Supplementary Material: Ordering-Based Causal Structure Learning in the Presence of Latent Variables

Daniel Irving Bernstein* MIT

Basil Saeed* MIT

Chandler Squires* MIT

Caroline Uhler MIT

A Graph Theory

This section provides additional graph-theoretic notations that are standard in the literature and are provided for ease of access. Let $G = (V, D, B)$ be a graph. If there is any edge between i and j, they are called *adjacent* which we may denote $i \sim j$. Otherwise they are called *non-adjacent* and we write $i \not\sim j$. We will use $\circ$ as a "wildcard" for edge marks, i.e. $i \circ\!\!\rightarrow j$ denotes that either $i \rightarrow j$ or $i \leftrightarrow j$. We will use subscripts on these vertex relations as a shorthand way to indicate the presence or absence of an edge, or the presence of a particular kind of edge. For example, $i \leftrightarrow_G j$ and $k \not\sim_G l$ respectively indicate that G has a bidirected edge between i and j, and no edge between k and l. A graph with only directed edges is called a *directed graph*.

A *path* $\gamma = \langle v_1, v_2, \ldots, v_k \rangle$ is a sequence of distinct nodes that such that v_i and v_{i+1} are adjacent. A *cycle* is a path together with any type of edge between v_k and $v_{k+1} = v_1$. A path or a cycle is called *directed* if all edges are directed toward later nodes, i.e. $v_i \rightarrow v_{i+1}$.

We extend the notation $\mathrm{pa}_G(i), \mathrm{sp}_G(i)$, and $\mathrm{an}_G(i)$ to allow arguments that are subsets of vertices by taking unions. For example, when $S \subseteq V$, we have

$$\mathrm{pa}_G(S) := \cup_{i \in S} \mathrm{pa}_G(i).$$

We add an asterisk to denote that the arguments are not included in the set, e.g.

$$\mathrm{pa}^*_G(S) := \mathrm{pa}_G(S) \setminus S.$$

The *colliders* on a path γ are the nodes where two arrowheads meet, i.e., v_i is a collider if $v_{i-1} \circ\!\!\rightarrow v_i \leftarrow\!\!\circ\, v_{i+1}$. A triple of nodes (i, j, k) is called a *v-structure* if j is a collider on the path $\langle i, j, k \rangle$ and $i \not\sim k$.

B Proof of Proposition 1

We will prove Proposition 1 via a sequence of intermediate Lemmas. Since our goal is to prove that all the m-separation statements of a given DMAG are satisfied by a given $\mathbb{P}$, it will be helpful to have the following lemma which reduces the number of m-separation statements we must consider.

Lemma 1. *Let G^* and H be DMAGs. Then $G^* \leq H$ if and only if whenever $i \not\sim_H j$, i is $\mathrm{an}_H(\{i,j\})$-separated from j in G^*, i.e. $i \perp\!\!\!\perp_{G^*} j \mid \mathrm{an}_H(\{i,j\})$.*

Proof. This is an immediate consequence of Theorem 3 in (Sadeghi and Lauritzen, 2014). □

Throughout the rest of this section, let it be understood that G^* is a DMAG that is restricted-faithful to some fixed joint distribution $\mathbb{P}$. We will not repeat this assumption. Moreover, we will suppress $\mathbb{P}$ in our notation and write G_π instead of G_π and $AG(\pi)$ instead of $AG(\pi, \mathbb{P})$. Also, note that when H is a DMAG, $\mathrm{po}(H) = \mathrm{po}(\overline{H})$ since $\overline{H}$ is obtained from H by adding only bidirected edges (Richardson and Spirtes, 2002). We will make repeated tacit use of this fact.

