \subseteq \mathcal{D}_2$ if $E_1 \subseteq E_2$. We say that a graph, $\mathcal{D}$, is a *greatest* element of its equivalence class, $[\mathcal{D}]$, if it is a supergraph of all members of the class, i.e., $\tilde{\mathcal{D}} \subseteq \mathcal{D}$ for all $\tilde{\mathcal{D}} \in [\mathcal{D}]$. We say that $\mathcal{D}$ is a *least* element if $\mathcal{D} \subseteq \tilde{\mathcal{D}}$ for all $\tilde{\mathcal{D}} \in [\mathcal{D}]$.

Theorem 25 (Mogensen and Hansen (2020))**.** Let $\mathcal{G}$ be a directed mixed graph. Then $[\mathcal{G}]$ has a greatest element (within the class of DMGs), i.e., there exists $\bar{\mathcal{G}} \in [\mathcal{G}]$ such that $\bar{\mathcal{G}}$ is a supergraph of all Markov equivalent DMGs.

Example 26. Consider the graph to the left on the first row of Figure 4. The edge $\alpha \mapsto\!\!\!\!\!\leftarrow \gamma$ can be added Markov equivalently and the edge $\beta \mapsto\!\!\!\!\!\leftarrow \gamma$ can be added Markov equivalently (middle and right graphs), but they cannot both be added Markov equivalently at the same time. This shows that the equivalence class of this graph does not contain a greatest element. Figure 4 also gives an example showing that an equivalence class of cDGs does not necessarily contain a least element.

4.1 A characterization of Markov equivalence of cDGs

The central result of this section is a characterization of Markov equivalence of cDGs. We define *collider equivalence* of graphs as a first step in stating this result.

Definition 27. Let $\mathcal{D}_1 = (V, E_1)$, $\mathcal{D}_2 = (V, E_2)$ be cDGs with the same directed edges, and let ω be a collider path in $\mathcal{D}_1$,

$$\alpha \sim \gamma_1 \sim \ldots \sim \gamma_{k_1} \sim \beta.$$

We say that ω is *covered* in $\mathcal{D}_2$ if there exists a collider path in $\mathcal{D}_2$

$$\alpha \sim \bar{\gamma}_1 \sim \ldots \sim \bar{\gamma}_{k_2} \sim \beta$$

such that for each $\bar{\gamma}_j$ we have $\bar{\gamma}_j \in \text{an}(\alpha, \beta)$ or $\bar{\gamma}_j \in \cup_i \text{an}(\gamma_i)$.

In the above definition $\{\gamma_j\}$ and $\{\bar{\gamma}_j\}$ may be the empty set, corresponding to α and β being adjacent, $\alpha \sim \beta$.

Definition 28 (Collider equivalence)**.** Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be cDGs on the same node set and with the same directed edges. We say that $\mathcal{D}_1$ and $\mathcal{D}_2$ are *collider equivalent* if every collider path in $\mathcal{D}_1$ is covered in $\mathcal{D}_2$ and every collider path in $\mathcal{D}_2$ is covered in $\mathcal{D}_1$.

In the above definition, it is crucial that we use the convention that every node is an ancestor of itself, $\gamma \in \text{an}(\gamma)$ for all $\gamma \in V$. Otherwise, a graph would not necessarily be collider equivalent with itself. With this convention, it follows immediately that every cDG is collider equivalent with itself. One should also note that a single edge, $\alpha \sim \beta$, constitutes a collider path between α and β and that a single edge covers any other collider path as it has no nonendpoint nodes.

We do not need to consider walks in the above definitions (only paths) as we assume that all loops are included and therefore all nodes are collider connected to themselves by assumption. If there is a collider walk between α and β, then there is also a collider path. Furthermore, if a collider walk is covered by a collider walk, then it is also covered by a collider path, and we see that one would obtain an equivalent definition by using collider walks instead of collider paths in Definitions 27 and 28.

If $\mathcal{D}$ is a cDG such that $\alpha \to_{\mathcal{D}} \alpha$ for all $\alpha \in V$, then we say that $\mathcal{D}$ *contains every loop*. We say that cDGs $\mathcal{D}_1 = (V, E_1)$ and $\mathcal{D}_2 = (V, E_2)$ *have the same directed edges* if for all $\alpha, \beta \in V$ it holds that $\alpha \to_{\mathcal{D}_1} \beta$ if and only if $\alpha \to_{\mathcal{D}_2} \beta$.

Remark 29. Collider equivalence implies that two graphs have the same weak inducing paths in the following sense. Assume ω is a weak inducing path between α and β in $\mathcal{D}_1$, and that $\mathcal{D}_1$ and $\mathcal{D}_2$ are collider equivalent and have the same directed edges. In $\mathcal{D}_2$, there exists a collider path, $\bar{\omega}$, such that every nonendpoint node is an ancestor of a node on ω, i.e, an ancestor of $\{\alpha, \beta\}$ using the fact that ω is a weak inducing path. This means that $\bar{\omega}$ is a weak inducing path in $\mathcal{D}_2$.

Lemma 30. Let $\mathcal{D}_1 = (V, E_1)$, $\mathcal{D}_2 = (V, E_2)$ be cDGs that contain every loop. If $\mathcal{D}_1$ and $\mathcal{D}_2$ are not collider equivalent, then they are not Markov equivalent.

Proof. Assume that $\mathcal{D}_1$ and $\mathcal{D}_2$ are not collider equivalent. Assume first that there exists $\alpha, \beta \in V$ such that there is a collider path between α and β in $\mathcal{D}_2$,

$$\alpha \sim \bar{\gamma}_1 \sim \ldots \sim \bar{\gamma}_k \sim \beta$$

which is not covered in $\mathcal{D}_1$. Both graphs contain all loops, so $\alpha \neq \beta$. This means that on every collider path between α and β in $\mathcal{D}_1$, there exists a collider γ such that $\gamma \notin \text{an}(\alpha, \beta)$ and $\gamma \notin \cup_j \text{an}(\bar{\gamma}_j)$. Now consider the set $D = \text{an}(\alpha, \beta) \cup [\cup_j \text{an}(\bar{\gamma}_j)] \setminus \{\alpha, \beta\}$. Note that β is not μ-separated from α given D in $\mathcal{D}_2$ as $\beta \to_{\mathcal{D}_2} \beta$, and we will argue that β is μ-separated from α given D in $\mathcal{D}_1$ showing that these graphs are not Markov equivalent. Consider any walk between α and β in $\mathcal{D}_1$. It suffices to consider paths between α and β composed with the edge $\beta \to \beta$ (as $\beta \notin D$). Assume first that it is a collider path. If it is open, then every nonendpoint node is an ancestor of α, β, or $\bar{\gamma}_j$ for some j, which is a contradiction. Assume instead that there exists a noncollider (different from

α and β) on the path. There must also exist a collider (otherwise it is closed), and the collider is a descendant of the noncollider. The collider is either closed, or it is an ancestor of either $\{\alpha, \beta\}$ or of $\cup_i \bar{\gamma}_i$. In the latter case, the path is closed in the noncollider. □

Proposition 31. Assume $\alpha, \beta \notin C$. If ω is a collider path from α to β given C such that every collider is in $\mathrm{an}(\{\alpha,\beta\} \cup C)$, then there is a walk between α and β such that no noncollider is in C and every collider is in $\mathrm{an}(C)$.

A similar and more general result was shown by Richardson (2003) in the case of m-separation in directed mixed graphs.

Proof. In the original graph, $\mathcal{D}$, we add directed edges such that every node in C is a parent of α. Now the path is a weak inducing path, in this larger graph $\mathcal{D}^+$. Using Proposition 20, we can find a μ-connecting walk from α to β given C in $\mathcal{D}^+$, and therefore a walk from α to β such that every noncollider is not in C and every collider is in $\mathrm{an}(C)$. This walk is also in $\mathcal{D}$ as it cannot contain an edge with a tail at $\gamma \in C$. In $\mathcal{D}$, we see that every collider is still in $\mathrm{an}(C)$ and the result follows. □

Theorem 32 (Markov equivalence of cDGs)**.** Let $\mathcal{D}_1 = (V, E_1)$ and $\mathcal{D}_2 = (V, E_2)$ be cDGs that contain every loop. The graphs $\mathcal{D}_1$ and $\mathcal{D}_2$ are Markov equivalent if and only if they have the same directed edges and are collider equivalent.

We give a direct proof of this theorem. One can also use the augmentation criterion to show this result.

Proof. Assume first that $\mathcal{D}_1$ and $\mathcal{D}_2$ have the same directed edges and are collider equivalent. Then $\mathrm{an}_{\mathcal{D}_1}(C) = \mathrm{an}_{\mathcal{D}_2}(C)$ for all $C \subseteq V$ so we will omit the subscript and write simply $\mathrm{an}(C)$. Let ω denote a μ-connecting walk from α to β given C in $\mathcal{D}_1$. We will argue that we can also find a μ-connecting walk in $\mathcal{D}_2$. We say that a nontrivial subwalk of ω is a *maximal collider segment* if all its nonendpoint nodes are colliders on ω, its endpoint nodes are not colliders, and it contains at least one blunt edge (note that on a general walk this should be read as *instances* of these nodes as nodes may be repeated on a walk). We can partition ω into a sequence of subwalks such that every subwalk is either a maximal collider segment, or a subwalk consisting of directed edges only. We note that maximal collider segment may be adjacent, i.e., share an endpoint. Every segment of ω that consists of directed edges only is also present in $\mathcal{D}_2$. Consider a maximal collider segment. This is necessarily a collider walk in $\mathcal{D}_1$. Then there exists a collider path in $\mathcal{D}_1$, and therefore a covering collider path in $\mathcal{D}_2$ using collider equivalence. Denote this path by ρ and assume that it is between δ and ε. It follows that $\delta, \varepsilon \notin C$ as they are noncolliders on ω, or equal to α or β. If they are equal to the final β, the final edge must point towards β and therefore the segment is directed. We will now find an open (given C) walk between δ and ε using ρ. We know that ρ is a collider path and that every nonendpoint node on ρ is an ancestor of $\{\alpha, \beta\}$ or of a collider in the original maximal collider segment, and therefore to C. It follows from Proposition 31 that we can find an m-connecting walk between δ and ε. We create a walk from α to β in $\mathcal{D}_2$ by simply substituting each maximal collider segment with the corresponding open walk. This walk is open in any node which is not an endpoint of a maximal collider segment. If an endpoint of maximal collider node changes collider status on this new walk, then it must be a noncollider on ω and a parent of a node in $\mathrm{an}(C)$, i.e., also in $\mathrm{an}(C)$ itself. Finally, we note that the last segment (into β) is not a maximal collider segment and therefore still has a head into β.

On the other hand, if they do not have the same directed edges, it follows from Proposition 16 that they are not Markov equivalent. If they are not collider equivalent, it follows from Lemma 30 that they are not Markov equivalent. □

We say that α and β are *adjacent* in the graph $\mathcal{D}$ if $\alpha \sim_{\mathcal{D}} \beta$. In the case of *directed acyclic graphs* it holds that Markov equivalent graphs have the same adjacencies, however, this is not true in the case of cDGs, and in fact, it is also not true among maximal cDGs (Definition 15) as seen in Figure 7.

Proposition 33. Let $\mathcal{D} = (V, E)$ be a cDG, and let $\alpha, \beta, \gamma \in V$. Let e denote the blunt edge between α and β. If α and β are connected by a weak inducing path consisting of blunt edges only, then $\mathcal{D} + e \in [\mathcal{D}]$.

Proof. Let ω be a μ-connecting walk between δ and ε in $\mathcal{D}+e$. In $\mathcal{D}$, consider the weak inducing path between α and β that consists of blunt edges only. Using a proof similar to that of Proposition 20, one can show that there exists an open walk between α and β in $\mathcal{D}$ which has necks at both end. This means that replacing $\alpha \mapsto\!\!\!\!\dashv \beta$ with this walk gives a μ-connecting walk in $\mathcal{D}$. □

4.2 Markov equivalent permutation of nodes

The example in Figure 7 shows a characteristic of some Markov equivalent cDGs. In the example, one can obtain one graph from the other by a permutation of the endpoints of blunt edges within the set $\{\gamma, \delta\}$.

Definition 34 (Cyclic component)**.** We say that $S \subseteq V$ is a cyclic component if for every $(\alpha, \beta) \in S \times S$, it holds that $\alpha \in \mathrm{an}(\beta)$.

Note that if two sets of nodes have the same descendants (using the convention that every node is a descendant of itself), they are necessarily contained in the same cyclic component as the graphs are assumed to have all loops. The following is a formal definition of a *permutation graph* as illustrated in the example of Figure 7.

Definition 35 (Permutation graph)**.** Let $\mathcal{D} = (V, E)$ and let ρ be a permutation of the node set V. We define $\mathcal{P}_\rho(\mathcal{D})$ as the cDG on nodes V such that

$$\alpha \rightarrow_{\mathcal{P}_\rho(\mathcal{D})} \beta \quad \text{if } \alpha \rightarrow_{\mathcal{D}} \beta, \tag{4}$$

$$\rho(\alpha) \mapsto\!\!\!\!\dashv_{\mathcal{P}_\rho(\mathcal{D})} \rho(\beta) \text{ if } \alpha \mapsto\!\!\!\!\dashv_{\mathcal{D}} \beta. \tag{5}$$

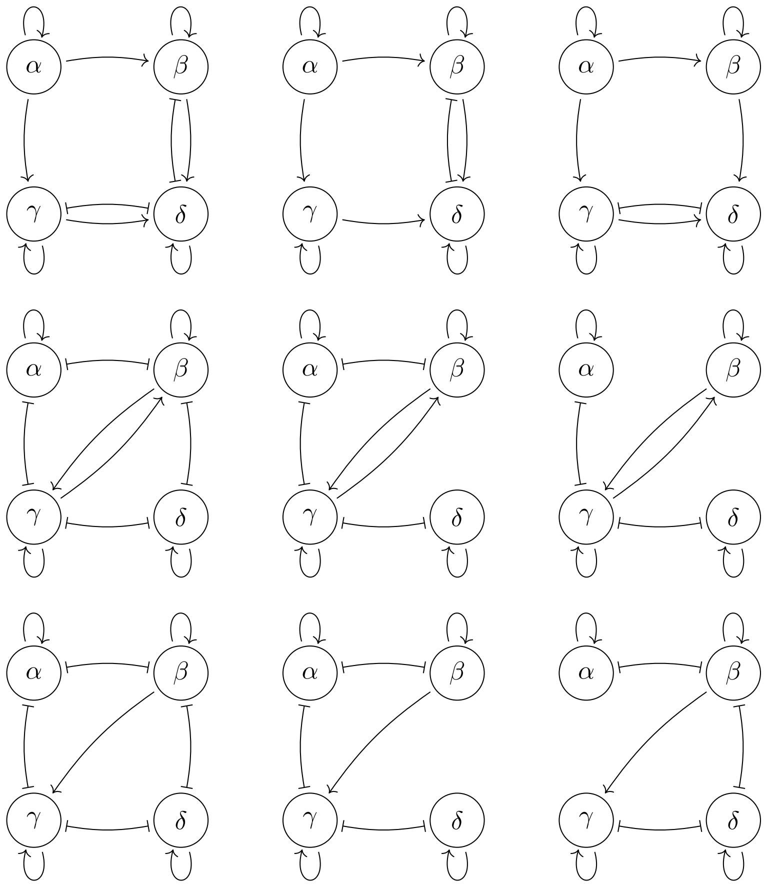


Figure 5: Markov equivalence in cDGs. First row: these are three members of a Markov equivalence class of size 21. The only restriction on 2^5 combinations of blunt edges (all but $\beta \vdash\!\dashv \gamma$ can be present) is the fact that we cannot have both $\alpha \vdash\!\dashv \beta$ and $\alpha \vdash\!\dashv \gamma$ present and that either (α, δ), (γ, δ), or (β, δ) are spouses as otherwise there would not be a weak inducing path between α and δ. Second row: these graphs are Markov equivalent. The collider path $\alpha \vdash\!\dashv \beta \vdash\!\dashv \delta$ in the first graph is 'covered' in the two others by the walk $\alpha \vdash\!\dashv \gamma \vdash\!\dashv \delta$ as $\gamma \in \mathrm{an}(\beta)$. The edge $\beta \vdash\!\dashv \delta$ is 'covered' by the inducing path $\delta \vdash\!\dashv \gamma \leftarrow \beta$ in the middle and right graphs of the row. The equivalence class of these graphs has cardinality 16 which is every combination of blunt edges (2^5) that makes the graph connected. Third row: the first graph is not collider equivalent with the following two: the collider path $\alpha \vdash\!\dashv \beta \vdash\!\dashv \delta$ is not covered by any collider path in the second graph. The collider path $\alpha \vdash\!\dashv \gamma$ is not covered by any collider path in the third.

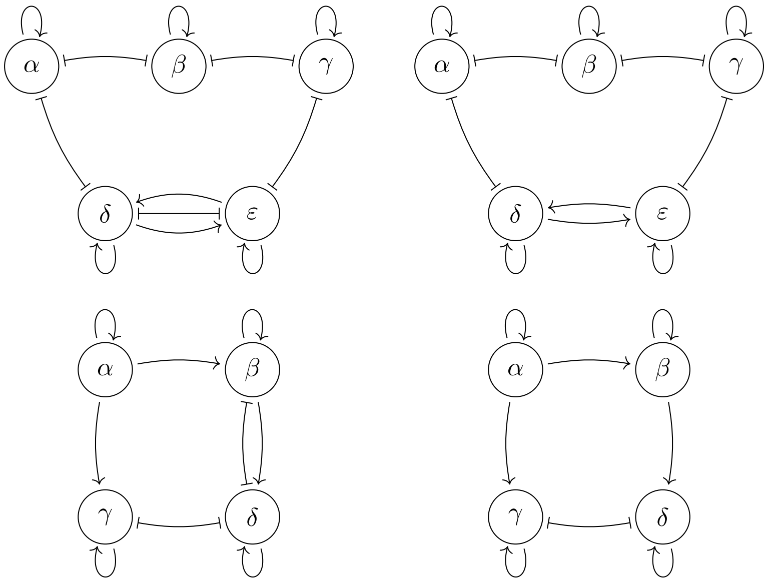


Figure 6: Examples of Markov equivalence in cDGs. First row: the two graphs have the same weak inducing paths, but are not Markov equivalent as the collider path $\alpha \vdash\!\dashv \delta \vdash\!\dashv \epsilon \vdash\!\dashv \gamma$ is not covered in the right graph. Second row: the two graphs are such that two nodes are collider connected in one if and only if they are collider connected in the other graph, however, they are not Markov equivalent.

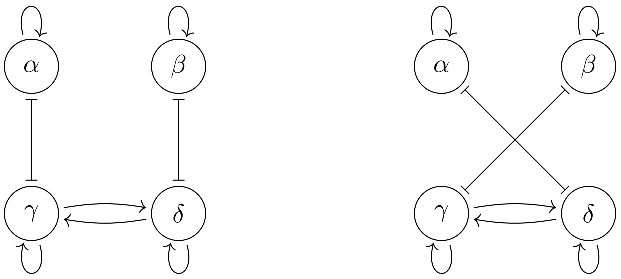


Figure 7: The two cDGs constitute a Markov equivalence class, and they are both seen to be maximal. However, they do not have the same adjacencies. A similar phenomenon can occur in DGs (without loops) under d-separation (Richardson, 1996a, 1997).

Proposition 36. Let $\mathcal{D} = (V, E)$ be a cDG and let $S \subseteq V$. Let ρ be a permutation of V such that $\rho(\alpha) = \alpha$ for all $\alpha \notin S$. If $\beta \to \gamma$ and $\text{pa}(\beta) = \text{pa}(\gamma)$ for all $\beta, \gamma \in S$, then $\mathcal{P}_\rho(\mathcal{D}) \in [\mathcal{D}]$.

Note that the condition that $\beta \to_\mathcal{D} \gamma$ for all $\beta, \gamma \in S$ implies that S is a cyclic component.

Proof. The graphs $\mathcal{D}$ and $\mathcal{P}_\rho(\mathcal{D})$ have the same directed edges so it suffices to show that they are collider equivalent (Theorem 32). Any permutation can be written as a composition of transpositions so it suffices to prove the result for a permutation, ρ, such that $\rho(\alpha) = \beta$, $\rho(\beta) = \alpha$, and $\rho(\gamma) = \gamma$ for all $\gamma \neq \alpha, \beta$. Let π be a collider path in $\mathcal{D}$,

$$\gamma \sim \delta_1 \sim \ldots \sim \delta_k \sim \varepsilon.$$

If $\gamma, \varepsilon \notin \{\alpha, \beta\}$, then the path

$$\gamma \sim \rho(\delta_1) \sim \ldots \sim \rho(\delta_k) \sim \varepsilon$$

is in the permutation graph and is covering, using that α and β have the same parent set. If, e.g., $\gamma = \alpha \mapsto \delta_1$ on the original path, then we can substitute this for $\alpha \to \beta \mapsto \gamma$ to obtain a covering walk in the permutation graph. Similar arguments in each case show that any collider path in $\mathcal{D}$ is covered in the permutation graph. Repeating the above argument starting from the permutation graph and using ρ^{-1} shows that the two graphs are Markov equivalent. □

Figure 7 shows two graphs that are Markov equivalent by Proposition 36. In some graphs one can find permutations, not fulfilling the assumptions of Proposition 36, that generate Markov equivalent graphs, and this proposition is therefore not a necessary condition for Markov equivalence under permutation of blunt edges. One example is in the first row of Figure 5. The middle and right graphs are Markov equivalent and one is generated from the other by permuting β and γ.

5 Deciding Markov equivalence

In this section, we will consider the problem of deciding Markov equivalence algorithmically. That is, given two cDGs on the same node set, how can we decide if they are Markov equivalent or not? A starting point is Theorem 32. While it is computationally easy to check whether the directed edges of two cDGs are the same (quadratic in the number of nodes in their common node set), collider equivalence could be hard as there may be exponentially many paths in a cDG.

