Supplementary material for Causal screening for dynamical systems

Søren Wengel Mogensen
Department of Mathematical Sciences
University of Copenhagen
Copenhagen, Denmark
swengel@math.ku.dk

This document contains additional graph theory, results, and definitions, as well as the proofs of the main paper.

1 Graph theory

The additional graph theory is useful in the proofs, and we can also give an alternative definition of the parent graph using a graphical marginalization operation that we will describe.

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 of the types $\alpha \rightarrow \beta$ and $\alpha \leftarrow \beta$ are *directed*, and that $\alpha \leftrightarrow \beta$ is *bidirected*.

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 $\mathrm{an}(\beta)$ denote the sets of nodes that are ancestors of β.

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 [1–4].

Definition S1 (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, $\langle \alpha, e_1, \ldots, e_n, \beta \rangle$, is μ-connecting given C if $\alpha \notin C$, the edge e_n has a head at β, every collider on the walk is in $\mathrm{an}(C)$ 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.

We use the class of DGs use to represent the underlying, data-generating structure. When only parts of the causal system is observed, the class of DMGs can be used for representing marginalized DGs [4]. This can be done using *latent projection* [5, 4] 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 [4] 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

$$A \perp_\mu B \mid C \; [\mathcal{G}_1] \Leftrightarrow A \perp_\mu B \mid C \; [\mathcal{G}_2],$$

Preprint. Under review.

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* [4], 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 S2. 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 $\alpha \in \text{an}_\mathcal{D}(\beta)$ if and only if $\alpha \in \text{an}_{D(\mathcal{G})}(\beta)$. 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 $\alpha \in \text{an}_\mathcal{D}(\beta)$ if and only if $\alpha \in \text{an}_\mathcal{G}(\beta)$ [4]. Ancestry is only defined by the directed edges, and it follows that $\alpha \in \text{an}_\mathcal{G}(\beta)$ if and only if $\alpha \in \text{an}_{D(\mathcal{G})}(\beta)$. 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 [4] 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 will simply 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 S3. 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 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 $\gamma \in \text{pa}_{\mathcal{D}_i}(\beta)$ for all $\mathcal{D}_i$ that occur during the algorithm, and therefore when $\langle \alpha, \beta \mid \text{pa}_{\mathcal{D}_i}(\beta) \setminus \{\alpha\} \rangle$ 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 $\alpha \rightarrow_{\mathcal{D}_i} \beta$ 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* α, β, γ *are all distinct* **do**
  if $\alpha \sim_{\mathcal{D}} \beta$*,* $\beta \rightarrow_{\mathcal{D}} \gamma$*, and* $\alpha \not\rightarrow_{\mathcal{D}} \gamma$ **then**
    if $\langle \alpha, \gamma \mid \emptyset \rangle \in \mathcal{I}^O$ **then**
      update $E_r = E_r \cup \{\beta \rightarrow \gamma\}$;
    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.

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 $\alpha \not\rightarrow_{\mathcal{D}} \beta$. Then $g^{\beta\alpha}$ is identically zero over the observation interval, and it follows directly from the functional form of λ_t^β that $\alpha \not\rightarrow \beta \mid V \setminus \{\alpha\}$. 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 $\inf_{s\in I}(g^{\beta\alpha}(s)) \geq g^{\beta\alpha}_{\min}$ and $g^{\beta\alpha}_{\min} > 0$. Let i_0 and i_1 denote the endpoints of this interval, $i_0 < i_1$. We consider now the events

$$D_k = (N^\alpha_{T-i_0} - N^\alpha_{T-i_1} = k, N^\gamma_T = 0 \text{ for all } \gamma \in V \setminus \{\alpha\}) \tag{1}$$

$k \in \mathbb{N}_0$. Then under Assumption 4, for all k

$$\lambda^\beta_T \mathbb{1}_{D_k} \geq \mathbb{1}_{D_k} \int_I g^{\beta\alpha}(T-s)\,\mathrm{d}N^\alpha_s \geq g^{\beta\alpha}_{\min} \cdot k \cdot \mathbb{1}_{D_k}.$$

Assume for contradiction that β is locally independent of α given $V \setminus \{\alpha\}$. Then $\lambda^\beta_T = \mathrm{E}(\lambda^\beta_T \mid \mathcal{F}^V_T) = \mathrm{E}(\lambda^\beta_T \mid \mathcal{F}^{V\setminus\{\alpha\}}_T)$ 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 11. 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 12. 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 11. □

Proof of Proposition 14. 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 $\gamma \in \mathrm{an}(\beta)$, and composing the subwalk from α to γ with the directed path from γ to β gives a directed trek. 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 16. 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 γ, thus implying $\alpha \to_{\mathcal{D}} \gamma$ or $\beta \to_{\mathcal{D}} \alpha$, respectively ($\mathcal{D}$ is the output graph). 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

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

[2] 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.

[3] 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.

[4] Søren Wengel Mogensen and Niels Richard Hansen. Markov equivalence of marginalized local independence graphs. 2019. To appear in the Annals of Statistics.

[5] Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. Technical Report R-150, University of California, Los Angeles, 1991.