Lemma 2. *Let π be a partial order on the random variables of $\mathbb{P}$ such that $G_\pi = \overline{AG(\pi)}$. Then G_π is an IMAP of $\mathbb{P}$.*

Proof. Lemma 1 implies that it suffices to show that whenever $i \not\sim_{G_\pi} j$, $X_i \perp\!\!\!\perp_{\mathbb{P}} X_j \mid X_{\mathrm{an}^*_{G_\pi}(i,j)}$. So assume $i \not\sim_{G_\pi} j$. Since $G_\pi = \overline{AG(\mathrm{po}(AG(\pi)))}$, $i \not\sim_{G_\pi} j$ implies $X_i \perp\!\!\!\perp_{\mathbb{P}} X_j \mid X_{\mathrm{pre}^*_{\mathrm{po}(AG(\pi))}(i,j)}$. But now we are done since $\mathrm{pre}^*_{\mathrm{po}(AG(\pi))}(i,j) = \mathrm{an}^*_{AG(\pi)}(i,j) = \mathrm{an}^*_{\overline{AG(\pi)}}(i,j)$ and we are assuming $G_\pi = \overline{AG(\pi)}$. □

Lemma 3. *Let π be a partial order on the random variables of $\mathbb{P}$. Then $\mathrm{po}(G_\pi) = \mathrm{po}(AG(\pi))$.*

Proof. We must show

$$\mathrm{po}(AG(\mathrm{po}(AG(\pi)))) = \mathrm{po}(AG(\pi))$$

If $i \leq j$ in $\mathrm{po}(AG(\mathrm{po}(AG(\pi))))$, then there exists a directed path $i = i_0 \rightarrow \cdots \rightarrow i_k = j$ in $AG(\mathrm{po}(AG(\pi)))$ and so $i = i_0 \leq \cdots \leq i_k = j$ in $\mathrm{po}(AG(\pi))$.

We now proceed to show that if $i \leq j$ in $\mathrm{po}(AG(\pi))$, then the same is true in $\mathrm{po}(AG(\mathrm{po}(AG(\pi))))$. We do this by showing that if $i \rightarrow_{AG(\pi)} j$, then $i \rightarrow_{AG(\mathrm{po}(AG(\pi)))} j$. So for the sake of contradiction, assume that $i \rightarrow_{AG(\pi)} j$ but not $i \rightarrow_{AG(\mathrm{po}(AG(\pi)))} j$. By the definition of AG, this implies that $i \not\sim_{AG(\mathrm{po}(AG(\pi)))} j$ and so i is m-separated from j given $\mathrm{an}^*_{AG(\pi)}(i,j)$ in G^*. But $i \rightarrow_{AG(\pi)} j$ implies that i is m-connected to j given $\mathrm{pre}^*_\pi(i,j)$ in G^*. Let P be an m-connecting path from i to j given $\mathrm{pre}^*_\pi(i,j)$ in G^*. Since $\mathrm{an}^*_{AG(\pi)}(i,j) \subseteq \mathrm{pre}^*_\pi(i,j)$, we can write

$$\mathrm{pre}^*_\pi(i,j) = \mathrm{an}^*_{AG(\pi)}(i,j) \cup S$$

for some nonempty set S, disjoint from $\mathrm{an}^*_{AG(\pi)}(i,j)$. Since i is m-separated from j given $\mathrm{an}^*_{AG(\pi)}(i,j)$ in G^*, P must contain a collider with a descendent in S, but no descendant in $\mathrm{an}^*_{AG(\pi)}(i,j)$. Let d be such a collider that is closest to j along P and let s be a $\mathrm{po}(G^*)$-minimal descendent of d from S.

We now construct a path Q in G^* that m-connects j and s given $\mathrm{pre}^*_\pi(j,s)$. Since $S \subseteq \mathrm{pre}_\pi(j)$, this would imply existence of the edge $s \rightarrow_{AG(\pi)} j$, contradicting $s \in S$. If $s = d$, we let Q be the subpath of P from j to s. Otherwise, we let Q be obtained by concatenating the subpath of P from j to d, followed by a directed path from d to s. Since P is m-connecting given $\mathrm{pre}^*_\pi(i,j)$ and $i, s \leq j$ in π, it follows that when Q is a subpath of P, Q is m-connecting given $\mathrm{pre}^*_\pi(j,s)$. When Q additionally has a directed path from d to s, Q is m-connecting given $\mathrm{pre}^*_\pi(j,s)$ since the non-P segment has no colliders, and assumptions on d and $\mathrm{po}(G^*)$-minimality of s imply that no element of this segment is in $\mathrm{an}^*_{AG(\pi)}(s,j)$. □

Proof of Proposition 1. Define $\tau := \mathrm{po}(AG(\pi))$. Since $\mathrm{po}(H) = \mathrm{po}(\overline{H})$ for any DMAG H, we have

$$G_\tau = \overline{AG(\mathrm{po}(G_\pi))}.$$

Lemma 3 implies that this is equal to $\overline{AG(\mathrm{po}(AG(\pi)))}$, which is equal to both G_π and $\overline{AG(\tau)}$. Thus we have shown that $G_\pi = G_\tau = \overline{AG(\tau)}$ and so Lemma 2 implies that G_π is an IMAP of $\mathbb{P}$.