Let $\mathcal{D} = (V, E)$ be a graph, and let $A \subseteq V$. We use $\mathcal{G}_A$ to denote the subgraph $\mathcal{G}$ *induced* by A, i.e., $\mathcal{G}_A = (A, E_A)$ where E_A is the set of nodes in E that are between $\alpha \in A$ and $\beta \in A$. The *directed part* of a cDG, $\mathbb{D}(\mathcal{D}) = (V, F)$, is the DG on nodes V such that $\alpha \to_{\mathbb{D}(\mathcal{D})} \beta$ if and only if $\alpha \to_\mathcal{D} \beta$. The *blunt* *part* of a cDG, $\mathbb{U}(\mathcal{D})$, is the cDG obtained by removed all directed edges. The *blunt components* of $\mathcal{D}$ are the connected components of $\mathbb{U}(\mathcal{D})$. We say that $\mathcal{D}_1 = (V, E_1)$ and $\mathcal{D}_2 = (V, E_2)$ have *the same collider connections* if it holds for all $\alpha \in V$ and $\beta \in V$ that α and β are collider connected in $\mathcal{D}_1$ if and only if they are collider connected in $\mathcal{D}_2$. We say that a subset of nodes, A, is *ancestral* if $A = \mathrm{an}(A)$. We will throughout only consider cDGs that contain every loop.

We start from the following result which is seen to be a reformulation of the augmentation criterion.

Theorem 37. Let $\mathcal{D}_1 = (V, E_1)$ and $\mathcal{D}_2 = (V, E_2)$ be cDGs such that $\mathbb{D}(\mathcal{D}_1) = \mathbb{D}(\mathcal{D}_2)$. $\mathcal{D}_1$ and $\mathcal{D}_2$ are Markov equivalent if and only if for every ancestral set, it holds that $(\mathcal{D}_1)_A$ and $(\mathcal{D}_2)_A$ have the same collider connections.

Proof. Assume that there exists an ancestral set $A \subseteq V$ such that α and β are collider connected in $(\mathcal{D}_1)_A$, but not in $(\mathcal{D}_2)_A$. There exists a collider path in $\mathcal{D}_1$ between α and β. Any covering path in $\mathcal{D}_2$ must by definition consist of nodes in $\mathrm{an}(A) = A$ and it follows that no such path can exists. By Theorem 32, it follows that $\mathcal{D}_1$ and $\mathcal{D}_2$ are not Markov equivalent.

On the other hand, assume that for every ancestral set $A \subseteq V$ and every $\alpha, \beta \in A$, it holds that α and β are collider connected in $(\mathcal{D}_1)_A$ if and only if α and β are collider connected in $(\mathcal{D}_2)_A$. Using Theorem 32, it suffices to show that $\mathcal{D}_1$ and $\mathcal{D}_2$ are collider equivalent. Consider a collider path between α and β in $\mathcal{D}_1$, and let C denote the set of nodes on this path. This path is also a collider path in $(\mathcal{D}_1)_{\mathrm{an}(\{\alpha,\beta\}\cup C)}$ and by assumption we can find a collider path between α and β in $(\mathcal{D}_2)_{\mathrm{an}(\{\alpha,\beta\}\cup C)}$ as well. This collider path is in $\mathcal{D}_2$ as well and is covering the path in $\mathcal{D}_1$. □

The above theorem can easily be turned into an algorithm for deciding if two cDGs are Markov equivalent (Algorithm 1). However, there may be exponentially many ancestral sets in a cDG. For instance, in the case where the only directed edges are loops all subsets of V are ancestral and therefore the algorithm would need to compare collider connections in 2^n pairs of graphs where n is the number of nodes in the graphs (or $2^n - 1$, omitting the empty set).

5.1 The condensation of a cDG

Let $\mathcal{D} = (V, E)$ be a cDG. We say that $\alpha, \beta \in V$ are *strongly connected* if there exists a directed path from α to β and a directed path from β to α, allowing trivial paths. Equivalently, α and β are strongly connected if and only if $\alpha \in \mathrm{an}(\beta)$ and $\beta \in \mathrm{an}(\alpha)$. This is an equivalence relation on the node set of a cDG. The definition of strong connectivity is often used in DGs (Cormen et al., 2009). We use the straight-forward generalization to the class of cDGs in which the directed part of the cDG simply determines strong connectivity.

The *condensation* of $\mathcal{D}$ (also known as the *acyclic component graph* of $\mathcal{D}$) is the directed acyclic graph obtained by contracting each strongly connected component to a single vertex. That is, if $C_1, \ldots, C_m$ are the cyclic components of $\mathcal{D}$, then the condensation of $\mathcal{D}$ has node set $\mathbb{C} = \{C_1, \ldots, C_m\}$ and $C_i \to C_j$

if $i \neq j$ and there exists $\alpha \in C_i, \beta \in C_j$ such that $\alpha \to_{\mathcal{D}} \beta$ (Cormen et al., 2009). We denote the condensation of $\mathcal{D}$ by $\mathcal{C}(\mathcal{D})$. We also define the *completed condensation* of $\mathcal{D}$, $\bar{\mathcal{C}}(\mathcal{D})$, which is the graph on nodes $\mathbb{C} \cup \{\varnothing\}$ such that $\bar{\mathcal{C}}(\mathcal{D})_{\mathbb{C}} = \mathcal{C}(\mathcal{D})$ and such that $\varnothing$ is a parent of every other node and a child of none. The condensation and the completed condensation are both DAGs. When $\mathcal{D}$ has d directed edges that are not loops, then strongly connected components can be found in linear time, that is, $O(n+d)$ (Cormen et al., 2009).

In the following, we will be considering sets of nodes in $\mathcal{D}$, i.e., subsets of V, as well as sets of nodes in $\mathcal{C}(\mathcal{D})$, that is, subsets of $\mathbb{C}$. We write the former as capital letters, A, B, C. We write the latter as capital letters in bold font, $\mathbf{A}, \mathbf{B}, \mathbf{C}$, to emphasize that they are subsets of $\mathbb{C}$, not of V.

Proposition 38. The ancestral sets in $\mathcal{D}$ are exactly the sets of the form $\{\alpha \in C_i : C_i \in \mathbf{C}\}$ for an ancestral set, $\mathbf{C}$, in $\mathcal{C}(\mathcal{D})$.

Proof. Consider an ancestral set $A \subseteq V$. We can write this as a union of strongly connected components, $A = \bigcup C_i$. These strongly connected components must necessarily constitute an ancestral set in $\mathcal{C}(\mathcal{D})$.

On the other hand, consider an ancestral set in $\mathcal{C}(\mathcal{D})$, $\mathbf{C}$, and consider $\alpha \in A = \{\alpha \in C_i : C_i \in \mathbf{C}\}$. Assume that $\alpha \in C \in \mathbf{C}$. If β is an ancestor of α in $\mathcal{D}$, then $\beta \in \tilde{C}$ such that $\tilde{C}$ is an ancestor of C in $\mathcal{C}(\mathcal{D})$. By assumption, $\mathbf{C}$ is ancestral, so $\tilde{C} \in \mathbf{C}$ and we see that A is ancestral. □

The above proposition shows that we can consider the condensation when finding ancestral sets in a cDG. We let $\mathbb{A}(\mathcal{D})$ denote the set of ancestral sets in $\mathcal{D}$. The correctness of Algorithm 1 follows from Theorem 37 and Proposition 38. The algorithm considers ancestral sets in the condensation, however, a version using ancestral sets directly in $\mathcal{D}_1$ is of course also possible. In the algorithm, one can decide collider connectivity by noting that α and β are collider connected in a cDG, $\mathcal{D}$, if and only if there exists a blunt component, $\mathcal{U} = (U, E_U)$, such that $\alpha \in \mathrm{pa}_{\mathcal{D}}(U)$ and $\beta \in \mathrm{pa}_{\mathcal{D}}(U)$, using that the graphs contain every loop.

Algorithm 1 Markov equivalence

```
Require: cDGs, $D_1 = (V, E_1), D_2 = (V, E_2)$
  if $\mathbb{D}(\mathcal{D}_1) \neq \mathbb{D}(\mathcal{D}_2)$ then
    return FALSE
  end if
  for $\mathbf{A} \in \mathbb{A}(\mathcal{C}(\mathcal{D}_1))$ do
    Define $A = \{\gamma \in C_i : C_i \in \mathbf{A}\}$
    if $(\mathcal{D}_1)_A$ and $(\mathcal{D}_2)_A$ do not have the same collider connections then
      return FALSE
    end if
  end for
  return TRUE
```

5.2 Virtual collider tripaths

This section describes a graphical structure that we will call *virtual collider tripaths.* We will use these to give a necessary condition for Markov equivalence.

Definition 39 (Virtual collider tripath)**.** Let $\alpha, \beta \in V$ and let C be a node in $\bar{\mathcal{C}}(\mathcal{D})$, i.e., C is a cyclic component or the empty set. We say that (α, β, C) is a *virtual collider tripath* if there exists a collider path $\alpha \sim \gamma_1 \sim \dots \gamma_m \sim \beta$ such that $\gamma_i \in \mathrm{an}(\{\alpha, \beta\} \cup C)$ for all $i = 1, \dots, m$.

Richardson (1996b) described *virtual adjacencies* in DGs equipped with d-separation. Those are structures that in terms of separation act as adjacencies. The idea behind virtual collider tripaths is essentially the same; for a fixed pair of nodes, α and β, a virtual collider tripath, (α, β, C), acts as if there exists $\gamma \in C$ such that $\alpha \sim \gamma \sim \beta$ is a collider walk. Note also that if α and β are adjacent, then (α, β, C) is virtual collider tripath for any cyclic component C. Finally, note that there are no restrictions on whether or not α, β, or both are elements in the set $C \subseteq V$.

Definition 40 (Maximal virtual collider tripath)**.** We say that a virtual collider tripath, (α, β, C), is *maximal* if there is no $\tilde{C} \neq C$ such that $(\alpha, \beta, \tilde{C})$ is a virtual collider tripath and $\tilde{C}$ is an ancestor of C in $\bar{\mathcal{C}}(\mathcal{D})$.

We say that two cDGs have the same (maximal) virtual collider tripaths if it holds that (α, β, C) is a (maximal) virtual collider tripath in $\mathcal{D}_1$ if and only if (α, β, C) is a (maximal) virtual collider tripath in $\mathcal{D}_2$.

Proposition 41. If (α, β, C) is not a virtual collider tripath, then β and α are m-separated by $\mathrm{an}(\{\alpha, \beta\} \cup C) \setminus \{\alpha, \beta\}$.

Proof. The contraposition follows from the definition of a virtual collider tripath. Assume that ω is an m-connecting walk between α and β given $\mathrm{an}(\{\alpha, \beta\} \cup C) \setminus \{\alpha, \beta\}$. If it is a single edge, then (α, β, C) is a virtual collider tripath for any C. Assume that it has length at least two. If there is a noncollider, δ, on ω, then δ must be an ancestor of $\{\alpha, \beta\}$ or of a collider. In the former case, ω is closed as in δ is in the condition set. In the latter case, either ω is closed in the collider or in δ. Assume therefore that ω is a collider walk. We can reduce ω to a path and we see from the definition that (α, β, C) is a virtual collider tripath. □

The next theorem gives a necessary condition for Markov equivalence of cDGs.

Theorem 42. Let $\mathcal{D}_1 = (V, E_1)$ and $\mathcal{D}_2 = (V, E_2)$ be cDGs. If they are Markov equivalent, then they have the same directed edges and the same maximal virtual collider tripaths.

Proof. We show this by contraposition. If α is a parent of β in $\mathcal{D}_1$, but not in $\mathcal{D}_2$, then it follows from Corollary 24 that they are not Markov equivalent.

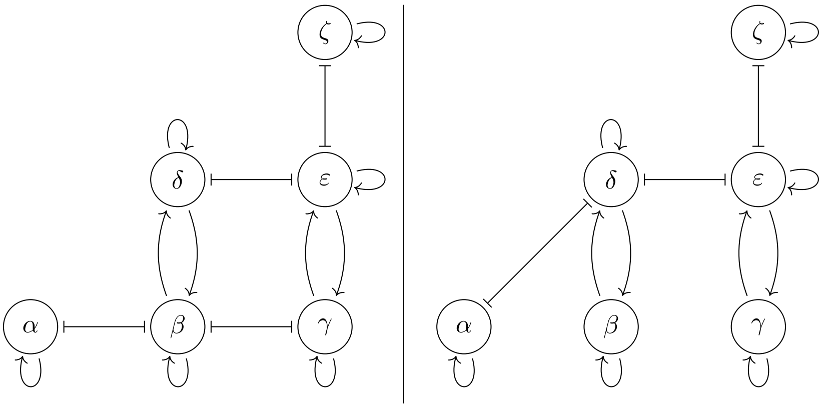


Figure 8: These cDGs on nodes $\{\alpha,\beta,\gamma,\delta,\varepsilon,\zeta\}$ have the same maximal virtual collider tripaths, however, disagree on whether ζ is μ-separated from α by $\{\beta,\gamma,\delta,\varepsilon\}$.

Assume instead that $\mathcal{D}_1$ and $\mathcal{D}_2$ have the same directed edges, and that (α,β,C) is a maximal virtual collider tripath in $\mathcal{D}_1$, but not in $\mathcal{D}_2$. Then it follows that $\alpha \neq \beta$ as we assume all directed loops to be present in both graphs. There are two cases; either (α,β,C) is not a virtual collider tripath in $\mathcal{D}_2$, or it is not maximal. In the first case, β is μ-separated from α by $\text{an}(\{\alpha,\beta\} \cup C) \setminus \{\alpha,\beta\}$ (Proposition 41) which is seen to not be the case in $\mathcal{D}_1$. In the second case, in $\mathcal{D}_2$ there is a virtual collider tripath $(\alpha,\beta,\tilde{C})$ such that $\tilde{C} \to C$ in $\bar{\mathcal{C}}(\mathcal{D}_1)$ (note that $\bar{\mathcal{C}}(\mathcal{D}_1) = \bar{\mathcal{C}}(\mathcal{D}_2)$) and $(\alpha,\beta,\tilde{C})$ is not a virtual collider tripath in $\mathcal{D}_1$. Repeating the above argument, we see that $\mathcal{D}_1$ and $\mathcal{D}_2$ are not Markov equivalent in this case either. □

The example in Figure 8 shows that having the same directed edges and the same maximal virtual collider tripaths is not a sufficient condition for Markov equivalence.

5.3 Complexity of deciding Markov equivalence

We have given two characterizations of Markov equivalence of cDGs and argued that they both use exponentially many conditions in the worst case. In this section, we prove that this, most likely, cannot be circumvented.

coNP is the class of decision problems for which a no-instance can be verified using a polynomial-length counterexample in polynomial time and a problem is in coNP if and only if its complement is in NP. If a problem is as hard as any problem in coNP, then we say that the problem is coNP-*hard*. If a problem is coNP-hard and also in coNP, we say that it is coNP-*complete* (Garey and Johnson, 1979; Sipser, 2013). Various inference problems in graphical models are known to be computationally hard (Meek, 2001; Chickering et al., 2004; Chandrasekaran et al., 2008; Koller and Friedman, 2009). On the other hand, there exists polynomial-time algorithms for deciding Markov equivalence in several classes of graphs, e.g., maximal ancestral graphs (Ali et al., 2009) and DGs under d-separation (Richardson, 1997). This is different in cDGs under μ-separation.

Theorem 43. Deciding Markov equivalence of cDGs is coNP-complete.

The complexity result implies that, unless P = coNP (which is commonly believed to not be the case), one cannot find a characterization of Markov equivalence of cDGs which can be verified in polynomial time in the size of the graph as this would allow us to decide Markov equivalence of two cDGs.

Proof. We first show that deciding Markov equivalence is in coNP. This is clear as given two graphs that are not Markov equivalence and a certificate indicating sets A, B, C such that we have separation in one but not in the other, we can use Proposition 17 to verify this no-instance in polynomial time.

In order to show that deciding Markov equivalence is coNP-hard, we use a reduction similar to one by Böhler et al. (2012) who study complexity of deciding equivalence of Boolean circuits, see in particular the proof of their Lemma 4.3. Consider Boolean variables $x_1, \ldots, x_n$. We say that x_l and $\neg x_l$ are *literals.* A Boolean formula is in *disjunctive normal form* (DNF) if it is a disjunction of conjuctions of literals. It is a 3DNF, if each conjunction has at most three literals. The 3DNF tautology is the problem of deciding if a 3DNF is satisfied for all inputs and this problem is known to be coNP-hard. We reduce 3DNF tautology to the problem of deciding Markov equivalence. Let H be a 3DNF formula on variables $x_1, \ldots, x_n$ consisting of literals

$$H = (z_1^1 \wedge z_2^1 \wedge z_3^1) \vee \ldots \vee (z_1^N \wedge z_2^N \wedge z_3^N)$$

such that z_i^j equals x_l or $\neg x_l$ for some $l = 1, \ldots, n$. In the former case, we say that z_i^j is a *positive* literal, and in the latter that z_i^j is a *negative* literal. We say that a conjunction, e.g., $z_1^j \wedge z_2^j \wedge z_3^j$, is a *term.* In the following, we will define graphs in which the nodes corresponds to literals, variables, and negated variables in this problem. We will use Greek alphabet letters for the nodes. Now define

$$V^- = \{\zeta_i^j\} \cup \{\chi_l, \upsilon_l\},$$

such that ζ_i^j corresponds to z_i^j, χ_l to x_l, and υ_l to the negation of x_l. We also define

$$V = \{\alpha, \beta\} \cup V^- \cup \{\gamma_\delta : \delta \in V^-\}.$$

We now construct a cDG on nodes V with the following edge set. Every node has a directed loop. Furthermore, for $\delta \in V^-$,

$$\alpha \rightarrow \gamma_\delta \overset{\leftarrow}{\rightarrow} \delta.$$

For every term (analogously if the term has fewer than three literals),

$$\alpha \to \zeta_1^j \mapsfromto \zeta_2^j \mapsfromto \zeta_3^j \mapsfromto x_1$$

and also $\zeta_3^j \mapsfromto \upsilon_1$. Furthermore, $\chi_l, \upsilon_l \mapsfromto \chi_{l+1}, \upsilon_{l+1}$ and $\chi_n, \upsilon_n \mapsfromto \beta$ (in the sense that there is a blunt edge between any pair of nodes on opposite sides of $\mapsfromto$). We let also $\chi_1 \mapsfromto \upsilon_1$. Finally, $\chi_l \overset{\rightarrow}{\leftarrow} \zeta_i^j$ if and only if z_i^j is a positive literal of the variable x_l and $\upsilon_l \overset{\rightarrow}{\leftarrow} \zeta_i^j$ if and only if z_i^j is a negative literal of the variable x_l. We let $\mathcal{G}$ denote the cDG on nodes V and with edges as described above. We define also $\mathcal{G}^+$ by adding edges $\alpha \mapsfromto \chi_1, \upsilon_1$ to $\mathcal{G}$.

We now argue that H is a tautology (that is, true for all inputs) if and only if $\mathcal{G}$ and $\mathcal{G}^+$ are Markov equivalent. Assume that H is a tautology. To argue that $\mathcal{G}$ and $\mathcal{G}^+$ are Markov equivalent it suffices to show that every collider path of $\mathcal{G}^+$ is covered in $\mathcal{G}$ (Theorem 32). Every collider path in $\mathcal{G}^+$ which is not in $\mathcal{G}$ either contains the subpath $\chi_1 \mapsfromto \alpha \mapsfromto \upsilon_1$ or is of the below form. If it contains $\chi_1 \mapsfromto \alpha \mapsfromto \upsilon_1$, then we can substitute this for $\chi_1 \mapsfromto \upsilon_1$ and obtain a covering path in $\mathcal{G}$. Assume instead a collider path of the following form,

$$\alpha \mapsfromto \varepsilon_1 \mapsfromto \ldots \sim \varepsilon_{k+1}.$$

If $\varepsilon_{k+1} \neq \beta$, then this is covered by $\alpha \to \gamma_{\varepsilon_{k+1}} \overset{\leftarrow}{\rightarrow} \varepsilon_{k+1}$, or by $\alpha \to \varepsilon_{k+1}$. Assume instead that $\varepsilon_{k+1} = \beta$. In this case, for all $i = 1 \ldots, n$ either $\chi_i \in \{\varepsilon_1 \ldots, \varepsilon_k\}$ or $\upsilon_i \in \{\varepsilon_1 \ldots, \varepsilon_k\}$. Consider now the following assignment of truth values to the variables: $x_l = 1$ if and only if $\chi_l \in \{\varepsilon_1 \ldots, \varepsilon_k\}$. By assumption, H is a tautology, so there is a term which equals 1 for this assignment, say the j'th (without loss of generality assuming the the j'th term contains three literals),

$$z_1^j \wedge z_2^j \wedge z_3^j.$$

If z_i^j is a positive literal, then it must correspond to a x_l such that $\chi_l \in \{\varepsilon_1 \ldots, \varepsilon_k\}$, and then in $\mathcal{G}$, ζ_i^j is a parent of $\chi_l \in \{\varepsilon_1 \ldots, \varepsilon_k\}$. If it is a negative literal, then it must correspond to x_l such that $\chi_l \notin \{\varepsilon_1 \ldots, \varepsilon_k\}$. Then $\upsilon_l \in \{\varepsilon_1 \ldots, \varepsilon_k\}$, and therefore ζ_i^j is a parent of $\{\varepsilon_1 \ldots, \varepsilon_k\}$. This means that the walk

$$\alpha \to \zeta_1^j \mapsfromto \zeta_2^j \mapsfromto \zeta_3^j \mapsfromto \phi_1 \mapsfromto \ldots \phi_n \mapsfromto \beta,$$

where $\phi_l = \chi_l$ if $\chi_l \in \{\varepsilon_1 \ldots, \varepsilon_k\}$ and $\phi_l = \upsilon_l \in \{\varepsilon_1 \ldots, \varepsilon_k\}$ else, is a covering path in $\mathcal{G}$. This implies that $\mathcal{G}$ and $\mathcal{G}^+$ are Markov equivalent.

On the other hand, assume that H is not a tautology. In this case, there exists some assignment of truth values such that every term of H is 0, and let I denote this assignment. We now define the following subset of nodes,

$$C = \{\chi_l : x_l = 1 \text{ in } I\} \cup \{\upsilon_l : x_l = 0 \text{ in } I\}.$$

We see that for all $l = 1, \ldots, n$, either $\chi_l \in C$ or $\upsilon_l \in C$, and this means that β is not μ-separated from α by C in $\mathcal{G}^+$. If we consider a term (again, without loss of generality assuming that the term has three literals),

$$z_1^j \wedge z_2^j \wedge z_3^j.$$

we know that (under assignment I) one of them must equal 0, say z_i^j. If it is a positive literal, then the corresponding variable equals 0 in the assignment and ζ_i^j is not an ancestor of C. If it is a negative literal, then the corresponding variable x_l equals 1 in the assignment, and therefore υ_l is not in C, and ζ_i^j is not an ancestor of C. In either case, we see that every path

$$\alpha \rightarrow \zeta_1^j \mapsfromto \zeta_2^j \mapsfromto \zeta_3^j \mapsfromto \phi_1$$

such that $\phi_1 \in \{\chi_1, \upsilon_1\}$ contains a nonendpoint node which is not an ancestor of C. This implies that the collider path in $\mathcal{G}^+$ between α and β which traverses exactly the nodes in C is not covered in $\mathcal{G}$ and therefore $\mathcal{G}$ and $\mathcal{G}^+$ are not Markov equivalent (Theorem 32).

The reduction from 3DNF tautology to the Markov equivalence problem is clearly done in polynomial time and is a many-one reduction. □

6 Conclusion

We have studied graphs that represent independence structures in stochastic processes that are driven by correlated error processes. We have characterized their equivalence classes in two ways and proven that deciding equivalence is coNP-complete. The characterizations of Markov equivalence do, however, suggest subclasses of cDGs in which deciding Markov equivalence is feasible, e.g., in cDGs with blunt components of bounded size, or in cDGs such that the length of the shortest blunt path between two nodes is bounded.

We have also shown a global Markov property in the case of Ornstein-Uhlenbeck processes driven by correlated Brownian motions. It is an open question if and how this can be extended to other or larger classes of continuous-time stochastic processes.

References

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

Odd O. Aalen and Håkon K. Gjessing. Survival models based on the Ornstein-Uhlenbeck process. *Lifetime Data Analysis*, 10:407–423, 2004.

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.

Krzysztof Bartoszek, Sylvain Glémin, Ingemar Kaj, and Martin Lascoux. Using the Ornstein-Uhlenbeck process to model the evolution of interacting populations. *Journal of Theoretical Biology*, 429, 2017.

Paul Beesack. Systems of multidimensional Volterra integral equations and inequalities. *Nonlinear Analysis: Theory, Methods & Applications*, 1985.

Elmar Böhler, Nadia Creignou, Matthias Galota, Steffen Reith, Henning Schnoor, and Heribert Vollmer. Complexity classifications for different equivalence and audit problems for Boolean circuits. *Logical Methods in Computer Science*, 8(3:27), 2012.

Giacomo Bormetti, Valentina Cazzola, and Danilo Delpini. Option pricing under Ornstein-Uhlenbeck stochastic volatility: A linear model. *International Journal of Theoretical and Applied Finance*, 13(7):1047–1063, 2010.

Venkat Chandrasekaran, Nathan Srebro, and Prahladh Harsha. Complexity of inference in graphical models. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2008.

David Maxwell Chickering, David Heckerman, and Christopher Meek. Large-sample learning of Bayesian networks is NP-hard. *Journal of Machine Learning Research*, 5, 2004.

Chiu H. Choi. A survey of numerical methods for solving matrix Riccati differential equations. 1990.

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.

Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. MIT Press, third edition, 2009.

Robert G. Cowell, Philip Dawid, Steffen L. Lauritzen, and David J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer New York, 1999.

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, 2006.

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.

Susanne Ditlevsen and Petr Lansky. Estimation of the input parameters in the Ornstein-Uhlenbeck neuronal model. *Physical Review E*, 71, 2005.

Michael Eichler. Granger causality and path diagrams for multivariate time series. *Journal of Econometrics*, 137:334–353, 2007.

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

Michael Eichler. *Causality: Statistical perspectives and applications*, chapter Causal inference in time series analysis, pages 327–354. Wiley, 2012b.

Michael Eichler. Causal inference with multiple time series: Principles and problems. *Philosophical Transactions of the Royal Society*, 371(1997):1–17, 2013.

Michael Eichler and Vanessa Didelez. Causal reasoning in graphical time series models. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 109–116, 2007.

Michael Eichler and Vanessa Didelez. On Granger causality and the effect of interventions in time series. *Lifetime Data Analysis*, 16(1):3–32, 2010.

Michael R. Garey and David S. Johnson. *Computers and Intractability: a Guide to the Theory of NP-Completeness.* 1979.

Clive W. J. Granger and Paul Newbold. *Forecasting economic time series.* Academic Press, 2nd edition, 1986.

Chun-Hua Guo and Peter Lancaster. Analysis and modification of Newton's method for algebraic Riccati equations. *Mathematics of Computation*, 67 (223), 1998.

Richard A. Heath. The Ornstein-Uhlenbeck model for decision time in cognitive tasks: An example of control of nonlinear network dynamics. *Psychological Research*, 63:183–191, 2000.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis.* Cambridge University Press, 1985.

Birgit Jacob and Hans J. Zwart. *Linear Port-Hamiltonian Systems on Infinite-dimensional Spaces.* Birkhäuser, 2012.

M. Jacobsen. A brief account of the theory of homogeneous Gaussian diffusions in finite dimensions. In Niemi, H. et.al, editor, *Frontiers in Pure and Applied Probability*, volume 1, pages 86–94, 1993.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques.* The MIT Press, 2009.

Vladimír Kučera. A review of the matrix Riccati equation. *Kybernetika*, 9(1), 1973.

Peter Lancaster and Leiba Rodman. *Algebraic Riccati Equations.* Clarendon Press, 1995.

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

Mei-Ling Ting Lee and G. A. Whitmore. Threshold regression for survival analysis: Modeling event times by a stochastic process reaching a boundary. *Statistical Science*, 21(4):501–513, 2006.

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

Marloes Maathuis, Mathias Drton, Steffen Lauritzen, and Martin Wainwright, editors. *Handbook of graphical models.* CRC Press, 2018.

Christopher Meek. Finding a path is harder than finding a tree. *Journal of Artificial Intelligence Research*, 15:383–389, 2001.

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

G.A. Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations.* Texts in Applied Mathematics. Springer New York, 2014.

Luigi M. Ricciardi and Laura Sacerdote. The Ornstein-Uhlenbeck process as a model for neuronal activity. *Biological Cybernetics*, 35:1–9, 1979.

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

Thomas S. Richardson. A polynomial-time algorithm for deciding Markov equivalence of directed cyclic graphical models. In *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence (UAI)*, 1996b.

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

Thomas S. Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 2003.

Thomas S. Richardson and Peter Spirtes. Ancestral graph markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.

Kjetil Røysland. Counterfactual analyses with graphical models based on local independence. *Annals of Statistics*, 40(4):2162–2194, 2012.

Rainer Schöbel and Jianwei Zhu. Stochastic volatility with an Ornstein-Uhlenbeck process: An extension. *European Finance Review*, 3:23–46, 1999.

T. Shimokawa, K. Pakdaman, T. Takahata, S. Tanabe, and S. Sato. A first-passage-time analysis of the periodically forced noisy leaky integrate-and-fire model. *Biological Cybernetics*, 83:327–340, 2000.

Michael Sipser. *Introduction to the theory of computation.* Thomson Course Technology, 3rd edition, 2013.

Elias M. Stein and Jeremy C. Stein. Stock price distributions with stochastic volatility: An analytic approach. *The Review of Financial Studies*, 4(4):727–752, 1991.

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. Technical Report R-150, University of California, Los Angeles, 1991.

A Proof of Theorem 13

We assume X is a regular Ornstein-Uhlenbeck process with drift

$$\lambda(x) = M(x - \mu)$$

and diffusion matrix σ and let $\Sigma = \sigma\sigma^T$. We let $a = -M\mu$. Let $V = U \dot\cup W$. We will use the following notation similar to that of Liptser and Shiryayev (1977),

$$s \circ s = \sigma_{UU}\sigma_{UU}^T + \sigma_{UW}\sigma_{UW}^T \tag{6}$$

$$s \circ S = \sigma_{UU}\sigma_{WU}^T + \sigma_{UW}\sigma_{WW}^T \tag{7}$$

$$S \circ S = \sigma_{WU}\sigma_{WU}^T + \sigma_{WW}\sigma_{WW}^T \tag{8}$$

Note that the above matrices are simply the block components of $\Sigma = \sigma\sigma^T$,

$$\Sigma = \begin{bmatrix} \sigma_{UU} & \sigma_{UW} \\ \sigma_{WU} & \sigma_{WW} \end{bmatrix} \begin{bmatrix} \sigma_{UU}^T & \sigma_{WU}^T \\ \sigma_{UW}^T & \sigma_{WW}^T \end{bmatrix} = \begin{bmatrix} s \circ s & s \circ S \\ (s \circ S)^T & S \circ S \end{bmatrix}. \tag{9}$$

We let m_t denote $E(X_t^U \mid \mathcal{F}_t^W)$. The following integral equation holds (Liptser and Shiryayev, 1977, Theorem 10.3),

$$m_t = m_0 \int_0^t a_U + M_{UU} m_s + M_{UW} X_s^W \,\mathrm{d}s \tag{10}$$

$$+ \int_0^t (s \circ S + \gamma_s M_{WU}^T)(S \circ S)^{-1}(\,\mathrm{d}X_s^W - (a_W + M_{WU} m_s + M_{WW} X_s^W)\,\mathrm{d}s) \tag{11}$$

where $m_0 = \mathrm{E}\big[X_0^U \mid \mathcal{F}_0^W\big]$ and γ_t is the solution of a differential equation given below. We can write this as

$$\begin{aligned} m_t = m_0 + \int_0^t & a_U + (M_{UU} + (s \circ S + \gamma_s M_{WU}^T)(S \circ S)^{-1} M_{WU}) m_s + M_{UW} X_s^W \, \mathrm{d}s \\ & + \int_0^t (s \circ S + \gamma_s M_{WU}^T)(S \circ S)^{-1} (\, \mathrm{d}X_s^W - (a_W + M_{WW} X_s^W) \, \mathrm{d}s). \end{aligned}$$

The process $\gamma(t)$ is given by the following equation (Liptser and Shiryayev, 1977, Theorem 10.3).

$$\dot{\gamma}(t) = M_{UU}\gamma(t) + \gamma(t) M_{UU}^T + s \circ s \tag{12}$$

$$- \big(s \circ S + \gamma(t) M_{WU}^T\big) \, [S \circ S]^{-1} \big(s \circ S + \gamma(t) M_{WU}^T\big)^T \tag{13}$$

$$= (M_{UU} - (s \circ S)[S \circ S]^{-1} M_{WU})\gamma(t) + \gamma(t)(M_{UU}^T - M_{WU}^T [S \circ S]^{-1} (s \circ S)^T) \tag{14}$$

$$+ s \circ s - (s \circ S)[S \circ S]^{-1}(s \circ S)^T - \gamma(t) M_{WU}^T [S \circ S]^{-1} M_{WU} \gamma(t) \tag{15}$$

This is known as a *differential Riccati equation.* The solution of these equations is unique when we restrict our attention to solutions such that γ_t is symmetric and nonnegative definite (Liptser and Shiryayev, 1977, Theorem 10.3). Essentially, we will show the global Markov property by arguing about the measurability of m_t, using the sparsity of the matrices that go into the integral equation. We will achieve this by first describing the sparsity in the solution of an associated *algebraic* Riccati equation and this will allow us to describe the sparsity in the solution of the differential Riccati equation.

For ease of notation, we now define the matrices

$$D = M_{UU}^T - M_{WU}^T [S \circ S]^{-1} (s \circ S)^T \tag{16}$$

$$E = M_{WU}^T [S \circ S]^{-1} M_{WU} \tag{17}$$

$$F = s \circ s - (s \circ S)[S \circ S]^{-1}(s \circ S)^T \tag{18}$$

and this allows us to write the equation as

$$\dot{\gamma}(t) = \gamma(t) D + D^T \gamma(t) - \gamma(t) E \gamma(t) + F.$$

Note that F is the Schur complement of $S \circ S$ in Σ. The matrix Σ is positive definite by assumption, and therefore so is F (Horn and Johnson, 1985, p. 472).

A.1 Sparsity of the solution of the algebraic Riccati equation

In order to solve the differential Riccati equation, we will first solve an algebraic Riccati equation (Equation (19)) - or rather argue that its solution has a certain sparsity structure.

$$0 = \Gamma D + D^T\Gamma - \Gamma E \Gamma + F \tag{19}$$

Proposition 44. Assume $V = U \dot{\cup} W$, and let $U_1, U_2 \subseteq U$. If there is no m-connecting walk between any $\alpha \in U_1$ and any $\beta \in U_2$ given W, then there exists V_i, $i = 1, \ldots, 6$, such that $U = \bar{V_1} \dot{\cup} \bar{V_2} \dot{\cup} \bar{V_3}, W = V_4 \dot{\cup} V_5 \dot{\cup} V_6$, $U_1 \subseteq V_1$, $U_2 \subseteq V_2$ and furthermore after a reordering of the rows and columns such that the order is consistent with $V_1, \ldots, V_6$, we have the following sparsity of the matrices M and Σ,

$$M = \begin{bmatrix} M_{11} & 0 & 0 & M_{14} & M_{15} & M_{16} \\ 0 & M_{22} & 0 & M_{24} & M_{25} & M_{26} \\ M_{31} & M_{32} & M_{33} & M_{34} & M_{35} & M_{36} \\ M_{41} & 0 & 0 & M_{44} & M_{45} & M_{46} \\ 0 & M_{52} & 0 & M_{54} & M_{55} & M_{56} \\ 0 & 0 & 0 & M_{64} & M_{65} & M_{66} \end{bmatrix},$$

$$\Sigma = \sigma\sigma^T = \begin{bmatrix} \Sigma_{11} & 0 & \Sigma_{13} & \Sigma_{14} & 0 & 0 \\ 0 & \Sigma_{22} & \Sigma_{23} & 0 & \Sigma_{25} & 0 \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} & \Sigma_{34} & \Sigma_{35} & \Sigma_{36} \\ \Sigma_{41} & 0 & \Sigma_{43} & \Sigma_{44} & 0 & 0 \\ 0 & \Sigma_{52} & \Sigma_{53} & 0 & \Sigma_{55} & 0 \\ 0 & 0 & \Sigma_{63} & 0 & 0 & \Sigma_{66} \end{bmatrix}.$$

For both matrices, the subscript ij corresponds to rows V_i and columns V_j.

The concept of μ-separation is similar to that of m-separation which has been used in acyclic graphs. In a graph and for disjoint node sets A, B, and C, we say that A and B are *m-separated* given C if there is no path between any $\alpha \in A$ and any $\beta \in B$ such that every collider is in $\text{an}(C)$ and no noncollider is in C. m-separation is, in contrast to μ-separation, a symmetric notion of separation in the sense that if B is m-separated from A given C, then A is also m-separated from B given C. We will use m-separation as a technical tool in our study of cDGs as some statements are most easily expressed using this symmetric notion.

Proof. We simply define sets of nodes, $V_1, \ldots, V_6$ such that the matrices M and Σ satisfy the above sparsity. Note first that a trivial walk is m-connecting, and it follows that U_1 and U_2 are disjoint. We use the convention that if $A \cap B \neq \emptyset$, then A and B are not m-separated by any subset of $V \setminus (A \cup B)$. For the purpose of this proof, we write $A \rightharpoondown B \mid C$ if there exists $\alpha \in A$ and $\beta \in B$ such that there is walk between α and β with every collider in $\text{an}(C)$ and no noncollider in C and furthermore there is a head or a blunt edge on the final edge at β.

$$\begin{aligned}
V_2 &= \{u \in U : u \perp_m U_1 \mid W\} \\
V_1 &= \{u \in U : u \perp_m V_2 \mid W \text{ and } u \not\perp_m U_1 \mid W\} \\
V_3 &= \{u \in U : u \not\perp_m U_1 \mid W \text{ and } u \not\perp_m V_2 \mid W\} \\
V_4 &= \{w \in W : V_1 \rightharpoondown w \mid W\} \\
V_5 &= \{w \in W : V_2 \rightharpoondown w \mid W\} \\
V_6 &= W \smallsetminus (V_4 \cup V_5)
\end{aligned}$$

Note that $U_1 \subseteq V_1$, and $U_2 \subseteq V_2$. We have that $U = V_1 \dot\cup V_2 \dot\cup V_3$. If $w \in V_4 \cap V_5 \neq \varnothing$, then there is an m-connecting walk between V_2 and V_1 which would be a contradiction, and thus, $W = V_4 \dot\cup V_5 \dot\cup V_6$. Note that Σ is symmetric so we only need to argue that the lower triangular part has the postulated sparsity pattern. Whenever we mention a m-connecting walk in this proof we tacitly mean 'given W'.

Any edge $V_1 \sim V_2$ would create an m-connecting walk and therefore $M_{21} = 0, M_{12} = 0, \Sigma_{21} = 0$. An edge $V_1 \to w \in V_5$ would also create an m-connecting walk between V_1 and V_2 given W as $V_5 \subseteq W$, and therefore $M_{51} = 0$. Similarly, we see that $M_{42} = 0$, $\Sigma_{51} = 0$, and $\Sigma_{42} = 0$. If $V_1 \to w \in V_6$, then w would have to be in V_4, and thus, $M_{61} = 0$. Similarly, $M_{62} = 0$, $\Sigma_{61} = 0$, $\Sigma_{62} = 0$. Let $u \in V_3$. Then there exists an m-connecting walk between u and V_2, and composing this walk with an edge $u \to V_1$ would give an m-connecting walk between V_1 and V_2 as $u \notin W$. This is a contradiction and $M_{13} = 0$. Similarly, $M_{23} = 0$. Consider any $u \in V_3$. There exists m-connecting walks between u and U_1 and u and V_2. None of them can have a tail at u as otherwise their composition would be connecting. Therefore, u is a collider on their composition, and from this it follows that $M_{43} = 0, M_{53} = 0, M_{63} = 0$. If $V_4 \mapsto V_5$, it would follow that there is an m-connecting walk between V_1 and V_2, a contradiction. It follows that $\Sigma_{54} = 0$. If $V_4 \mapsto w$, then $w \in V_4$, and it follows that $\Sigma_{64} = 0$. Similarly, $\Sigma_{65} = 0$. □

The matrices D, E, and F all have their rows and columns indexed by $U = V_1 \dot\cup V_2 \dot\cup V_3$. The above proposition and the definition of the matrices D, E, and F give the following.

Corollary 45. The matrix D has the sparsity structure

$$\begin{bmatrix} * & 0 & * \\ 0 & * & * \\ 0 & 0 & * \end{bmatrix},$$

i.e., $D_{V_2V_1} = 0, D_{V_3V_1} = 0, D_{V_1V_2} = 0$, and $D_{V_3V_2} = 0$. The matrix F is such that $F_{V_1V_2} = 0$ and $F_{V_2V_1} = 0$. The matrix E is block diagonal and $E_{V_3V_3} = 0$.

Lemma 46. If N is an invertible matrix with the sparsity of D, then so is N^{-1}.

Proof. This is easily seen from the Schur complement representation of N^{-1}, using the first two blocks as one component, and the third as the second component. □

Lemma 47. Consider the Lyapunov equation

$$LZ + ZL^T + Q = 0,$$

and let Z_0 denote its solution. If L is stable and has the sparsity pattern of D^T and Q is such that $Q_{V_1V_2} = 0$, $Q_{V_2V_1} = 0$, then $(Z_0)_{V_1V_2} = 0$ and $(Z_0)_{V_2V_1} = 0$.

Proof. The result follows from the explicit solution of a Lyapunov equation when L is stable (Lancaster and Rodman, 1995),

$$Z = \int_0^\infty e^{Ls} Q e^{L^T s}\, \mathrm{d}s.$$

□

Definition 48 (Stabilizable pair of matrices)**.** Let G and H be matrices, $n \times n$ and $n \times m$, respectively. We say that the pair (G, H) is *stabilizable* if there exists an $m \times n$ matrix, K, such that $G + HK$ is stable.

In the literature, stabilizability is used in both the context of continuous-time and discrete-time systems. The above definition is that of a continuous-time system (Lancaster and Rodman, 1995, p. 90). The following is proven in Jacob and Zwart (2012).

Lemma 49. The pair (A, B) is stabilizable if and only if for every eigenvector of the matrix A^T with eigenvalue λ such that $Re(\lambda) \geq 0$ it holds that $v^T B \neq 0$.

Lemma 50. The pair (D, E) is stabilizable.

Proof. We will prove this using Lemma 49. To obtain a contradiction, assume that there exists an eigenvector v of D^T with corresponding eigenvalue λ such that $Re(\lambda) \geq 0$, and assume furthermore that $v^T E = 0$. The matrix $(S \circ S)^{-1}$ is positive definite (since Σ is positive definite), and $v^T M_{WU}^T (S \circ S)^{-1} M_{WU} v = 0$. It follows that $M_{WU} v = 0$. Let o be the column vector of zeros of length l. Note that $\lambda v = D^T v = M_{UU} v$. Then,

$$M \begin{pmatrix} v \\ o \end{pmatrix} = \begin{pmatrix} M_{UU} & M_{UW} \\ M_{WU} & M_{WW} \end{pmatrix} \begin{pmatrix} v \\ o \end{pmatrix} = \lambda \begin{pmatrix} v \\ o \end{pmatrix}$$

It follows that λ is an eigenvalue of M which is a contradiction as M is stable by assumption. □

Corollary 51. There exists a symmetric $k \times k$ matrix X_0 such that $(X_0)_{V_1V_2} = 0$, $(X_0)_{V_2V_1} = 0$ and such that $D - EX_0$ is stable.