We now show that G_π is a *minimal* IMAP of $\mathbb{P}$, i.e. that removing any edge results in a directed ancestral graph that is either not maximal, or not an IMAP of $\mathbb{P}$. Let i, j be such that $i \sim_{G_\pi} j$ and let G' be the graph obtained from G_π by removing the edge between i and j. If G' is still maximal, then Lemma 1 implies that i is m-separated from j given $\mathrm{an}^*_{G'}(i,j)$ in G'. If $G^* \leq G'$, then i is m-separated from j given $\mathrm{an}^*_{G'}(i,j)$ in G^*. Note that $\mathrm{an}^*_{G'}(i,j) = \mathrm{an}^*_{G_\pi}(i,j)$, and that Lemma 2 implies that $\mathrm{an}^*_{G_\pi}(i,j) = \mathrm{pre}^*_{\mathrm{po}(AG(\pi))}(i,j)$. But if i were $\mathrm{pre}^*_{\mathrm{po}(AG(\pi))}(i,j)$-separated from j in G^*, then $X_i \perp\!\!\!\perp_{\mathbb{P}} X_j \mid X_{\mathrm{pre}^*_{\mathrm{po}(AG(\pi))}(i,j)}$ and so $i \not\sim_{AG(\pi)} j$. This would imply that $AG(\pi)$ is a subgraph of G'. Since G' is maximal, G_π would be a subgraph as well contradicting $i \sim_{G_\pi} j$. □

C Proof of Theorem 1

We begin by proving the following lemma, which extends classic results for the case of DAGs and deals with discriminating paths.

Lemma 4. *Let G^* and H be DMAGs and let $\mathbb{P}$ be a distribution that is Markov to both G^* and H.If $\mathbb{P}$ is adjacency-faithful to G^*, then*

(a) $\mathrm{skel}(G^*) \subseteq \mathrm{skel}(H)$.

If $\mathbb{P}$ is furthermore orientation-faithful to G^*, then

(b) If $i \circ\!\!\rightarrow k \leftarrow\!\!\circ j$ is a v-structure in G^*, then either $i \circ\!\!\rightarrow k \leftarrow\!\!\circ j$ is a v-structure in H or $i \sim_H j$.

(c) If $i \circ\!\!\rightarrow k \leftarrow\!\!\circ j$ is a v-structure in H, then either $i \circ\!\!\rightarrow k \leftarrow\!\!\circ j$ is a v-structure in G^*, or $i \not\sim_{G^*} k$ or $j \not\sim_{G^*} k$.

Finally, if $\mathbb{P}$ is also discriminating-faithful to G^*, then

(d) If $\gamma := \langle i, \ldots, k, j\rangle$ is a discriminating path in both H and G^*, then k is a collider in γ in H iff k is a collider in γ in G^*.

Proof. (a) If $i \not\sim_H j$, then by the pairwise Markov property (Richardson and Spirtes, 2002), $X_i \perp\!\!\!\perp_{\mathbb{P}} X_j \mid X_{\mathrm{an}^*_H(i,j)}$, and by adjacency-faithfulness, $i \not\sim_{G^*} j$ in G^*.

(b) Let $i \not\sim_H j$, so $X_i \perp\!\!\!\perp_{\mathbb{P}} X_j \mid X_{\mathrm{an}^*_H(\{i,j\})}$. Suppose k is a parent of either i or j. Since $k \in \mathrm{an}^*_H(\{i,j\})$, i is m-connected to j in G^* given $\mathrm{an}^*_H(\{i,j\})$ by the path $i \circ\!\!\rightarrow k \leftarrow\!\!\circ j$, and thus $X_i \not\perp\!\!\!\perp_{\mathbb{P}} X_j \mid X_{\mathrm{an}_H(\{i,j\})}$ by orientation faithfulness. Hence, H is not an I-MAP of $\mathbb{P}$.