Proof. From the above lemma it follows that there exists a $k \times k$ matrix $\bar{X}$ such that $D - E\bar{X}$ is stable. From the sparsity of D and E it follows that for any $k \times k$ matrix, X, $D - EX$ is stable if and only if $D_{\{V1,V2\}\{V1,V2\}} - E_{\{V1,V2\}\{V1,V2\}}X_{\{V1,V2\}\{V1,V2\}}$ and $D_{V_3V_3}$ are stable. The matrices $D_{\{V1,V2\}\{V1,V2\}}$ and $E_{\{V1,V2\}\{V1,V2\}}$ are both block diagonal and thus both pairs of blocks are stabilizable (Lemma 49). It follows that $X_{\{V1,V2\}\{V1,V2\}}$ can be chosen as block diagonal. We need to argue that X_0 can be chosen to be symmetric. The blocks in the diagonal of E are positive semidefinite and are stabilizable (when paired with their corresponding D blocks). Therefore X_0 can be chosen to also be positive definite (Lancaster and Rodman, 1995, Lemma 4.5.4). □

Matrices E and F are both positive semidefinite and there exist unique positive semidefinite matrices $\bar{E}$ and $\bar{F}$ such that $E = \bar{E}\bar{E}$ and such that $F = \bar{F}\bar{F}$ (Horn and Johnson, 1985, Theorem 7.2.6).

Corollary 52. The pair $(D, \bar{E})$ is stabilizable.

Proof. This follows from the fact that (D, E) is stabilizable (Lemma 50). □

Proposition 53. The pair $(\bar{F}, D)$ is detectable, i.e, there exists X such that $X\bar{F} + D$ is stable. The pair (F, D) is also detectable.

Proof. Observe that $\bar{F}$ is invertible. This means that we can choose $X = (I - D)\bar{F}^{-1}$. With this choice of X, the matrix $X\bar{F} + D$ is stable. □

Lemma 54 (Sparsity in solution of algebraic Riccati equation)**.** If there is no m-connecting walk between $\alpha \in U_1$ and $\beta \in U_2$ given W, then $\bar{\Gamma}_{V_1V_2} = 0$ when $\bar{\Gamma}$ is the unique, nonnegative solution of Equation (19).

Proof. We will first argue that there exists a unique, nonnegative solution of Equation (19). We have that E and F are positive semidefinite, that $(D, \bar{E})$ is stabilizable, and that $(\bar{F}, D)$ is detectable. This means that there exists a unique nonnegative solution (Kučera, 1973, Theorem 5), and this is necessarily the maximal solution in the terminology of Lancaster and Rodman (1995). In this proof, we denote this matrix by X_+.

Using Corollary 51, there exists a symmetric $k \times k$ matrix, X_0, such that $(X_0)_{V_1V_2} = 0$, $(X_0)_{V_2V_1} = 0$, and such that $D - EX_0$ is stable. From this matrix, we will define a sequence of matrices that converge to X_+. With this purpose in mind, we define a Newton step as the operation that takes a matrix X_i to the solution of (this is an equation in X)

$$(D - EX_i)^T X + X(D - EX_i) + X_iEX_i + F = 0.$$

Assume now that X_i is such that $(X_i)_{V_1V_2} = 0$ and $(X_i)_{V_2V_1} = 0$. Note first that by Corollary 45, $\bar{Q} = X_iEX_i + F$ is also such that $\bar{Q}_{V_1V_2} = 0$ and $\bar{Q}_{V_2V_1} = 0$. The matrix EX_i has the sparsity pattern of D, and the matrix D does too. By induction and using Lemma 47, it follows that X_i is such that $(X_i)_{V_1V_2} = 0$ and $(X_i)_{V_2V_1} = 0$ for all $i \geq 0$. Note that for all i it holds that $D - EX_i$ is stable (Guo and Lancaster, 1998). Theorem 1.2 of Guo and Lancaster (1998) now gives that

$X_+ = \lim X_i$ is the solution of the algebraic Riccati equation, and it follows from the above that $(X_+)_{V_1V_2} = 0$ and $(X_+)_{V_2V_1} = 0$. □

A.2 Sparsity in the solution of the differential Riccati equation

We will use the above results on the algebraic Riccati equation to describe zero entries of the solution to the differential Riccation equation. From Choi (1990), it follows that if Γ_0 is positive definite, then

$$\Gamma(t) = \bar{\Gamma} + \mathrm{e}^{tK^T}(\Gamma_0 - \bar{\Gamma})\left(I + \int_0^t \mathrm{e}^{sK}E\mathrm{e}^{sK^T}\ \mathrm{d}s(\Gamma_0 - \bar{\Gamma})\right)^{-1}\mathrm{e}^{tK} \qquad (20)$$

where $K = D - E\bar{\Gamma}$ and $\bar{\Gamma}$ is the unique nonnegative definite solution of the algebraic Riccati equation (Equation (19)).

Proof of Equation 20. From Choi (1990), we have that Equation (20) holds under whenever Γ_0 is positive definite as $(D, \bar{E})$ is stabilizable (Corollary 52), and $(\bar{F}, D)$ is detectable (Proposition 53). □

Lemma 55. Assume that Γ_0 is a positive definite matrix such that $(\Gamma_0)_{V_1V_2} = 0$, and let $\Gamma(t)$ denote the solution of the differential Riccati equation (Equation (20)) with initial condition Γ_0. If there is no m-connecting walk between $\alpha \in U_1$ and $\beta \in U_2$ given W, then $(\Gamma(t))_{V_1V_2} = 0$ for all $t \geq 0$.

Proof. This follows directly from the expression in Equation (20) and the sparsity of the matrices that go into that expression: e^{tK} has the sparsity of D and e^{tK^T} has that of D^T. From Lemma 54 we know that $\bar{\Gamma}_{V_1V_2} = 0$. The matrix

$$I + \int_0^t \mathrm{e}^{sK}E\mathrm{e}^{sK^T}\ \mathrm{d}s(\Gamma_0 - \bar{\Gamma})$$

has the sparsity of D and so does its inverse (Lemma 46). This result follows immediately by matrix multiplication. □

Proof of Theorem 13. Let $\beta \in B$ and let $t \in I$. We need to show that $E(\lambda_t^\beta \mid \mathcal{F}_t^{A\cup C})$ is almost surely equal to an $\mathcal{F}_t^C$-measurable random variable. We can without loss of generality assume that A and C are disjoint. The fact that B is μ-separated from A given C implies that $M_{\beta A} = 0$,

$$\begin{aligned} E(\lambda_t^\beta \mid \mathcal{F}_t^{A\cup C}) &= \sum_{\gamma\in A\cup C} M_{\beta\gamma}X_t^\gamma + \sum_{\delta\notin A\cup C} M_{\beta\delta}E(X_t^\delta \mid \mathcal{F}_t^{A\cup C}) \\ &= \sum_{\gamma\in C} M_{\beta\gamma}X_t^\gamma + \sum_{\delta\in \mathrm{pa}(\beta)\smallsetminus(A\cup C)} M_{\beta\delta}E(X_t^\delta \mid \mathcal{F}_t^{A\cup C}). \end{aligned}$$

Let $U = V \smallsetminus A \cup C$. Consider now $V_1 = \{u \in U : u \perp_\mu A \mid C\}$, $V_2 = \{u \in U : u \perp_\mu V_1 \mid A \cup C,\ u \not\perp_m A \mid C\}$, and $V_3 = \{u \in U : u \not\perp_m V_1 \mid A \cup C,\ u \not\perp_m A \mid C\}$. This is a partition of U and we partition $W = A \dot{\cup} C$ as in Proposition 44. This gives the same sparsity structure as in Proposition 44 and the later proofs apply. We see that $\delta \in V_1$ whenever $M_{\beta\delta} \neq 0$. The matrix $M_{UU} + (s \circ S + \gamma_t M_{WU}^T)(S \circ S)^{-1} M_{WU}$ in the integral equation for the conditional expectation process has the sparsity of D^T and it follows that one can solve for $m_t^{V_1}$ independently of $m_t^{U \setminus V_1}$ as the solution of the smaller system is unique (Beesack, 1985). We see that processes X_t^A do not enter into these equations. This follows from the sparsity of $s \circ S$, $S \circ S$, and of $\gamma_t M_{WU}^T$, and the fact that $M_{V_4 A} = 0$ and $M_{V_1 A} = 0$. □

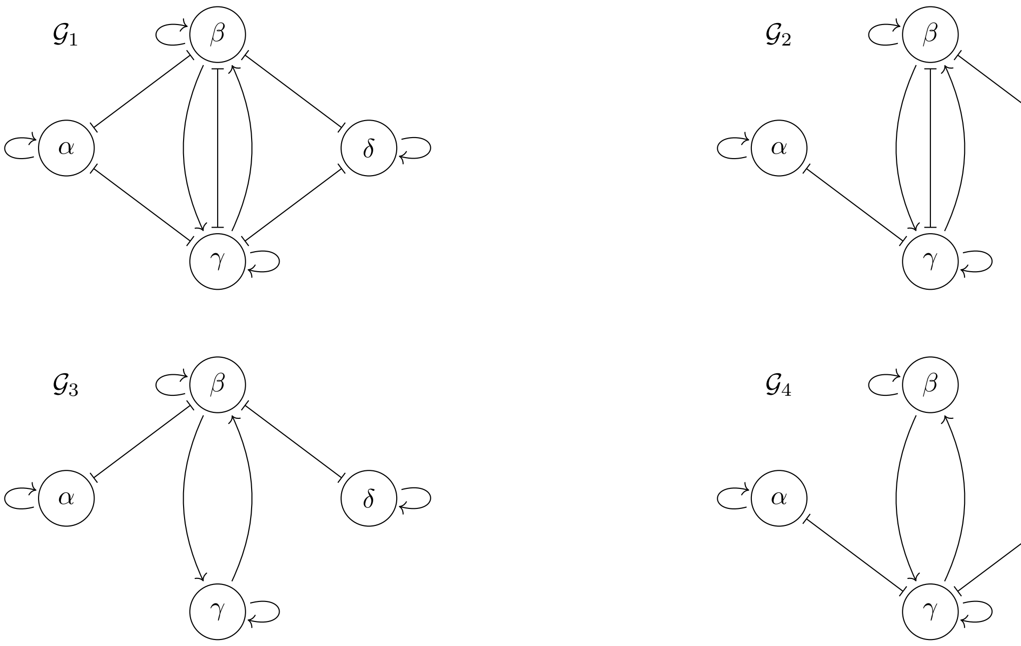


Figure 4.2: Four Markov equivalent cDGs. $\mathcal{G}_1$ is a greatest element of the equivalence class (the entire class is not shown), and therefore the set of minimal representations of $\mathcal{G}_1$ equals the set of minimal representations of $[\mathcal{G}_1]$. The minimal reprensentations of $\mathcal{G}_1$ (or equivalently, for this particular graph, of $[\mathcal{G}_1]$) are shown in the bottom row ($\mathcal{G}_3$ and $\mathcal{G}_4$). $\mathcal{G}_2$ and $\mathcal{G}_1$ are Markov equivalent, and therefore $\mathcal{G}_3$ and $\mathcal{G}_4$ are also the minimal representations of $[\mathcal{G}_2]$. However, $\mathcal{G}_2$ equals the minimal representation of $\mathcal{G}_2$ as there is no Markov equivalent proper subgraph of $\mathcal{G}_2$.

Finding minimal representations of cDGs

We saw in Paper **B** that an equivalence class of cDGs need not have a least element. Given a cDG, $\mathcal{G}$, it is still reasonable to ask for an equivalent graph with fewer edges; maybe even one with as few as possible. We say that $\mathcal{G}^- = (V, E^-)$ is a *minimal representation* of $[\mathcal{G}]$ if $\mathcal{G}^- \in [\mathcal{G}]$ and $|E^-| \leq |\tilde{E}|$ for all $\tilde{\mathcal{G}} = (V, \tilde{E}) \in [\mathcal{G}]$. We say that $\mathcal{G}^-$ is a minimal representation of $\mathcal{G}$ if $\mathcal{G}^- \in [\mathcal{G}]$, $\mathcal{G}^- \subseteq \mathcal{G}$, and $|E^-| \leq |\tilde{E}|$ for any $\tilde{\mathcal{G}} \in [\mathcal{G}]$ such that $\tilde{\mathcal{G}} \subseteq \mathcal{G}$. For any $\mathcal{G}$, minimal representations exist of both $[\mathcal{G}]$ and $\mathcal{G}$, though they are not necessarily unique (see Figure 4.2 for examples).

In this section, we prove that finding minimal representations of $[\mathcal{G}]$ and $\mathcal{G}$ are NP-hard problems.

Theorem 4.2. Let $\mathcal{G} = (V, E)$ be a cDG. It is NP-hard to find a minimal representation of $[\mathcal{G}]$, that is, **Smallest Markov equivalent graph in (cDG, μ)** is NP-hard.

Note that **Smallest Markov equivalent graph** is not a decision problem, however, it is NP-hard as we can find a (Turing) reduction from an NP-complete problem to **Smallest Markov equivalent graph**, thus showing that it is at least as hard as any problem in **NP**.

Proof. We do a reduction from the *minimum set-covering problem.* An instance of the minimum set-covering problem consists of a finite set $\mathcal{U} = \{u_1, \ldots, u_m\}$ and a family, $\mathcal{F} = \{S_1, \ldots, S_l\}$, of subsets of $\mathcal{U}$ such that $\cup_{S \in \mathcal{F}} S = \mathcal{U}$. A *covering* is a subfamily, $\mathcal{C} \subseteq \mathcal{F}$, such that $\cup_{S \in \mathcal{C}} S = \mathcal{U}$, and the *size* of a covering $\mathcal{C}$ is $|\mathcal{C}|$. The minimum set-covering problem asks for a covering of minimal size and is known to be NP-hard as the corresponding decision problem is NP-complete.

From an instance of the minimum set-covering problem, we will construct a graph, $\mathcal{D} = (V, E)$. For each set S_j we include a corresponding node Σ_j, and for each element of the universe, u_k, we include a node v_k. We include auxiliary nodes $\alpha, \sigma_1, \sigma_2$, and σ_3. For each j, we include the edges

$$\alpha \mapsto \Sigma_j \to \sigma_1 \to \sigma_2 \to \sigma_3 \to \Sigma_j$$

We also include the edge $\alpha \mapsto \sigma_1$. Finally, we include $\Sigma_j \mapsto v_k$ if and only if $u_k \in S_j$, and we include all directed loops.

We argue first that $\mathcal{D}$ is a greatest element of $[\mathcal{D}]$ and that if $\bar{\mathcal{D}} = (V, \bar{E}) \in [\mathcal{D}]$ then $\bar{E} = E \cup \{\alpha \mapsto \Sigma_j\}_{j \in J}$ for some $J \subseteq \{1, \ldots, l\}$. Let $\bar{\mathcal{D}} \in [\mathcal{D}]$. For all $\beta, \gamma \in V$, it holds that $\beta \to_{\bar{\mathcal{D}}} \gamma$ if and only if $\beta \to_{\mathcal{D}} \gamma$, using Theorem **B**.32. Note first that for each $i = 1, \ldots, m$, v_i is only weakly inseparable in $\mathcal{D}$ from σ_3 and Σ_j for j such that $u_i \in S_j$. This means that v_i cannot be adjacent with α, σ_1, σ_2, nor with Σ_j for j such that $u_i \notin S_j$, and also not with v_{i_0}, $i \neq i_0$. It also cannot be adjacent with σ_3 as this would make it weakly inseparable from σ_2. Now we consider the possibility of an edge $\Sigma_{j_0} \vdash\!\!\dashv \Sigma_{j_1}$ in $\bar{\mathcal{D}}$. We assume without loss of generality that no set is contained in another, so assume that $u_{i_0} \in S_{j_0} \setminus S_{j_1}$ and $u_{i_1} \in S_{j_1} \setminus S_{j_0}$. υ_{i_0} is weakly inseparable from Σ_{j_0} in $\mathcal{D}$, and υ_{i_1} from Σ_{j_1}. There is a weak inducing path between υ_{i_0} and Σ_{j_0} and if σ_1 is on this weak inducing path, then υ_i and σ_1 would be weakly inseparable which is a contradiction. This means that there must exist a blunt path between υ_{i_0} and Σ_{j_0}, and one between υ_{i_1} and Σ_{j_1}. Every nonendpoint node must be a node in the cyclic component of the Σ's, and composing these paths with the edge $\Sigma_{j_0} \vdash\!\!\dashv \Sigma_{j_1}$ would give a connecting walk in $\bar{\mathcal{D}}$ from υ_{i_0} to υ_{i_1} given $V \setminus \{\alpha\} \cup \{\upsilon_i\}_i$ which is a contradiction. This means that in $\bar{\mathcal{D}}$ there are no edges $\Sigma_{j_0} \vdash\!\!\dashv \Sigma_{j_1}$. In $\bar{\mathcal{D}}$ there must be a weakly inducing path from υ_i to Σ_j. From the above, if they are not adjacent in $\bar{\mathcal{D}}$, we see that it must pass through σ_1, σ_2, or σ_3. This would create an inducing path from υ_i to σ_1 or σ_2 which is a contradiction. This means that $\upsilon_i \vdash\!\!\dashv \Sigma_j$ in $\bar{\mathcal{D}}$ if and only if $u_i \in S_j$. We know that for each Σ_j there is some i such that $\upsilon_i \vdash\!\!\dashv \Sigma_j$ must be in $\bar{\mathcal{D}}$. Therefore, the edges $\Sigma_j \vdash\!\!\dashv \sigma_1$, $\Sigma_j \vdash\!\!\dashv \sigma_2$, $\Sigma_j \vdash\!\!\dashv \sigma_3$ cannot be in $\bar{\mathcal{D}}$ as they would make υ_i weakly inseparable from σ_1, σ_2, and σ_2, respectively. $\sigma_1 \vdash\!\!\dashv \sigma_2$ cannot be in $\bar{\mathcal{D}}$ as this would make Σ_1 weakly inseparable from σ_2. The edge $\sigma_1 \vdash\!\!\dashv \sigma_3$ would make the two weakly inseparable in $\bar{\mathcal{D}}$, and so would $\sigma_2 \vdash\!\!\dashv \sigma_3$. α is weakly inseparable from σ_1 in $\mathcal{D}$ and any weakly inducing path, different from $\alpha \vdash\!\!\dashv \sigma_1$, in $\bar{\mathcal{D}}$ must have a blunt edge between σ_1 and σ_2, σ_3, or Σ_j for some j. From the above, this is impossible, so $\alpha \vdash\!\!\dashv \sigma_1$ must be in $\bar{\mathcal{D}}$. The edges $\alpha \vdash\!\!\dashv \sigma_2$ and $\alpha \vdash\!\!\dashv \sigma_3$ cannot be in $\bar{\mathcal{D}}$ as they would both make α and σ_2 weakly inseparable. This shows that $\mathcal{D}$ is a greatest element of $[\mathcal{D}]$ and that if $\bar{\mathcal{D}} \in [\mathcal{D}]$ and some edge, e, is in $\mathcal{D}$ but not in $\bar{\mathcal{D}}$, then e is a blunt edge between α and Σ_j for some j. The reduction from the instance of the set-covering problem to this graph is done in polynomial time.

Assume that we have a minimal representation of $[\mathcal{D}]$ which we denote by $\mathcal{D}^-$. We define $\mathcal{C}^- = \{S_j : \Sigma_j \vdash\!\!\dashv_{\mathcal{D}^-} \alpha\}$. We argue that $\mathcal{C}$ is a covering. Let $u_k \in \mathcal{U}$, $k = 1, \ldots, m$. Then $u_k \in S_{j_0}$ for some $j_0 = 1, \ldots, l$, and $\alpha \vdash\!\!\dashv \Sigma_{j_0} \vdash\!\!\dashv \upsilon$ in $\mathcal{D}$. There must be a covering collider path between α and υ_k since $\mathcal{D}^-$ is Markov equivalent with $\mathcal{D}$. The only possibility is that $\upsilon_k \vdash\!\!\dashv \Sigma_{j_1} \vdash\!\!\dashv \alpha$, and this means that $u_k \in S_{j_1} \in \mathcal{C}^-$.

Assume that $\mathcal{C}$ is any covering and define $\mathcal{D}_\mathcal{C}$ to be graph obtained from $\mathcal{D}$ by removing every edge $\alpha \vdash\!\!\dashv S_j$ such that $S_j \notin \mathcal{C}$. We will argue that that $\mathcal{D}_\mathcal{C} \in [\mathcal{D}]$ and for this it suffices to argue that any collider path in $\mathcal{D}$ is covered in $\mathcal{D}_\mathcal{C}$ (Theorem **B**.32). Every superset of a covering is also a covering and if we show that we can remove an edge, $\alpha \vdash\!\!\dashv S_j$, Markov equivalently as long as the smaller graph still corresponds to a covering, then the result follows as we can create a sequence of Markov equivalent graphs, $\mathcal{D}_0, \mathcal{D}_1, \ldots, \mathcal{D}_M$ such that $\mathcal{D}_0 = \mathcal{D}$ and $\mathcal{D}_M = \mathcal{D}_\mathcal{C}$. So assume that $\mathcal{D}_2$ corresponds to a covering and that $\mathcal{D}_1 = \mathcal{D}_2 + \{\alpha \vdash\!\!\dashv S_{j_0}\}$. Consider a collider path in $\mathcal{D}_1$ such that $\alpha \vdash\!\!\dashv \Sigma_{j_0}$ is on the path. We assume first that α appears before Σ_{j_0} and divide into cases depending on the subpath after Σ_{j_0}. If Σ_{j_0} is the final node of the path, then we can substitute the blunt edge between α and Σ_{j_0} with $\alpha \vdash\!\!\dashv \sigma_1 \leftarrow \Sigma_{j_0}$ to obtain a covering path in $\mathcal{D}_2$. If instead,

$$\alpha \vdash\!\!\dashv \Sigma_{j_0} \leftarrow \sigma_3$$

then there must be some j_1 such that $\alpha \vdash\!\!\dashv \Sigma_{j_1}$ (otherwise the universe is empty) and we obtain a covering walk. Finally, if

$$\alpha \vdash\!\!\dashv \Sigma_{j_0} \vdash\!\!\dashv \upsilon_k$$

then there exists some j_1 such that $u_k \in S_{j_1}$ in the covering corresponding to $\mathcal{D}_2$ and therefore we can substitute the above subpath for

$$\alpha \vdash\!\!\dashv \Sigma_{j_1} \vdash\!\!\dashv \upsilon_k$$

to obtain a covering walk in $\mathcal{D}_2$. In each case, if we obtain a covering collider walk, we can reduce it to a covering collider path. Similarly, if Σ_{j_0} appears before α on the original collider path. This shows that $\mathcal{D}_\mathcal{C} \in [\mathcal{D}]$. If $|\mathcal{C}| < |\mathcal{C}^-|$, then $\mathcal{D}_\mathcal{C}$ would have strictly fewer edges than $\mathcal{D}^-$, and this shows that $\mathcal{C}^-$ is in fact a minimum set-covering. This proves the correctness of the polynomial-time (Turing) reduction.

□

The sets of minimal representations of $[\mathcal{G}]$ and $\mathcal{G}$ may be different, but they may also be equal. If $\mathcal{G}$ is a greatest element they are in fact equal, and we state this as a proposition.

Proposition 4.3. If $\mathcal{G}$ is a greatest element of $[\mathcal{G}]$, then every minimal representation of $[\mathcal{G}]$ is also a minimal representation of $\mathcal{G}$ and vice versa.

Proof. This follows from the fact that every element of $[\mathcal{G}]$ is a subgraph of $\mathcal{G}$. □

In the graph constructed in the proof of Theorem 4.2, $\mathcal{G}$ is a greatest element and the sets of minimal representations of $\mathcal{G}$ and of $[\mathcal{G}]$ are equal.

Theorem 4.4. Let $\mathcal{G} = (V, E)$ be a cDG. It is NP-hard to find a minimal representation of $\mathcal{G}$, i.e., **Smallest Markov equivalent subgraph in (cDG,** μ) is NP-hard.

Proof. This follows directly from noting that $\mathcal{G}^-$ in the above proof is also a minimal representation of $\mathcal{G}$. □

Tractable Markov equivalence of cDGs

Paper **B** argued that deciding Markov equivalence of two cDGs is coNP-hard. One should keep in mind that this is a worst-case complexity result and the proof of the result exploits arbitrarily long collider paths with no shortcuts. We can devise a polynomial-time algorithm if we restrict ourselves to cDGs without such malicious structures. We say that a walk is *blunt* if it consists of blunt edges only.

Definition 4.5 (The blunt diameter of a cDG)**.** Let $\mathcal{G}$ be a cDG. Its *blunt* diameter is the length of the longest blunt path,

$$\gamma_1 \vdash\!\dashv \gamma_2 \vdash\!\dashv \ldots \vdash\!\dashv \gamma_m,$$

such that $\gamma_i \vdash\!\dashv \gamma_j$ implies $j = i - 1$ or $j = i + 1$.

By imposing a restriction on the blunt diameter, we can find a polynomial-time algorithm to decide Markov equivalence. Assume that $\mathcal{D}_1 = (V, E_1)$ and $\mathcal{D}_2 = (V, E_2)$ have blunt diameters less than or equal to a fixed k, and that their directed parts are equal (otherwise they are surely not Markov equivalent). In this case, every collider path in $\mathcal{D}_i$ is covered by a collider path in $\mathcal{D}_i$ of length $k + 2$ or less, and therefore it is enough to compare collider paths of length $k + 2$ or less. Under the restriction on the blunt diameters of $\mathcal{D}_1$ and $\mathcal{D}_2$, the number of those collider paths scales as a polynomial in $|V|$ and a polynomial-time algorithm follows straightforwardly by checking if every collider path of length $k + 2$ or less is covered in the other graph and vice versa (e.g., using Proposition 4.6).

We finish this section by stating a result that can be turned into an algorithm for checking if a collider path in $\mathcal{D}_1$ is covered in $\mathcal{D}_2$. When $\mathcal{D} = (V, E)$ and $\alpha, \beta \in V$, we define the *collider* graph of $\{\alpha, \beta\}$ to be the cDG $\mathcal{C}_{\mathcal{D}}^{\alpha,\beta} = (V, E^{\mathcal{C}})$ such that for all $\gamma, \delta \in V$

$$\gamma \vdash\!\dashv \delta \in E^{\mathcal{C}} \text{ if and only if } \gamma \vdash\!\dashv \delta \in E$$

and

$$\gamma \to \delta \in E^{\mathcal{C}} \text{ if and only if } \gamma \to \delta \in E \text{ and } \gamma \in \{\alpha, \beta\}.$$

We say that two nodes are *connected* if there exists a walk between them.

Proposition 4.6. Let $\mathcal{D}_1 = (V, E)$ and $\mathcal{D}_2 = (V, E)$ be cDGs with the same directed edges. Consider a collider path, π, in $\mathcal{D}_1$,

$$\alpha \sim \gamma_1 \vdash\!\dashv \ldots \vdash\!\dashv \gamma_m \sim \beta.$$

There exists a collider path in $\mathcal{D}_2$ which covers π if and only if α and β are connected in $(\mathcal{C}_{\mathcal{D}_2}^{\alpha,\beta})_{\mathrm{an}_{\mathcal{D}_2}(\alpha,\gamma_i,\beta_i)}$.

Proof. Assume first there exists a path in $\mathcal{D}_2$ which covers π,

$$\alpha \sim \bar{\gamma}_1 \vdash\!\dashv \ldots \vdash\!\dashv \bar{\gamma}_l \sim \beta.$$

This path is in $\mathcal{C}_{\mathcal{D}_2}^{\alpha,\beta}$, and by assumption, every $\tilde{\gamma}_i$ is an ancestor of $\{\alpha, \beta, \gamma_1, \ldots, \gamma_m\}$, so this path is also in $(\mathcal{C}_{\mathcal{D}_2}^{\alpha,\beta})_{\mathrm{an}_{\mathcal{D}_2}(\alpha,\gamma_i,\beta_i)}$. On the other hand, assume that α and β are connected in $(\mathcal{C}_{\mathcal{D}_2}^{\alpha,\beta})_{\mathrm{an}_{\mathcal{D}_2}(\alpha,\gamma_i,\beta_i)}$. Then there is a collider path

$$\alpha \sim \bar{\gamma}_1 \vdash\!\dashv \ldots \vdash\!\dashv \bar{\gamma}_l \sim \beta$$

such that every node is an ancestor of $\{\alpha, \beta, \gamma_1, \ldots, \gamma_m\}$. This path is also in $\mathcal{D}_2$ and is covering π. □

Chapter 5

Structure learning

Structure learning can be viewed as a type of model selection. In classical model selection we will often have a family of distributions and we choose among these, i.e., we select a model, by optimizing or approximately optimizing some data-dependent criterion. In structure learning, we choose among a family of structures. We will use graphs to represent local independence structures, and the task is to choose between these graphs.

There are different approaches to structure learning (see, e.g., Glymour and Cooper (1999); Spirtes et al. (2000) or Spirtes and Zhang (2018)). The algorithms that we describe in this chapter are all *constraint-based* in the sense that they test if some local independences hold or not and in the end output a graph which is consistent with the outcome of those tests. In this way, we look for graphs that satisfy the constraints that are given by the local independence model. These tests should be thought of as decision procedures; they will lead us to believe that the graph we are looking for looks like this or like that - and this means that they are not reject/no reject tests, but rather reject/accept tests.

Assumptions

We will in this chapter assume that we have access to a local independence *oracle*, a mechanism that will return the correct answer to a local independence query without uncertainty. Though highly unrealistic, this allows us to separate the learning into two separate components: a learning algorithm and a statistical test. One advantage of this is that we can evaluate the performance of a learning algorithm without conflating it with the performance of a specific test of local independence. Furthermore, while the learning algorithms apply to any class of stochastic processes in which we can define local independence, the tests will be specific to a model class.

We say that an independence model, $\mathcal{I}$, is *Markov* with respect to a graph $\mathcal{G}$ using the criterion c if whenever B is c-separated from A by C, it holds that $\langle A, B \mid C\rangle \in \mathcal{I}$. If the opposite implication holds, we say that the independence model is *faithful* to the graph. We will throughout this section assume that we have an independence model $\mathcal{I}$ which is Markov with respect to a DMG $\mathcal{G}_0$, and furthermore assume that $\mathcal{G}_0$ is faithful to $\mathcal{I}$. These assumptions allow us to translate back and forth between separation and independence statements as these assumptions establish a one-to-one correspondence between the graphical and the probabilistic independence models. Especially faithfulness is a quite strong assumption to make.

The graphical learning algorithms we consider will have a specified target, i.e., an unknown graph. We will say that a learning algorithm is *sound* if it is guaranteed to output a supergraph of the target graph in the oracle case and say that it is *complete* if it is guaranteed to output the target graph itself in the oracle case. Note that the target graph differs between algorithms as some algorithms attempt to recover targets that are strictly more informative than others.

We will throughout assume that there exists some underlying multivariate stochastic process which generates the distribution and we assume that the coordinate processes of this underlying multivariate process are indexed by V. Some coordinate processes may be unobserved, however, and we assume that we observe the coordinate processes in $O \subseteq V$. The set O is thus known while V is unknown, apart from the fact that it is a superset of O. We start by considering the easier case where $V = O$ is assumed.

Definition 5.1 (Causal sufficiency)**.** In the context of structure learning, we will say that the observed processes are *causally sufficient* if $O = V$. [1]

[1] One could relax the definition as there could, e.g., be unobserved stochastic processes that were independent from the processes in O in which case one could still reasonably call this system *causally sufficient*. This is not central to this chapter and we will settle for the above simple definition.

Causal minimality

Definition 5.2 (Causal minimality (Schölkopf et al., 2017))**.** We say that a DG, $\mathcal{G}$, is *causally minimal* with respect to a local independence model, $\mathcal{I}$, if

$$\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}$$

and there is no $\bar{\mathcal{G}} \subsetneq \mathcal{G}$ such that $\bar{\mathcal{G}}$ satisfies the above.

Sadeghi (2017) refers to causal minimality as *minimal Markovness*. For a collection of local independences, $\mathcal{I}$, the *induced local independence graph* is simply the DG, $\mathcal{D}$, such that $\alpha \to_{\mathcal{D}} \beta$ if and only if the statement $\alpha \not\to \beta \mid V \setminus \{\alpha\}$ is not in $\mathcal{I}$.

Proposition 5.3. Let $\mathcal{I}$ be a local independence model, and let $\mathcal{D}$ be the induced local independence graph. Then $\mathcal{D}$ is causally minimal with respect to $\mathcal{I}$ whenever we have equivalence of the pairwise and global Markov properties.

Proof. The model $\mathcal{I}$ satisfies the pairwise Markov property with respect to $\mathcal{D}$ by construction. By equivalence of the pairwise and global Markov properties it follows that $\mathcal{I}(\mathcal{D}) \subseteq \mathcal{I}$. Consider instead a proper subgraph of $\mathcal{D}$ and denote this subgraph by $\mathcal{D}^-$. Then there exists $\alpha, \beta \in V$ such that $\alpha \to_{\mathcal{D}} \beta$ and $\alpha \not\to_{\mathcal{D}^-} \beta$. This implies $\alpha \perp_\mu \beta \mid V \setminus \{\alpha\}\ [\mathcal{D}^-]$ and therefore $\mathcal{I}(\mathcal{D}^-) \nsubseteq \mathcal{I}$. □

The above shows that whenever we have causal sufficiency and equivalence of the Markov properties, we can simply test if β is locally independent of α given $V \setminus \{\alpha\}$, and if so leave out the edge $\alpha \to \beta$. The resulting graph will be causally minimal. Without causal sufficiency, the graph is not necessarily maximally informative about the local independences and this motivates learning DMGs instead. In this chapter, Paper **C** considers the problem of learning a maximal Markov equivalent graph from an independence oracle using a constraint-based algorithm. This is computationally hard, and in Paper **D** we will lower our ambitions and aim only to learn the directed part of this graph.

Learning a maximal DMG

Paper **C** attempts to output a graph which (using μ-separation) represents the observed local independence model over O. Under the Markov and faithfulness assumptions, this independence model is represented by an unknown DMG $\mathcal{G}_0 = (O, F)$. When $[\mathcal{G}_0]$ is not a singleton, several DMGs represent the same separation model as $\mathcal{G}_0$. From Paper **A**, we know that each equivalence class has a *maximal* element, denoted by $\mathcal{N}$, and we will output this graph. Distinguishing between the maximal Markov equivalent graph and $\mathcal{G}_0$ is impossible when using only tests of local independence. Having obtained this maximal element, Paper **A** also shows how to concisely represent the Markov equivalence class by computing the DMEG which encodes which edges must be in $\mathcal{G}_0$, which cannot be in $\mathcal{G}_0$, and for which we cannot determine their presence/absence from the local independence model alone.

The most naive approach to learning the supergraph is to simply use the definitions of potential parents and siblings, testing for each (ordered) pair of nodes all the conditions in those definitions. Paper **C** describes a slightly less naive approach in which we first determine which nodes are separable from each other, and which are not. The output, a *separability graph*, is a supergraph of $\mathcal{N}$. We then use some heuristics and the separating sets that are known to us to remove more edges. We can stop at this point to obtain a sound algorithm, or we can for each edge, if needed, use the definitions of potential parents and potential siblings to obtain a complete algorithm. The algorithm is reminiscent of classical constraint-based algorithms in DAG-based structure learning, such as the PC- and FCI-algorithms (Spirtes et al., 2000; Glymour and Cooper, 1999; Colombo et al., 2012). An important difference is the fact that in the asymmetric case there is no need for an orientation phase.

Paper **C**

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

+ supplementary material

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.

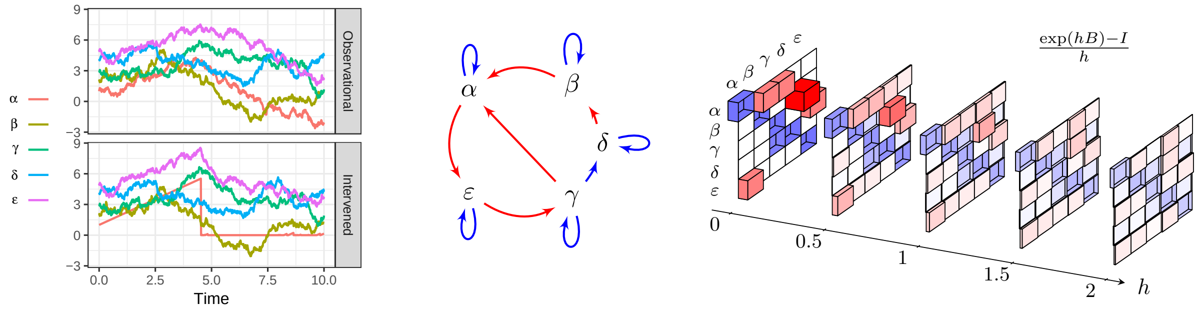


Figure 1: Simulated sample paths (left) for the linear SDE determined by B in (1). The sample paths are from the observational distribution started in the stationary mean as well as under an intervention regime on α. For the local independence graph (middle) the color of the edge $j \to i$ indicates 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 $A = (1,2,3,4,5)^T$,

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

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

$$\mathrm{d}X_t^\alpha = 1(X_t^\beta > 1)\mathrm{d}t, \quad X_t^\alpha - X_{t-}^\alpha = -X_{t-}^\alpha 1(X_t^\beta \leq 1).$$

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 ν_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 $C \subseteq \mathrm{an}(C)$.

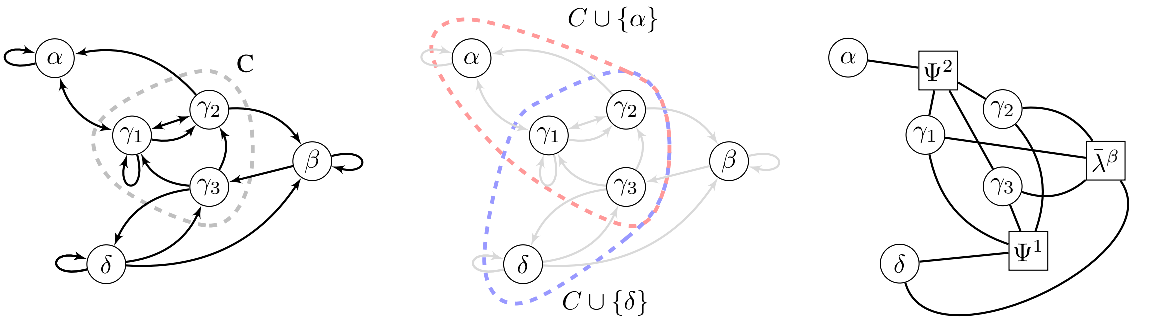


Figure 2: A DMG $\mathcal{G}$ (left) with sets $\{\alpha\}$ and $\{\delta\}$ that factorize w.r.t. $C = \{\gamma_1, \gamma_2, \gamma_3\}$ such that $\alpha \perp_\mu \beta \mid C \cup \{\delta\}$. Any node is μ-separated from either α by $C \cup \{\delta\}$ or δ by $C \cup \{\alpha\}$ (middle), and as $\mathcal{I}(\mathcal{G})$ 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 between 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

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

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

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

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 $(X_s^\alpha)_{0\leq s\leq t}$ 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)

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

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

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

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 $\lambda_{s-}^{A,\alpha} = \lim_{r \to s-} \lambda_r^{A,\alpha}$ denotes the left continuous (and thus predictable) version of the intensity process $\lambda_t^{A,\alpha} = E(\lambda_t^\alpha \mid \mathcal{F}_t^A)$. 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 $\alpha \rightarrow \beta$ 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 $\alpha \rightarrow \beta \in E$ do
        foreach $C \subseteq \mathrm{pa}_{\mathcal{G}}(\beta) \setminus \{\alpha\}$, $|C| = n$ do
            if $\alpha \not\rightarrow_\lambda \beta \mid C$ then
                delete $\alpha \rightarrow \beta$ 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 $\alpha \rightarrow \beta \in E$ do
        foreach $C \subseteq D_{\mathcal{G}}(\alpha, \beta)$, $|C| = n$ do
            if $\alpha \not\rightarrow_\lambda \beta \mid C$ then
                delete $\alpha \rightarrow \beta$ 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 $\beta \to \gamma$ is in $\mathcal{G}$ then
            delete $\beta \to \gamma$ from $\mathcal{G}$;
        end
    end
end
return $\mathcal{G}$
```

Subalgorithm 2: Pruning 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 $\alpha \to \beta$ 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,

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

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 $\operatorname{pa}_{\mathcal{G}}(\beta) \setminus \{\alpha\}$ 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).

```
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 $\alpha \overset{e}{\to} \beta \in E$ do
    if $\mathcal{I}(\mathcal{G} - e) \cap \mathcal{L}_n = \emptyset$ then
        if $\alpha \notin \operatorname{pp}(\beta, \mathcal{I})$ then
            delete $\alpha \to \beta$ 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 $\alpha \notin \operatorname{ps}(\beta, \mathcal{I})$ then
            delete $\alpha \leftrightarrow \beta$ in $\mathcal{G}$;
        end
    end
end
return $\mathcal{G}$
```

Subalgorithm 3: Potential step

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 $\alpha \to \beta \ast\!\!\to \gamma$. We denote such a triple by ${}_w(\alpha, \beta, \gamma)$. We will say that a W-structure is *unshielded* if the edge $\alpha \to \gamma$ 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.

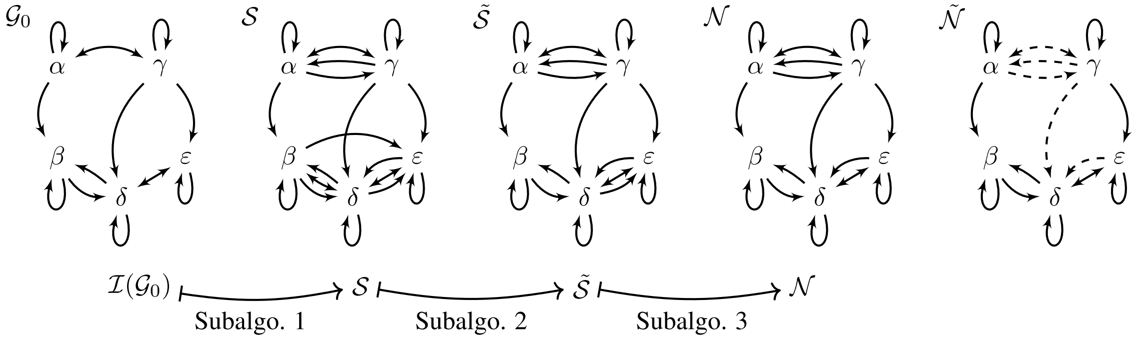


Figure 3: Illustration of the learning algorithm. The DMG $\mathcal{G}_0$ is the underlying graph and we have access to $\mathcal{I} = \mathcal{I}(\mathcal{G}_0)$. Subalgorithm 1 outputs $\mathcal{S}$, the separability graph of $\mathcal{I}(\mathcal{G}_0)$. Subalgorithm 2 prunes $\mathcal{S}$ and outputs $\tilde{\mathcal{S}}$. 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 $\mathcal{N}$.

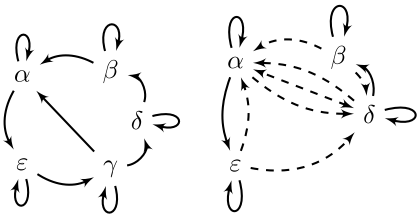


Figure 4: Left: linear SDE example (see Figure 1). Right: the DMEG after marginalization over γ. It is not possible to decide if a loop is directed or bidirected from the independence model only and we choose only to draw the directed loop and to not present it as dashed.

That is, in the case of a directed edge we check each of 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 $\alpha \in \text{ps}(\beta, \mathcal{I})$, then also $\beta \in \text{ps}(\alpha, \mathcal{I})$.

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 for ”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

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 nontrivial. 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, $\gamma \sim \ldots \sim \theta \sim \ldots \rightarrow \beta$. 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 $A_1 \cup A_2 \not\rightarrow_\lambda B \mid C$. We wish to show that $A_1 \not\rightarrow_\lambda B \mid C$.

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

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 $\alpha \ast\!\!\rightarrow \beta$ 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 $\alpha \rightarrow \beta$ 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.

Paper **D**

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

+ supplementary material

Causal screening in dynamical systems

Søren Wengel Mogensen
Department of Mathematical Sciences
University of Copenhagen
Copenhagen, Denmark
swengel@math.ku.dk

Abstract

Many classical algorithms output graphical representations of causal structures by testing conditional independence among a set of random variables. In dynamical systems, local independence can be used analogously as a testable implication of the underlying data-generating process. We suggest some inexpensive methods for causal screening which provide output with a sound causal interpretation under the assumption of ancestral faithfulness. The popular model class of linear Hawkes processes is used to provide an example of a dynamical causal model. We argue that for sparse causal graphs the output will often be close to complete. We give examples of this framework and apply it to a challenging biological system.

1 INTRODUCTION

Constraint-based causal learning is computationally and statistically challenging. There is a large literature on learning structures that are represented by directed acyclic graphs (DAGs) or marginalizations thereof (see Maathuis et al. (2019) for references). The fast causal inference algorithm (FCI, Spirtes et al., 2000) provides in a certain sense maximally informative output (Zhang, 2008), but at the cost of using a large number of conditional independence tests (Colombo et al., 2012). To reduce the computational cost, other methods provide output which has a sound causal interpretation, but may be less informative. Among these are the anytime FCI (Spirtes, 2001) and RFCI (Colombo et al., 2012). A recent algorithm, ancestral causal inference (ACI, Magliacane et al., 2016), aims to learn only the directed part of the underlying graphical structure which allows for a sound causal interpretation even though some information is lost.

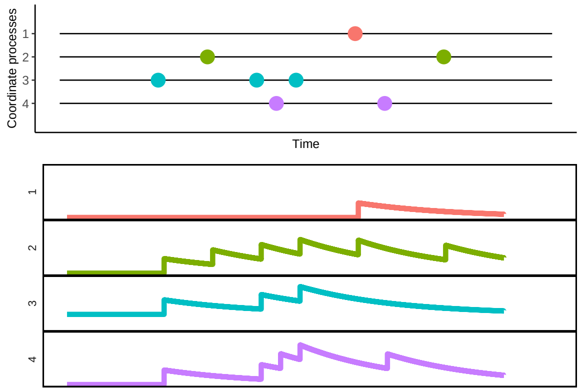
In this paper, we describe some simple methods for learning causal structure in dynamical systems represented by stochastic processes. Many authors have described frameworks and algorithms for learning structure in systems of time series, ordinary differential equations, stochastic differential equations, and point processes. However, most of these methods do not have a clear causal interpretation when the observed processes are part of a larger system and most of the current literature is either non-causal in nature, or requires that there are no unobserved processes.

Analogously to testing conditional independence when learning DAGs, one can use tests of local independence in the case of dynamical systems. Eichler (2013), Meek (2014), and Mogensen et al. (2018) propose algorithms for learning graphs that represent local independence structures. We show empirically that we can recover features of their graphical learning target using considerably fewer tests of local independence. First, we suggest a learning target which is easier to learn, though still conveys useful causal information, analogously to ACI (Magliacane et al., 2016). Second, the proposed algorithm is only guaranteed to provide a supergraph of the learning target and this also reduces the number of local independence tests drastically. A central point is that our proposed methods retain a causal interpretation in the sense that absent edges in the output correspond to implausible causal connections.

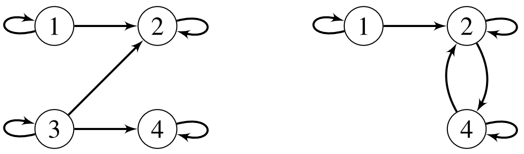
Meek (2014) suggests learning a directed graph to represent a causal dynamical system and gives a learning algorithm which we will describe as a *simple screening algorithm* (Section 4.2). We show that this algorithm can be given a sound interpretation under a weaker faithfulness assumption than that of Meek (2014). We also provide a simple interpretation of the output of this algorithm and we show that similar screening algorithms can give comparable results using considerably fewer tests of local independence.

All proofs are provided in the supplementary material.

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.



(a) Top: Example data from a four-dimensional Hawkes process. Bottom: The corresponding intensities. The time axis is aligned between the two plots.



(b) Left: The causal graph (see Section 2.1) of a four-dimensional Hawkes process. Right: Learning output of standard approach (see Section 2) when 3 is unobserved. When 3 is unobserved, 2 is predictive of 4 and vice versa (heuristically, more events in process 2 indicate more events in 3 which in turn indicates more events in 4). However, they are not causally connected and using local independence one can learn that 2 is not a parent of 4. This is important to predict what would happen under interventions in the system as the right-hand graph indicates that an intervention on 2 would change the distribution of 4 even though this is not the case as $g^{\alpha 2} = 0$ for $\alpha \in \{1, 3, 4\}$.

Figure 1: Subfigure 1a shows data generated from the system in 1b (left). Until the first event all intensities are constant (equal to μ_α for the α-process). The first event occurs in process 3. We see that g^{23}, g^{33}, and g^{43} are different from zero as encoded by the graph in 1b (left). Therefore the event makes the intensity processes of 2, 3, and 4 jump, making new events in these processes more likely in the immediate future (1a, bottom).

2 HAWKES PROCESSES

Local independence can be defined in a wide range of discrete-time and continous-time dynamical models (e.g., point processes (Didelez, 2000), time series (Eichler, 2012), and diffusions (Mogensen et al., 2018). See also Commenges and Gégout-Petit (2009)), and the algorithmic results we present apply to all these classes of models. However, the causal interpretation will differ between these model classes, and we will use the *linear Hawkes processes* to exemplify the framework. Laub et al. (2015) give an accessible introduction to this continuous-time model class and Liniger (2009), Bacry et al. (2015), and Daley and Vere-Jones (2003) provide more background. Hawkes processes have also been studied in the machine learning community in recent years (Zhou et al., 2013a,b; Luo et al., 2015; Xu et al., 2016; Etesami et al., 2016; Achab et al., 2017; Tan et al., 2018; Xu et al., 2018; Trouleau et al., 2019). It is important to note that these papers all consider the case of full observation, i.e., every coordinate process is observed. In causal systems that are not fully observed that assumption may lead to false conclusions (see Figure 1b). Our work addresses the learning problem without the assumption of full observation, hence there can be unknown and unobserved confounding processes.

On a filtered probability space, $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathrm{P})$, we consider an n-dimensional multivariate point process, $X = (X^1, \ldots, X^n)$. $\mathcal{F}_t$ is a filtration, i.e., a nondecreasing family of σ-algebras, and it represents the information which is available at a specific time point. Each coordinate process X^α is described by a sequence of positive, stochastic event times $T_1^\alpha, T_2^\alpha, \ldots$ such that $T_j^\alpha > T_i^\alpha$ almost surely for $j > i$. We let $V = \{1, \ldots, n\}$. This can also be formulated in terms of a counting process, N, such that $N_s^\alpha = \sum_i \mathbb{1}_{(T_i \leq s)}$, $\alpha \in V$. There exists so-called *intensity processes*, $\lambda = (\lambda^1, \ldots, \lambda^n)$, such that

$$\lambda_t^\alpha = \lim_{h \to 0} \frac{1}{h} \mathrm{P}(N_{t+h}^\alpha - N_t^\alpha = 1 \mid \mathcal{F}_t)$$

and the intensity at time t can therefore be thought of as describing the probability of a jump in the immediate future after time t conditionally on the history until time t as captured by the $\mathcal{F}_t$-filtration. In a linear Hawkes model, the intensity of the α-process, $\alpha \in V$, is of the simple form

$$\begin{aligned} \lambda_t^\alpha &= \mu_\alpha + \sum_{\gamma \in V} \int_0^t g^{\alpha\gamma}(t-s)\, \mathrm{d}N_s^\gamma \\ &= \mu_\alpha + \sum_{\gamma \in V} \sum_{i: T_i^\gamma < t} g^{\alpha\gamma}(t - T_i^\gamma) \end{aligned}$$

where $\mu_\alpha \geq 0$ and the function $g^{\alpha\gamma} : \mathbb{R}_+ \to \mathbb{R}$ is nonnegative for all $\alpha, \gamma \in V$. From the above formula, we see that if $g^{\beta\alpha} = 0$, then the α-process does not enter directly into the intensity of the β-process and we will formalize this observation in subsequent sections. The intensity processes determine how the Hawkes process evolves and if $g^{\beta\alpha} = 0$ then the α-process does not directly influence the evolution of the β-process (it may of course have an indirect influence which is mediated by other processes). Figure 1a provides an example of data from a linear Hawkes process and an illustration of its intensity processes.

2.1 A DYNAMICAL CAUSAL MODEL

We will in this section define what we mean by a *dynamical causal model* in the case of a linear Hawkes process and also define a graph (V, E) which represents the causal structure of the model. The node set V is the index set of the coordinate processes of the multivariate Hawkes process, thus identifying each node with a coordinate process. If we first consider the case where $X = (X_1, \ldots, X_n)$ is a multivariate random variable, it is common to define a *causal* model in terms of a DAG, $\mathcal{D}$, and a structural causal model (Pearl, 2009; Peters et al., 2017) by assuming that there exists functions f_i and error terms ϵ_i such that

$$X_i = f_i(X_{\mathrm{pa}_{\mathcal{D}}(X_i)}, \epsilon_i)$$

for $i = 1, \ldots, n$. The causal assumption amounts to assuming that the functional relations are stable under interventions. This idea can be transferred to dynamical systems (see also Røysland (2012); Mogensen et al. (2018)). In the case of a linear Hawkes process as described above, we can consider intervening on the α-process and force events to occur at the deterministic times $t_1, \ldots, t_k$, and at these times only. In this case, the causal assumption amounts to assuming that the distribution of the intervened system is governed by the intensities

$$\lambda_t^\beta = \mu_\beta + \int_0^t g^{\beta\alpha}(t-s)\,\mathrm{d}\bar{N}_s^\alpha + \sum_{\gamma \in V \setminus \{\alpha\}} \int_0^t g^{\beta\gamma}(t-s)\,\mathrm{d}N_s^\gamma$$

for all $\beta \in V \setminus \{\alpha\}$ and where $\bar{N}_t^\alpha = \sum_{i=1}^k \mathbb{1}_{(t_i \leq t)}$. We will not go into a discussion of the existence of these interventional stochastic processes. The above is a *hard* intervention in the sense that the α-process is fixed to be a deterministic function of time. Note that one could easily imagine other types of interventions such as *soft* interventions where the intervention process, α, is not deterministic. One can also extend this to interventions on more than one process. It holds that $N_{t+h}^\beta - N_t^\beta \sim \mathrm{Pois}(\lambda_t^\beta \cdot h)$ in the limit $h \to 0$, and we can think of this as a simulation scheme in which we generate the points in one small interval in accordance to some distribution depending on the history of the process. As such the intensity describes a structural causal model at infinitesimal time steps and the $g^{\alpha\beta}$-functions are in a causal model stable under interventions in the sense that they also describe how the intervention process $\bar{N}^\alpha$ enters into the intensity of the other processes.

We use the set of functions $\{g^{\beta\alpha}\}_{\alpha,\beta\in V}$ to define the *causal graph* of the Hawkes process. A *graph* is a pair (V, E) where V is a set of nodes and E is a set of edges between these nodes. We assume that we observe the Hawkes process in the time interval $J = [0, T]$, $T \in \mathbb{R}$. The causal graph has node set V (the index set of the coordinate processes) and the edge $\alpha \to \beta$ is in the causal graph if and only if $g^{\beta\alpha}$ is not identically zero on J. We call this graph *causal* as it is defined using $\{g^{\beta\alpha}\}_{\alpha,\beta\in V}$ which is a set of mechanisms assumed stable under interventions, and this causal assumption is therefore analogous to that of a classical structural causal model as briefly introduced above.

2.2 PARENT GRAPHS

In principle, we would like to recover the causal graph, $\mathcal{D}$, using local independence tests. Often, we will only have partial observation of the dynamical system in the sense that we only observe the processes in $O \subsetneq V$. We will then aim to learn the *parent graph* of $\mathcal{D}$ on nodes O.

Definition 1 (Parent graph)**.** Let $\mathcal{D} = (V, E)$ be a causal graph and let $O \subseteq V$. The *parent graph* of $\mathcal{D}$ on nodes O is the graph (O, F) such that for $\alpha, \beta \in O$, the edge $\alpha \to \beta$ is in F if and only if the edge $\alpha \to \beta$ is in the causal graph or there is a path $\alpha \to \delta_1 \to \ldots \to \delta_k \to \beta$ in the causal graph such that $\delta_1, \ldots, \delta_k \notin O$, for some $k > 0$.

We denote the parent graph of the causal graph by $\mathcal{P}_O(\mathcal{G})$, or just $\mathcal{P}(\mathcal{G})$ if the set O used is clear from the context. In applications, a parent graph may provide answers to important questions as it tells us the causal relationships between the observed nodes. A similar idea was applied in DAG-based models by Magliacane et al. (2016), though that paper describes an exact method and not a screening procedure. In large systems, it can easily be infeasible to learn the complete independence structure of the observed system, and we propose instead to estimate the parent graph which can be done efficiently. In the supplementary material, we give another characterization of a parent graph. Figure 2 contains an example of a causal graph and a corresponding parent graph.

2.3 LOCAL INDEPENDENCE

Local independence has been studied by several authors and in different classes of continuous-time models as well as in time series (Aalen, 1987; Didelez, 2000, 2008;

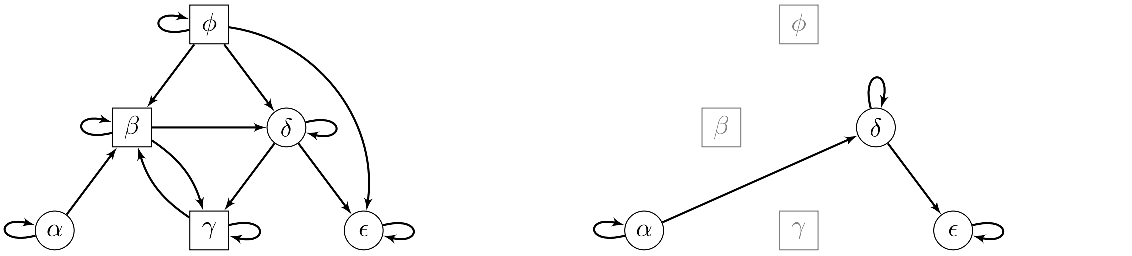


Figure 2: Left: A causal graph on nodes $V = \{\alpha, \beta, \gamma, \delta, \epsilon, \phi\}$. Right: The corresponding parent graph on nodes $O = \{\alpha, \delta, \epsilon\}$. Note that causal graphs and parent graphs may contain cycles. The parent graph does not contain information on the confounder process ϕ as it only encodes 'causal ancestors'. One can also *marginalize* the causal graph to obtain a *directed mixed graph* from which one can read off the parent graph (see the supplementary material).

Eichler and Didelez, 2010). We give an abstract definition of local independence, following the exposition by Mogensen et al. (2018).

Definition 2 (Local independence)**.** Let X be a multivariate stochastic process and let V be an index set of its coordinate processes. Let $\mathcal{F}_t^D$ denote the complete and right-continuous version of the σ-algebra $\sigma(\{X_s^\alpha : s \leq t, \alpha \in D\})$, $D \subseteq V$. Let λ be a multivariate stochastic process (assumed to be integrable and càdlàg) such that its coordinate processes are indexed by V. For $A, B, C \subseteq V$, we say that X^B is λ-locally independent of X^A given X^C (or simply B is λ-locally independent of A given C) if the process

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

has an $\mathcal{F}_t^C$-adapted version for all $\beta \in B$. We write this as $A \not\rightarrow_\lambda B \mid C$, or simply $A \not\rightarrow B \mid C$.

In the case of Hawkes processes, the intensities will be used as the λ-processes in the above definition. Didelez (2000), Mogensen et al. (2018), and Mogensen and Hansen (2020) provide technical details on the definition of local independence. Local independence can be thought of as a dynamical system analogue to the classical conditional independence. It is, however, asymmetric which means that $A \not\rightarrow B \mid C$ does not imply $B \not\rightarrow A \mid C$. This is a natural and desirable feature of an independence relation in a dynamical system as it helps us distinguish between the past and the present. It is important to note that by testing local independences we can obtain more information about the underlying parent graph than by simply assuming full observation and fitting a model to the observed data (see Figure 1b).

2.3.1 Local Independence and the Causal Graph

To make progress on the learning task, we will in this subsection describe the link between the local independence model and the causal graph.

Definition 3 (Pairwise Markov property (Didelez, 2008))**.** We say that a local independence model satisfies the *pairwise Markov property* with respect to a directed graph, $\mathcal{D} = (V, E)$, if the absence of the edge $\alpha \rightarrow \beta$ in $\mathcal{D}$ implies $\alpha \not\rightarrow_\lambda \beta \mid V \setminus \alpha$ for all $\alpha, \beta \in V$.

We will make the following technical assumption throughout the paper. In applications, the functions $g^{\alpha\beta}$ are often assumed to be of the below type (Laub et al. (2015)).

Assumption 4. Assume that N is a multivariate Hawkes process and that we observed N over the interval $J = [0, T]$ where $T > 0$. For all $\alpha, \beta \in V$, the function $g^{\beta\alpha} : \mathbb{R}_+ \rightarrow \mathbb{R}$ is continuous and $\mu_\alpha > 0$.

A version of the following result was also stated by Eichler et al. (2017) but no proof was given and we provide one in the supplementary material. If $\mathcal{G}_1 = (V, E_1)$ and $\mathcal{G}_2 = (V, E_2)$ are graphs, we say that $\mathcal{G}_1$ is a *proper subgraph* of $\mathcal{G}_2$ if $E_1 \subsetneq E_2$.

Proposition 5. The local independence model of a linear Hawkes process satisfies the pairwise Markov property with respect to the causal graph of the process and no proper subgraph of the causal graph has the property.

3 GRAPH THEORY AND INDEPENDENCE MODELS

A *graph* is a pair (V, E) where V is a finite set of nodes and E a finite set of edges. We will use $\sim$ to denote a generic edge. Each edge is between a pair of nodes (not necessarily distinct), and for $\alpha, \beta \in V$, $e \in E$, we will write $\alpha \overset{e}{\sim} \beta$ to denote that the edge e is between α and β. We will in particular consider the class of *directed graphs* (DGs) where between each pair of nodes $\alpha, \beta \in V$ one has a subset of the edges $\{\alpha \rightarrow \beta, \alpha \leftarrow \beta\}$, and we say that these edges are *directed*.

Let $\mathcal{G}_1 = (V, E_1)$ and $\mathcal{G}_2 = (V, E_2)$ be graphs. We say that $\mathcal{G}_2$ is a *supergraph* of $\mathcal{G}_1$, and write $\mathcal{G}_1 \subseteq \mathcal{G}_2$, if $E_1 \subseteq E_2$. For a graph $\mathcal{G} = (V, E)$ such that $\alpha, \beta \in V$, we write $\alpha \rightarrow_{\mathcal{G}} \beta$ to indicate that the directed edge from α to β is contained in the edge set E. In this case we say that α is a *parent* of β. We let $\text{pa}_{\mathcal{G}}(\beta)$ denote the set of nodes in V that are parents of β. We write $\alpha \not\rightarrow_{\mathcal{G}} \beta$ to indicate that the edge is *not* in E. Earlier work allowed loops, i.e., self-edges $\alpha \rightarrow \alpha$, to be either present or absent in the graph (Meek, 2014; Mogensen et al., 2018; Mogensen and Hansen, 2020). We assume that all loops are present, though this is not an essential assumption.

A *walk* is a finite sequence of nodes, $\alpha_i \in V$, and edges, $e_i \in E$, $\langle \alpha_1, e_1, \alpha_2, \ldots, \alpha_k, e_k, \alpha_{k+1} \rangle$ such that e_i is between α_i and α_{i+1} for all $i = 1, \ldots, k$ and such that an orientation of each edge is known. We say that a walk is *nontrivial* if it contains at least one edge. A *path* is a walk such that no node is repeated. A *directed* path from α to β is a path such that all edges are directed and point in the direction of β.

Definition 6 (Trek, directed trek)**.** A *trek* between α and β is a (nontrivial) path $\langle \alpha, e_1, \ldots, e_k, \beta \rangle$ with no colliders (Foygel et al., 2012). We say that a trek between α and β is *directed* from α to β if e_k has a head at β.

We will formulate the following properties using a general *independence model*, $\mathcal{I}$, on V. Let $\mathbb{P}(\cdot)$ denote the power set of some set. An independence model on V is simply a subset of $\mathbb{P}(V) \times \mathbb{P}(V) \times \mathbb{P}(V)$ and can be thought of as a collection of independence statements that hold among the processes/variables indexed by V. In subsequent sections, the independence models will be defined using the notion of local independence. In this case, for $A, B, C \subseteq V$, $A \not\rightarrow_{\lambda} B \mid C$ is equivalent to writing $\langle A, B \mid C \rangle \in \mathcal{I}$ in the abstract notation, and we use the two interchangeably. We do not require $\mathcal{I}$ to be symmetric, i.e., $\langle A, B \mid C \rangle \in \mathcal{I}$ does not imply $\langle B, A \mid C \rangle \in \mathcal{I}$. In the following, we also use μ-separation which is a ternary relation and a dynamical model (and asymmetric) analogue to d-separation or m-separation.

Definition 7 (μ-separation)**.** Let $\mathcal{G} = (V, E)$ be a DMG, and let $\alpha, \beta \in V$ and $C \subseteq V$. We say that a (nontrivial) walk from α to β, $\langle \alpha, e_1, \ldots, e_k, \beta \rangle$, is μ-connecting given C if $\alpha \notin C$, the edge e_k has a head at β, every collider on the walk is in $\text{an}(C)$ and no noncollider is in C. Let $A, B, C \subseteq V$. We say that B is μ-separated from A given C if there is no μ-connecting walk from any $\alpha \in A$ to any $\beta \in B$ given C. In this case, we write $A \perp_{\mu} B \mid C$, or $A \perp_{\mu} B \mid C \ [\mathcal{G}]$ if we wish to emphasize the graph to which the statement relates.

More graph-theoretical definitions and references are given in the supplementary material.

Definition 8 (Global Markov property)**.** We say that an independence model $\mathcal{I}$ satisfies the *global Markov property* with respect to a DG, $\mathcal{G} = (V, E)$, if $A \perp_{\mu} B \mid C \ [\mathcal{G}]$ implies $\langle A, B \mid C \rangle \in \mathcal{I}$ for all $A, B, C \subseteq V$.

From Proposition 5, we know that the local independence model of a linear Hawkes process satisfies the pairwise Markov property with respect to its causal graph, and using the results in Didelez (2008) and Mogensen et al. (2018) it also satisfies the global Markov property with respect to this graph.

Definition 9 (Faithfulness)**.** We say that $\mathcal{I}$ is *faithful* with respect to a DG, $\mathcal{G} = (V, E)$, if $\langle A, B \mid C \rangle \in \mathcal{I}$ implies $A \perp_{\mu} B \mid C \ [\mathcal{G}]$ for all $A, B, C \subseteq V$.

4 NEW LEARNING ALGORITHMS

In this section, we state a very general class of algorithms which is easily seen to provide sound causal learning and we describe some specific algorithms. We throughout assume that there is some underlying, true DG, $\mathcal{D}_0 = (V, E)$, describing the causal model and we wish to output $\mathcal{P}_O(\mathcal{D}_0)$. However, this graph is not in general identifiable from the local independence model. In the supplementary material, we argue that for an equivalence class of parent graphs, there exists a unique member of the class which is a supergraph of all other members. Denote this unique graph by $\bar{\mathcal{D}}$. Our algorithms will output supergraphs of $\bar{\mathcal{D}}$, and the output will therefore also be supergraphs of the true parent graph.

We assume that we are in the 'oracle case', i.e., have access to a local independence oracle that provides the correct answers. We will say that an algorithm is *sound* if it in the oracle case outputs a supergraph of $\bar{\mathcal{D}}$ and that it is *complete* if it outputs $\bar{\mathcal{D}}$. We let $\mathcal{I}^O$ denote the local independence model restricted to subsets of O, i.e., this is the observed part of the local independence model. We provide algorithms that are guaranteed to be sound, but only complete in particular cases. Naturally, one would wish for completeness as well. However, complete algorithms can easily be computationally infeasible whereas sound algorithms can be very inexpensive (e.g., Mogensen et al., 2018). We think of these sound algorithms as *screening procedures* as they rule out some causal connections, but do not ensure completeness.

4.1 ANCESTRAL FAITHFULNESS

Under the faithfulness assumption, every local independence implies μ-separation in the graph. We assume a weaker, but similar, property to show soundness. For learning marginalized DAGs, weaker types of faithfulness have also been explored, see Zhang and Spirtes (2008); Zhalama et al. (2017a,b).

Definition 10 (Ancestral faithfulness)**.** Let $\mathcal{I}$ be an independence model and let $\mathcal{D}$ be a DG. We say that $\mathcal{I}$ satisfies *ancestral faithfulness* with respect to $\mathcal{D}$ if for every $\alpha, \beta \in V$ and $C \subseteq V \setminus \{\alpha\}$, $\langle \alpha, \beta \mid C \rangle \in \mathcal{I}$ implies that there is no μ-connecting directed path from α to β given C in $\mathcal{D}$.

Ancestral faithfulness is a strictly weaker requirement than faithfulness. We conjecture that local independence models of linear Hawkes processes satisfy ancestral faithfulness with respect to their causal graphs. Heuristically, if there is a directed path from α to β which is not blocked by any node in C, then information should flow from α to β, and this cannot be 'cancelled out' by other paths in the graph as the linear Hawkes processes are self-excitatory, i.e., no process has a dampening effect on any process. This conjecture is supported by the so-called *Poisson cluster representation* of a linear Hawkes process (see Jovanović et al. (2015)).

4.2 SIMPLE SCREENING ALGORITHMS

As a first step in describing a causal screening algorithm, we will define a very general class of learning algorithms that simply test local independences and sequentially remove edges. It is easily seen that under the assumption of ancestral faithfulness every algorithm in this class gives sound learning in the oracle case. The *complete* DG on nodes V is the DG with edge set $\{\alpha \to \beta \mid \alpha, \beta \in V\}$.

Definition 11 (Simple screening algorithm)**.** We say that a learning algorithm is a *simple screening algorithm* if it starts from a complete DG on nodes O and removes an edge $\alpha \to \beta$ only if a conditioning set $C \subseteq O \setminus \{\alpha\}$ has been found such that $\langle \alpha, \beta \mid C \rangle \in \mathcal{I}^O$.

The next results describe what can be learned from absent edges in the output of a simple screening algorithm.

Proposition 12. Assume that $\mathcal{I}$ satisfies ancestral faithfulness with respect to $\mathcal{D}_0 = (V, E)$. The output of any simple screening algorithm is sound in the oracle case.

Corollary 13. Assume ancestral faithfulness of $\mathcal{I}$ with respect to $\mathcal{D}_0$ and let $A, B, C \subseteq O$. If every directed path from A to B goes through C in the output graph of a simple screening algorithm, then every directed path from A to B goes through C in $\mathcal{D}_0$.

Corollary 14. If there is no directed path from A to B in the output graph, then there is no directed path from A to B in $\mathcal{D}_0$.

4.3 PARENT LEARNING

In the previous section, it was shown that if edges are only removed when a separating set is found the output is sound under the assumption of ancestral faithfulness. In this section we give a specific algorithm. The key observation is that we can easily retrieve structural information from a rather small subset of local independence tests. Let $\mathcal{D}^t$ denote the output from Subalgorithm 1 (see below). The following result shows that under the assumption of faithfulness, $\alpha \to_{\mathcal{D}^t} \beta$ if and only if there is a directed trek from α to β in $\mathcal{D}_0$.

Proposition 15. There is no directed trek from α to β in $\mathcal{D}_0$ if and only if $\alpha \perp_\mu \beta \mid \beta$ $[\mathcal{D}_0]$.

Note that above, β in the conditioning set represents the β-past while the other β represents the present of the β-process. While there is no distinction in the graph, this interpretation follows from the definition of local independence and the global Markov property. We will refer to running first Subalgorithm 1 and then Subalgorithm 2 (using the output DG from the first as input to the second) as the causal screening (CS) algorithm. Intuitively, Subalgorithm 2 simply tests if a candidate set (the parent set) is a separating set and other candidate sets could be chosen.

Proposition 16. The CS algorithm is a simple screening algorithm.

It is of course of interest to understand under what conditions the edge $\alpha \to \beta$ is guaranteed to be removed by the CS algorithm when it is not in the underlying target graph. In the supplementary material we state and prove a result describing one such condition.

```
input  : a local independence oracle for $\mathcal{I}^O$
output : a DG on nodes $O$
initialize $\mathcal{D}$ as the complete DG on $O$;
foreach $(\alpha, \beta) \in V \times V$ do
    if $\alpha \not\to_\lambda \beta \mid \beta$ then
        delete $\alpha \to \beta$ from $\mathcal{D}$;
    end
end
return $\mathcal{D}$
```

Subalgorithm 1: Trek step

```
input  : a local independence oracle for $\mathcal{I}^O$ and a
         DG, $\mathcal{D} = (O, E)$
output : a DG on nodes $O$
foreach $(\alpha, \beta) \in V \times V$ such that $\alpha \to_{\mathcal{D}} \beta$ do
    if $\alpha \not\to_\lambda \beta \mid \mathrm{pa}_{\mathcal{D}}(\beta) \setminus \{\alpha\}$ then
        delete $\alpha \to \beta$ from $\mathcal{D}$;
    end
end
return $\mathcal{D}$
```

Subalgorithm 2: Parent step

4.4 ANCESTRY PROPAGATION

In this section, we describe an additional step which propagates ancestry by reusing the output of Subalgorithm 1 to remove further edges. This comes at a price as one needs faithfulness to ensure soundness. The idea is similar to ACI (Magliacane et al., 2016).

```
input  : a DG, $\mathcal{D} = (O, E)$
output : a DG on nodes $O$
initialize $E_r = \emptyset$ as the empty edge set;
foreach $(\alpha, \beta, \gamma) \in V \times V \times V$ such that $\alpha, \beta, \gamma$ are all distinct do
    if $\alpha \to_{\mathcal{D}} \beta$, $\beta \not\to_{\mathcal{D}} \alpha$, $\beta \to_{\mathcal{D}} \gamma$, and $\alpha \not\to_{\mathcal{D}} \gamma$ then
        update $E_r = E_r \cup \{\beta \to \gamma\}$;
    end
end
Update $\mathcal{D} = (V, E \setminus E_r)$;
return $\mathcal{D}$
```

Subalgorithm 3: Ancestry propagation

In ancestry propagation, we exploit the fact that any trek between α and β (such that γ is not on this trek) composed with the edge $\beta \to \gamma$ gives a directed trek from α to γ. We only use the trek between α and β 'in one direction', as a directed trek from α to β. In Subalgorithm 4 (supplementary material), we use a trek between α and β twice when possible, at the cost of an additional test.

We can construct an algorithm by first running Subalgorithm 1, then Subalgorithm 3, and finally Subalgorithm 2 (using the output of one subalgorithm as input to the next). We will call this the CSAPC algorithm. If we use Subalgorithm 4 (in the supplementary material) instead of Subalgorithm 3, we will call this the CSAP.

Proposition 17. If $\mathcal{I}$ is faithful with respect to $\mathcal{D}_0$, then CSAP and CSAPC both provide sound learning.

5 APPLICATION AND SIMULATIONS

When evaluating the performance of a sound screening algoritm, the output graph is guaranteed to be a supergraph of the true parent graph, and we will say that edges that are in the output but not in the true graph are *excess edges*. For a node in a directed graph, the *indegree* is the number of directed edges adjacent with and pointed into the node, and the *outdegree* is the number of directed edges adjacent with and pointed away from the node.

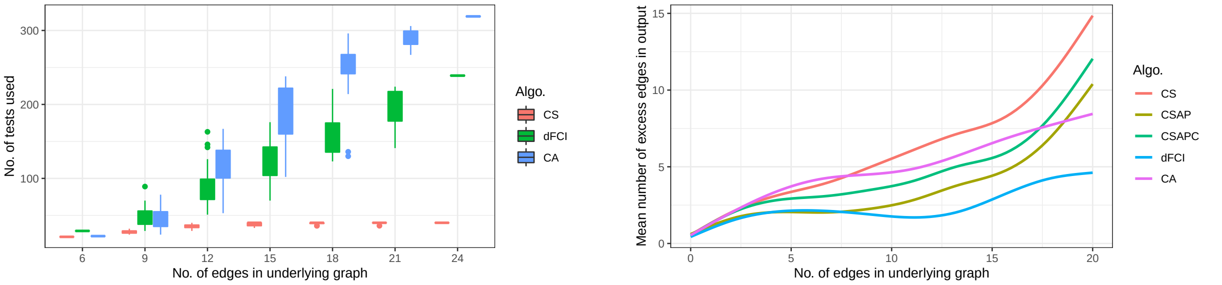
One should note that all our experiments are done using an *oracle test*, i.e., instead of using real or synthetic data, the algorithms simply query an oracle for each local independence and receive the correct answer. This tests whether or not an algorithm can give good results using an efficient testing strategy (i.e., a low number of queries to the oracle) and therefore it evaluates the algorithms. This approach separates the algorithm from the specific test of local independence and evaluates only the algorithm. As such this is highly unrealistic as we would never have access to an oracle with real data, however, we should think of these experiments as a study of efficiency. The oracle approach to evaluating graphical learning algorithms is common in the DAG-based case, see Spirtes (2010) for an overview.

Also note that the comparison is only made with other constraint-based learning algorithms that can actually solve the problem at hand. Learning methods that assume full observation (such as the Hawkes methods mentioned in Section 2) would generally not output a graph with the correct interpretation even in the oracle case (see the example in Figure 1b).

5.1 C. ELEGANS NEURONAL NETWORK

Caenorhabditis elegans is a roundworm in which the network between neurons has been mapped completely (Varshney et al., 2011). We apply our methods to this network as an application to a highly complex network. It consists of 279 neurons which are connected by both non-directional *gap junctions* and directional chemical synapses. We will represent the former as an unobserved process and the latter as a direct influence which is consistent with the biological system (Varshney et al., 2011). From this network, we sampled subnetworks of 75 neurons each (details in the supplementary material) and computed the output of the CS algorithm. These subsampled networks had on average 1109 edges (including bidirected edges representing unobserved processes, see the supplementary material) and on average 424 directed edges. The output graphs had on average 438 excess edges which is explained by the fact that there are many unobserved nodes in the graphs. To compare the output to the true parent graph, we computed the rank correlation between the indegrees of the nodes in the output graph and the indegrees of the nodes in the true parent graph, and similarly for the outdegree (indegree correlation: 0.94, outdegree correlation: 0.52). Finally, we investigated the method's ability to identify the observed nodes of highest directed connectivity (i.e., highest in- and outdegrees). The neuronal network of c. elegans is inhomogeneous in the sense that some neurons are extremely highly connected while others are only very sparsely connected. We considered the 15 nodes of highest indegree/outdegree (out of the 75 observed nodes). On average, the CS algorithm placed 13.4 (in) and 9.2 (out) of these 15 among the 15 most connected nodes.

From the output of the CS algorithm, we can find areas of the neuronal network which mediates information from one area to another, e.g., using Corollary 13.



(a) Comparison of number of tests used. For each level of sparsity (number of edges in true graph), we generated 500 graphs, all on 5 nodes. The number of tests required quickly rises for dFCI and CA while CS spends no more than $2 \cdot 5(5-1)$ tests. The output of dFCI and CA is not considerably more informative as measured by the mean number of excess edges: CS 0.96, dFCI 0.07, CA 0.81 (average over all levels of sparsity).

(b) Mean number of excess edges in output graphs for varying numbers of edges (bidirected and directed) in the true graph (all graphs are on 10 nodes), not counting loops.

Figure 3: Comparison of performance.

5.2 COMPARISON OF ALGORITHMS

In this section we compare the proposed causal screening algorithms with previously published algorithms that solve similar problems. Mogensen et al. (2018) propose two algorithms, one of which is sure to output the correct graph when an oracle test is available. They note that this complete algorithm is computationally very expensive and adds little extra information, and therefore we will only consider their other algorithm for comparison. We will call this algorithm *dynamical* FCI (dFCI) as it resembles FCI (Mogensen et al., 2018). dFCI actually solves a harder learning problem (see details in the supplementary material), however, it is computationally infeasible for many problems.

The Causal Analysis (CA) algorithm of Meek (2014) is a simple screening algorithm and we have in this paper argued that it is sound for learning the parent graph under the weaker assumption of ancestral faithfulness. Even though this algorithm uses a large number of tests, it is not guaranteed to provide complete learning as there may be inseparable nodes that are not adjacent (Mogensen et al., 2018; Mogensen and Hansen, 2020).

For the comparison of these algorithms, two aspects are important. As they are all sound, one aspect is the number of excess edges. The other aspect is of course the number of tests needed. The CS and CSAPC algorithms use at most $2n(n-1)$ tests and empirically the CSAP uses roughly the same number as the two former. This makes them feasible in large graphs. The quality of their output is dependent on the sparsity of the true graph, though the CSAP and CSAPC algorithms can deal considerably better with less sparse graphs (Subfigure 3b).

6 DISCUSSION

We suggested inexpensive constraint-based methods for learning causal structure based on testing local independence. An important observation is that local independence is asymmetric while conditional independence is symmetric. In a certain sense, this may help when constructing learning algorithms as there is no need of something like an 'orientation phase' as in the FCI. This facilitates using very simple methods to give sound causal learning as we do not need the independence structure in full to give interesting output. Simple screening algorithms may be either adaptive or nonadaptive. We note that nonadaptive algorithms may be more robust to false conclusions from statistical tests of local independence.

The amount of information in the output of the screening algorithms depends on the sparsity of the true graph. However, even in examples with very little sparsity interesting structural information can be learned.

We showed that the proposed algorithms have a computational advantage over previously published algorithms within this framework. This makes it feasible to consider causal learning in large networks with unobserved processes. We obtained this gain in efficiency in part by outputting only the directed part of the causal structure. This means that we may be able to answer structural questions, but not questions relating to causal effect estimation.

Acknowledgments

This work was supported by VILLUM FONDEN (research grant 13358). We thank Niels Richard Hansen and the anonymous reviewers for their helpful comments that improved this paper.

References

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

Massil Achab, Emmanuel Bacry, Stéphane Gaïffas, Iacopo Mastromatteo, and Jean-François Muzy. Uncovering causality from multivariate Hawkes integrated cumulants. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Emmanuel Bacry, Iacopo Mastromatteo, and Jean-François Muzy. Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(1), 2015.

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.

Daryl J. Daley and David D. Vere-Jones. *An introduction to the theory of point processes*. New York: Springer, 2nd edition, 2003.

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

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.

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

Michael Eichler. Causal inference with multiple time series: Principles and problems. *Philosophical Transactions of the Royal Society*, 371(1997):1–17, 2013.

Michael Eichler and Vanessa Didelez. On Granger causality and the effect of interventions in time series. *Lifetime Data Analysis*, 16(1):3–32, 2010.

Michael Eichler, Rainer Dahlhaus, and Johannes Dueck. Graphical modeling for multivariate Hawkes processes with nonparametric link functions. *Journal of Time Series Analysis*, 38:225–242, 2017.

Jalal Etesami, Negar Kiyavash, Kun Zhang, and Kushagra Singhal. Learning network of multivariate Hawkes processes: A time series approach. In *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence (UAI)*, 2016.

Rina Foygel, Jan Draisma, and Mathias Drton. Half-trek criterion for generic identifiability of linear structural equation models. *The Annals of Statistics*, 40(3):1682–1713, 2012.

Stojan Jovanović, John Hertz, and Stefan Rotter. Cumulants of Hawkes point processes. *Physical Review E*, 91(4), 2015.

Patrick J. Laub, Thomas Taimre, and Philip K. Pollett. Hawkes processes. 2015. URL `https://arxiv.org/pdf/1507.02822.pdf`.

Thomas Josef Liniger. *Multivariate Hawkes processes*. PhD thesis, ETH Zürich, 2009.

Dixin Luo, Hongteng Xu, Yi Zhen, Xia Ning, Hongyuan Zha, Xiaokang Yang, and Wenjun Zhang. Multi-task multi-dimensional Hawkes processes for modeling event sequences. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, 2015.

Marloes Maathuis, Mathias Drton, Steffen Lauritzen, and Martin Wainwright. *Handbook of graphical models*. Chapman & Hall/CRC handbooks of modern statistical methods, 2019.

Sara Magliacane, Tom Claassen, and Joris M. Mooij. Ancestral causal inference. In *Proceedings of the 29th Conference on Neural Information Processing Systems (NIPS)*, 2016.

Christopher Meek. Toward learning graphical and causal process models. In *CI'14 Proceedings of the UAI 2014 Conference on Causal Inference: Learning and Prediction*, 2014.

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

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Jonas Christopher Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference, foundations and learning algorithms*. MIT Press, 2017.

Kjetil Røysland. Counterfactual analyses with graphical models based on local independence. *The Annals of Statistics*, 40(4):2162–2194, 2012.

Peter Spirtes. An anytime algorithm for causal inference. In *Proceedings of the 8th International Workshop on Artificial Intelligence and Statistics (AISTATS)*, 2001.

Peter Spirtes. Introduction to causal inference. *Journal of Machine Learning Research*, 11, 2010.

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

Xi Tan, Vinayak Rao, and Jennifer Neville. Nested CRP with Hawkes-Gaussian processes. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

William Trouleau, Jalal Etesami, Matthias Grossglauser, Negar Kiyavash, and Patrick Thiran. Learning Hawkes processes under synchronization noise. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Lav R. Varshney, Beth L. Chen, Eric Paniagua, David H. Hall, and Dmitri B. Chklovskii. Structural properties of the Caenorhabditis elegans neuronal network. *PLoS Computational Biology*, 7(2), 2011.

Hongteng Xu, Mehrdad Farajtabar, and Hongyuan Zha. Learning Granger causality for Hawkes processes. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016.

Hongteng Xu, Dixin Luo, Xu Chen, and Lawrence Carin. Benefits from superposed Hawkes processes. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

Zhalama, Jiji Zhang, Frederick Eberhardt, and Wolfgang Mayer. SAT-based causal discovery under weaker assumptions. In *Proceedings of the 33th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017a.

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

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.

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

Ke Zhou, Hongyuan Zha, and Le Song. Learning triggering kernels for multi-dimensional Hawkes processes. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013a.

Ke Zhou, Hongyuan Zha, and Le Song. Learning social infectivity in sparse low-rank networks using multidimensional Hawkes processes. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2013b.

Supplementary material for Causal screening in dynamical systems

Søren Wengel Mogensen
Department of Mathematical Sciences
University of Copenhagen
Copenhagen, Denmark
swengel@math.ku.dk

This supplementary material contains additional graph theory, results, and definitions, as well as the proofs of the main paper.

1 GRAPH THEORY

In the main paper, we introduce the class of DGs to represent causal structures. One can represent marginalized DGs using the larger class of DMGs. A *directed mixed graph* (DMG) is a graph such that any pair of nodes $\alpha, \beta \in V$ is joined by a subset of the edges $\{\alpha \rightarrow \beta, \alpha \leftarrow \beta, \alpha \leftrightarrow \beta\}$.

We say that edges $\alpha \rightarrow \beta$ and $\alpha \leftarrow \beta$ are *directed*, and that $\alpha \leftrightarrow \beta$ is *bidirected*. We say that the edge $\alpha \rightarrow \beta$ has a *head* at β and a *tail* at α. $\alpha \leftrightarrow \beta$ has heads at both α and β. We also introduced a walk $\langle \alpha_1, e_1, \alpha_2, \ldots, \alpha_n, e_n, \alpha_{n+1} \rangle$. We say that α_1 and α_{n+1} are endpoint nodes. A nonendpoint node α_i on a walk is a *collider* if e_{i-1} and e_i both have heads at α_i, and otherwise it is a *noncollider.* A cycle is a path $\langle \alpha, e_1, \ldots, \beta \rangle$ composed with an edge between α and β. We say that α is an *ancestor* of β if there exists a directed path from α to β. We let an(β) denote the set of nodes that are ancestors of β. For a node set C, we let an$(C) = \cup_{\beta \in C}$an(β). By convention, we say that a trivial path (i.e., with no edges) is directed and this means that $C \subseteq$ an(C).

For DAGs d-separation is often used for encoding independences. We use the analogous notion of μ-separation which is a generalization of δ-separation Didelez (2000, 2008); Meek (2014); Mogensen and Hansen (2020).

We use the class of DGs to represent the underlying, data-generating structure. When only parts of the causal system is observed, the class of DMGs can be used to represent marginalized DGs Mogensen and Hansen (2020). This can be done using *latent projection* Verma and Pearl (1991); Mogensen and Hansen (2020) which is a map that for a DG (or more generally, for a DMG), $\mathcal{D} = (V, E)$, and a subset of observed nodes/processes, $O \subseteq V$, provides a DMG, $m(\mathcal{D}, O)$, such that for all $A, B, C \subseteq O$,

$$A \perp_\mu B \mid C \; [\mathcal{D}] \Leftrightarrow A \perp_\mu B \mid C \; [m(\mathcal{D}, O)].$$

See Mogensen and Hansen (2020) for details on this graphical marginalization. We say that two DMGs, $\mathcal{G}_1 = (V, E_1), \mathcal{G}_2 = (V, E_2)$, are *Markov equivalent* if

$$A \perp_\mu B \mid C \; [\mathcal{G}_1] \Leftrightarrow A \perp_\mu B \mid C \; [\mathcal{G}_2],$$

for all $A, B, C \subseteq V$, and we let $[\mathcal{G}_1]$ denote the Markov equivalence class of $\mathcal{G}_1$. Every Markov equivalence class of DMGs has a unique *maximal element* Mogensen and Hansen (2020), i.e., there exists $\mathcal{G} \in [\mathcal{G}_1]$ such that $\mathcal{G}$ is a supergraph of all other graphs in $[\mathcal{G}_1]$.

For a DMG, $\mathcal{G}$, we will let $D(\mathcal{G})$ denote the *directed part* of $\mathcal{G}$, i.e., the DG obtained by deleting all bidirected edges from $\mathcal{G}$.

Proposition 1. Let $\mathcal{D} = (V, E)$ be a DG, and let $O \subseteq V$. Consider $\mathcal{G} = m(\mathcal{D}, O)$. For $\alpha, \beta \in O$ it holds that $\alpha \in$ an$_\mathcal{D}(\beta)$ if and only if $\alpha \in$ an$_{D(\mathcal{G})}(\beta)$. Furthermore, the directed part of $\mathcal{G}$ equals the parent graph of $\mathcal{D}$ on nodes O, i.e., $D(\mathcal{G}) = \mathcal{P}_O(\mathcal{D})$.

Proof. Note first that $\alpha \in$ an$_\mathcal{D}(\beta)$ if and only if $\alpha \in$ an$_\mathcal{G}(\beta)$ Mogensen and Hansen (2020). Ancestry is only defined by the directed edges, and it follows that $\alpha \in$ an$_\mathcal{G}(\beta)$ if and only if $\alpha \in$ an$_{D(\mathcal{G})}(\beta)$. For the second statement, the definition of the latent projection gives that there is a directed edge from α to β in $\mathcal{G}$ if and only if there is a directed path from α to β in $\mathcal{D}$ such that no nonendpoint node is in O. By definition, this is the parent graph, $\mathcal{P}_O(\mathcal{D})$. □

In words, the above proposition says that if $\mathcal{G}$ is a marginalization (done by latent projection) of $\mathcal{D}$, then the ancestor relations of $\mathcal{D}$ and $D(\mathcal{G})$ are the same among the observed nodes. It also says that our learning target, the parent graph, is actually the directed part of the latent projection on the observed nodes. In the next subsection, we use this to describe what is actually identifiable from the induced independence model of a graph.

1.1 MAXIMAL GRAPHS AND PARENT GRAPHS

Under faithfulness of the local independence model and the causal graph, we know that the maximal DMG is a correct representation of the local independence structure in the sense that it encodes exactly the local independences that hold in the local independence model. From the maximal DMG, one can use results on equivalence classes of DMGs to obtain every other DMG which encodes the observed local independences (Mogensen and Hansen, 2020) and from this graph one can find the parent graph as simply the directed part. However, it may require an infeasible number of tests to output such a maximal DMG. This is not surprising, seeing that the learning target encodes this complete information on local independences.

Assume that $\mathcal{D}_0 = (V, E)$ is the underlying causal graph and that $\mathcal{G}_0 = (O, F), O \subseteq V$ is the marginalized graph over the observed variables, i.e., the latent projection of $\mathcal{D}_0$. In principle, we would like to output $\mathcal{P}(\mathcal{D}_0) = D(\mathcal{G}_0)$, the directed part of $\mathcal{G}_0$. However, no algorithm can in general output this graph by testing only local independences as Markov equivalent DMGs may not have the same parent graph. Within each Markov equivalence class of DMGs, there is a unique maximal graph. Let $\bar{\mathcal{G}}$ denote the maximal graph which is Markov equivalent of $\mathcal{G}_0$. The DG $D(\bar{\mathcal{G}})$ is a supergraph of $D(\mathcal{G}_0)$ and we will say that a learning algorithm is complete if it is guaranteed to output $D(\bar{\mathcal{G}})$ as no algorithm testing local independence only can identify anything more than the equivalence class.

2 COMPLETE LEARNING

The CS algorithm provides sound learning of the parent graph of a general DMG under the assumption of ancestral faithfulness. For a subclass of DMGs, the algorithm actually provides complete learning. It is of interest to find sufficient graphical conditions to ensure that the algorithm removes an edge $\alpha \to \beta$ which is not in the true parent graph. In this section, we state and prove one such condition which can be understood as 'the true parent set is always found for unconfounded processes'. We let $\mathcal{D}$ denote the output of the CS algorithm.

Proposition 2. If $\alpha \not\to_{\mathcal{G}_0} \beta$ and there is no $\gamma \in V \setminus \{\beta\}$ such that $\gamma \leftrightarrow_{\mathcal{G}_0} \beta$, then $\alpha \not\to_{\mathcal{D}} \beta$.

Proof. Let $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_N$ denote the DGs that are constructed when running the algorithm by sequentially removing edges, starting from the complete DG, $\mathcal{D}_1$. Consider a connecting walk from α to β in $\mathcal{G}_0$. It must be of the form $\alpha \sim \ldots \sim \gamma \to \beta$, $\gamma \neq \alpha$. Under ancestral faithfulness, the edge $\gamma \to \beta$ is in $\mathcal{D}$, thus $\gamma \in \mathrm{pa}_{\mathcal{D}_i}(\beta)$ for all $\mathcal{D}_i$ that occur during the algorithm, and therefore when $\langle \alpha, \beta \mid \mathrm{pa}_{\mathcal{D}_i}(\beta) \setminus \{\alpha\} \rangle$ is tested, the walk is closed. Any walk from α to β is of this form, thus also closed, and we have that $\alpha \perp_{\mu} \beta \mid \mathrm{pa}_{\mathcal{D}_i}(\beta)$ and therefore $\langle \alpha, \beta \mid \mathrm{pa}_{\mathcal{D}_i}(\beta) \setminus \{\alpha\} \rangle \in \mathcal{I}$. The edge $\alpha \to_{\mathcal{D}_i} \beta$ is removed and thus absent in the output graph, $\mathcal{D}$. □

3 ANCESTRY PROPAGATION

We state Subalgorithm 4 here.

```
input  : a local independence oracle for $\mathcal{I}^O$ and a
         DG, $\mathcal{D} = (O, E)$
output : a DG on nodes $O$
initialize $E_r = \emptyset$ as the empty edge set;
foreach $(\alpha, \beta, \gamma) \in V \times V \times V$ such that $\alpha, \beta, \gamma$
  are all distinct do
    if $\alpha \sim_{\mathcal{D}} \beta$, $\beta \to_{\mathcal{D}} \gamma$, and $\alpha \not\to_{\mathcal{D}} \gamma$ then
        if $\langle \alpha, \gamma \mid \emptyset \rangle \in \mathcal{I}^O$ then
            update $E_r = E_r \cup \{\beta \to \gamma\}$;
        end
    end
end
Update $\mathcal{D} = (V, E \setminus E_r)$;
return $\mathcal{D}$
```

Subalgorithm 4: Ancestry propagation

Composing Subalgorithm 1, Subalgorithm 4, and Subalgorithm 2 is referred to as the causal screening, ancestry propagation (CSAP) algorithm. If we use Subalgorithm 3 instead of Subalgorithm 4, we call it the CSAPC algorithm (C for cheap as this does not entail any additional independence tests compared to CS).

4 APPLICATION AND SIMULATIONS

In this section, we provide some additional details about the c. elegans neuronal network and the simulations.

4.1 C. ELEGANS NEURONAL NETWORK

For each connection between two neurons a different number of synapses are present (ranging from 1 to 37). We only consider connections with more than 4 synapses when we define the true underlying network. When sampling the subnetworks, highly connected neurons were sampled with higher probability to avoid a fully connected subnetwork when marginalizing.

4.2 COMPARISON OF ALGORITHMS

As noted in the main paper, the dFCI algorithm solves a strictly harder problem. By using the additional graph theory in the supplementary material, we can understand the output of the dFCI algorithm as a supergraph of the maximal DMG, $\bar{\mathcal{G}}$. There is also a version of the dFCI which is guaranteed to output not only a supergraph of $\bar{\mathcal{G}}$, but the graph $\bar{\mathcal{G}}$ itself. Clearly, from the output of the dFCI algorithm, one can simply take the directed part of the output and this is a supergraph of the underlying parent graph.

5 PROOFS

In this section, we provide the proofs of the result in the main paper.

Proof of Proposition 5. Let $\mathcal{D}$ denote the causal graph. Assume first that $\alpha \not\to_{\mathcal{D}} \beta$. Then $g^{\beta\alpha}$ is identically zero over the observation interval, and it follows directly from the functional form of λ_t^β that $\alpha \not\to \beta \mid V \setminus \{\alpha\}$. This shows that the local independence model satisfies the pairwise Markov property with respect to $\mathcal{D}$.

If instead $g^{\beta\alpha} \neq 0$ over J, there exists $r \in J$ such that $g^{\beta\alpha}(r) \neq 0$. From continuity of $g^{\beta\alpha}$ there exists a compact interval of positive measure, $I \subseteq J$, such that $\inf_{s \in I}(g^{\beta\alpha}(s)) \geq g_{\min}^{\beta\alpha}$ and $g_{\min}^{\beta\alpha} > 0$. Let i_0 and i_1 denote the endpoints of this interval, $i_0 < i_1$. We consider now the events

$$D_k = (N_{T-i_0}^\alpha - N_{T-i_1}^\alpha = k, N_T^\gamma = 0 \text{ for all } \gamma \in V \setminus \{\alpha\})$$

$k \in \mathbb{N}_0$. Then under Assumption 4, for all k

$$\lambda_T^\beta \mathbb{1}_{D_k} \geq \mathbb{1}_{D_k} \int_I g^{\beta\alpha}(T-s)\, \mathrm{d}N_s^\alpha \geq g_{\min}^{\beta\alpha} \cdot k \cdot \mathbb{1}_{D_k}.$$

Assume for contradiction that β is locally independent of α given $V \setminus \{\alpha\}$. Then $\lambda_T^\beta = \mathrm{E}(\lambda_T^\beta \mid \mathcal{F}_T^V) = \mathrm{E}(\lambda_T^\beta \mid \mathcal{F}_T^{V\setminus\{\alpha\}})$ is constant on $\cup_k D_k$ and furthermore $\mathrm{P}(D_k) > 0$ for all k. However, this contradicts the above inequality when $k \to \infty$. □

Proof of Proposition 12. Let $\mathcal{D}$ denote the DG which is output by the algorithm. We should then show that $\mathcal{P}(\mathcal{D}_0) \subseteq \mathcal{D}$. Assume that $\alpha \to_{\mathcal{P}(\mathcal{D}_0)} \beta$. In this case, there is a directed path from α to β in $\mathcal{D}_0$ such that no nonendpoint node on this directed walk is in O (the observed coordinates). Therefore for any $C \subseteq O \setminus \{\alpha\}$ there exists a directed μ-connecting walk from α to β in $\mathcal{D}_0$ and by ancestral faithfulness it follows that $\langle \alpha, \beta \mid C \rangle \notin \mathcal{I}$. The algorithm starts from the complete directed graph, and the above means that the directed edge from α to β will not be removed. □

Proof of Corollary 13. Consider some directed path from α to β in $\mathcal{D}_0$ on which no node is in C. Then there is also a directed path from α to β on which no nodes is in C in the graph $\mathcal{P}(\mathcal{D}_0)$, and therefore also in the output graph using Proposition 12. □

Proof of Proposition 15. Assume that there is a μ-connecting walk from α to β given $\{\beta\}$. If this walk has no colliders, then it is a directed trek, or can be reduced to one. Otherwise, assume that γ is the collider which is the closest to the endpoint α. Then $\gamma \in \text{an}(\beta)$, and composing the subwalk from α to γ with the directed path from γ to β gives a directed trek, or it can be reduced to one. On the other hand, assume there is a directed trek from α to β. This is μ-connecting from α to β given $\{\beta\}$. □

Proof of Proposition 17. Assume $\beta \to_{\mathcal{P}(\mathcal{D}_0)} \gamma$. Subalgorithms 1 and 2 are both simple screening algorithms, and they will not remove this edge. Assume for contradiction that $\beta \to \gamma$ is removed by Subalgorithm 3. Then there must exist $\alpha \neq \beta, \gamma$ and a directed trek from α to β in $\mathcal{D}_0$. On this directed trek, γ does not occur as this would imply a directed trek either from α to γ or from β to α, thus implying $\alpha \to_{\mathcal{D}} \gamma$ or $\beta \to_{\mathcal{D}} \alpha$, respectively ($\mathcal{D}$ is the output graph of Subalgorithm 1). As γ does not occur on the trek, composing this trek with the edge $\beta \to \gamma$ would give a directed trek from α to γ. By faithfulness, $\langle \alpha, \gamma \mid \gamma \rangle \notin \mathcal{I}$, and this is a contradiction as $\alpha \to \gamma$ would not have been removed during Subalgorithm 1.

We consider instead CSAP. Assume for contradiction that $\beta \to \gamma$ is removed during Subalgorithm 4. There exists in $\mathcal{D}_0$ either a directed trek from α to β or a directed trek from β to α. If γ is on this trek, then γ is not μ-separated from α given the empty set (recall that there are loops at all nodes, therefore also at γ), and using faithfulness we conclude that γ is not on this trek. Composing it with the edge $\beta \to \gamma$ would give a directed trek from α to γ and using faithfulness we obtain a contradiction. □

References

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

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.

Christopher Meek. Toward learning graphical and causal process models. In *CI'14 Proceedings of the UAI 2014 Conference on Causal Inference: Learning and Prediction*, 2014.

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

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. Technical Report R-150, University of California, Los Angeles, 1991.

This thesis as a commutative diagram

As a final component in this chapter, we consider the commutative diagram below to summarize the content of the thesis. Our starting point is a distribution of a stochastic process, P_X. Using the concept of local independence, we obtain a set of independences, $\mathcal{I}$, that hold in the distribution. Often we will assume that we only observed some of the coordinate processes, i.e., we have access to the marginal distribution P_{X^O}. From a distribution, P_X, one can define a graph $\mathcal{D}$ such that $\mathcal{D}$ in a certain sense encodes the independence structure (or some of it) in the distribution. In Paper **B** we saw that in the case of a regular Ornstein-Uhlenbeck process, one can read off $\mathcal{D}$ from the matrix M in the drift of the process. We also needed a graph for representing the system under partial observation, $\mathcal{G}$ in the diagram. Using μ-separation, we obtained a separation model. The global Markov property shows that in certain classes of processes, we have $\mathcal{I}(\mathcal{D}) \subseteq \mathcal{I}$ and therefore $\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}^O$. Under Markov and faithfulness assumptions, such that $\mathcal{I}(\mathcal{G}) = \mathcal{I}^O$, we considered the learning problem in which one tries to learn something about $\mathcal{G}$ from $\mathcal{I}^O$.

$$\begin{tikzcd}
P_{X^O} \arrow[d, "\text{li}"'] & P_X \arrow[l] \arrow[r, "M"] \arrow[d, "\text{li}"'] & \mathcal{D} \arrow[r, "\text{latent proj}"] \arrow[d, "\mu\text{-sep}"] & \mathcal{G} \arrow[d, "\mu\text{-sep}"] \\
\mathcal{I}^O & \mathcal{I} \arrow[l] & \mathcal{I}(\mathcal{D}) \arrow[r] & \mathcal{I}(\mathcal{G})
\end{tikzcd}$$

Bibliography

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

Theodore B. Achacoso, Victor Fernandez, Duc C. Nguyen, and William S. Yamamoto. Computer representation of the synaptic connectivity of Caenorhabditis elegans. In *Proceedings of the Annual Symposium on Computer Applications in Medical Care (SCAMC)*, 1989.

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

Venkat Chandrasekaran, Nathan Srebro, and Prahladh Harsha. Complexity of inference in graphical models. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2008.

David Maxwell Chickering, David Heckerman, and Christopher Meek. Large-sample learning of Bayesian networks is NP-hard. *Journal of Machine Learning Research*, 5, 2004.

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 (UAI)*, 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.

Steven J. Cook, Travis A. Jarrell, Christopher A. Brittin, Yi Wang, Adam E. Bloniarz, Maksim A. Yakovlev, Ken C. Q. Nguyen, Leo T.-H. Tang, Emily A. Bayer, Janet S. Duerr, Hannes E. Bülow, Oliver Hobert, David H. Hall, and Scott W. Emmons. Whole-animal connectomes of both Caenorhabditis elegans sexes. *Nature*, 571(7763), 2019.

Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. The MIT Press, 3rd edition, 2009.

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

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.

Michael Eichler. *Causality: Statistical Perspectives and Applications*, chapter 22 (Causal Inference in Time Series Analysis), page 327–354. John Wiley & Sons, Ltd, 2012.

Michael Eichler. Causal inference with multiple time series: Principles and problems. *Philosophical Transactions of the Royal Society*, 371(1997):1–17, 2013.

Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. 1979.

Clark Glymour and Gregory F. Cooper, editors. *Computation, Causation, and Discovery*. AAAI Press/MIT Press, 1999.

Oded Goldreich. *P, NP, and NP-Completeness: The Basics of Computational Complexity*. Cambridge University Press, 2010.

Jonathan L. Gross, Jay Yellen, and Ping Zhang, editors. *Handbook of Graph Theory*. Chapman & Hall/CRC, 2013.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.

Jan T.A. Koster. On the validity of the Markov interpretation of path diagrams of Gaussian structural equations systems with correlated errors. *Scandinavian Journal of Statistics*, 26:413–431, 1999.

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

Marloes Maathuis, Mathias Drton, Steffen Lauritzen, and Martin Wainwright, editors. *Handbook of Graphical Models*. CRC Press, 2018.

Christopher Meek. Finding a path is harder than finding a tree. *Journal of Artificial Intelligence Research*, 15:383–389, 2001.

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. Causal screening in dynamical systems. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2020. (to appear).

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

Søren Wengel Mogensen and Niels Richard Hansen. Graphical modeling of stochastic processes driven by correlated errors. 2020b.

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.

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

Thomas S. Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 2003.

Thomas S. Richardson and Peter Spirtes. Ancestral graph Markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.

Thomas S. Richardson, Robin J. Evans, James M. Robins, and Ilya Shpitser. Nested Markov properties for acyclic directed mixed graphs. 2017. URL `https://arxiv.org/pdf/1701.06686.pdf`.

Kayvan Sadeghi. Stable mixed graphs. *Bernoulli*, 19(5B):2330–2358, 2013.

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

Bernhard Schölkopf, Dominik Janzing, and Jonas Peters. *Elements of Causal Inference*. The MIT Press, 2017.

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

Michael Sipser. *Introduction to the theory of computation*. Thomson Course Technology, 3rd edition, 2013.

Peter Spirtes and Kun Zhang. *Handbook of Graphical Models*, chapter 18 (Search for Causal Models), pages 439–469. CRC Press, 2018.

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

L. R. Varshney, B. L. Chen, E. Paniagua, D. H. Hall, and D. B. Chklovskii. Structural properties of the Caenorhabditis elegans neuronal network. *PLOS Computational Biology*, 7(2), 2011.

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. Technical Report R-150, University of California, Los Angeles, 1991.

Thomas Verma and Judea Pearl. Deciding morality of graphs is NP-complete. In *Proceedings of the 9th Conference on Uncertainty in Artificial Intelligence (UAI)*, 1993.

J.G. White, E. Southgate, J.N. Thomson, and S. Brenner. The structure of the nervous system of the nematode Caenorhabditis elegans. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 314(1165), 1986.