(c) Suppose $i \sim_G k$ and $j \sim_G k$. We have $X_i \perp\!\!\!\perp_{\mathbb{P}} X_j \mid X_{\mathrm{an}^*_H(\{i,j\})}$, and thus by orientation faithfulness i and j are m-separated given $\mathrm{an}^*_H(\{i,j\})$ in G^*. Since H is ancestral, $k \notin \mathrm{an}^*_H(\{i,j\})$. Thus, to ensure the required m-separation in G, k must be a collider in G on the path $i - k - j$.

(d) Assume $\gamma = \langle i, C_1, \ldots, C_l, k, j\rangle$. If k is a non-collider in γ in G^*, then i is m-connected to j given S for every S containing $C_1, \ldots, C_l$ but not k. Discriminating faithfulness implies $X_i \not\perp\!\!\!\perp_{\mathbb{P}} X_j \mid X_S$ for every such S. Then k must also be a non-collider in γ in H,

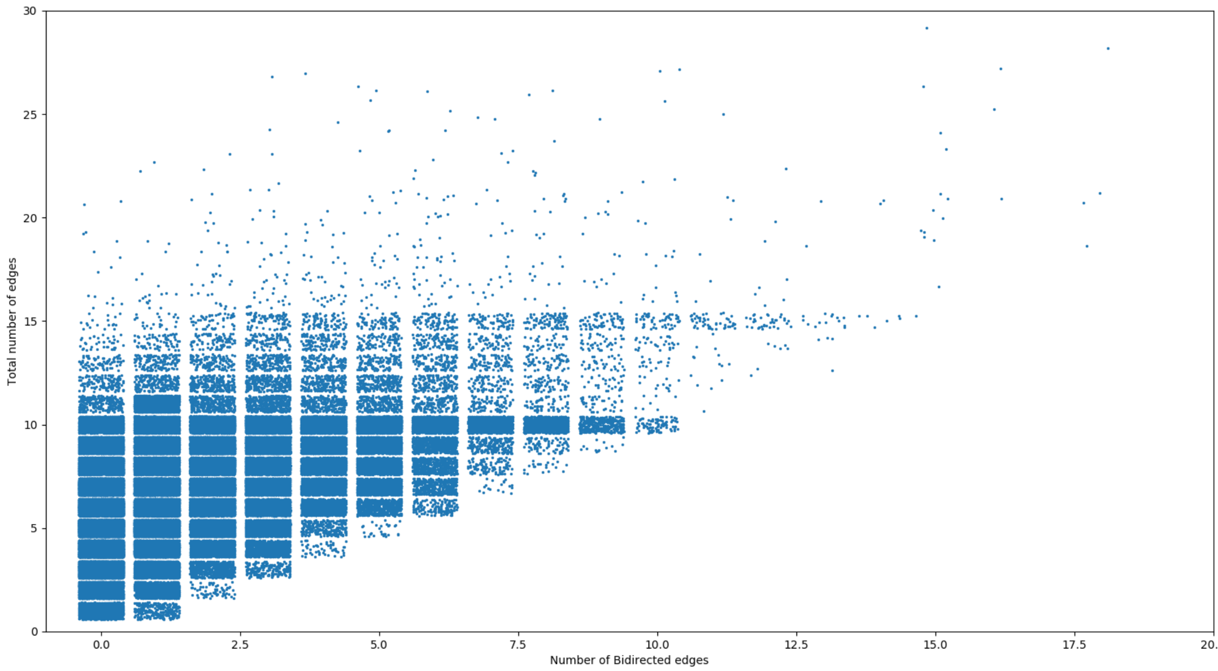

Figure 1: A scatter plot of the number of edges of the graphs that we tested the oracle version of our algorithm on. The plot includes over 200,000 points, representing graphs with varying number of bidirected edges and total number of edges.

since otherwise there would exist some S containing $C_1, \dots, C_l$ but not k such that i is m-separated from j given S in H^*, contradicting $\mathcal{I}(H) \subseteq \mathcal{I}(\mathbb{P})$. If K is a collider in γ in G^*, then i is m-connected to j given S for every S containing $C_1, \dots, C_l, k$. Again, discriminating faithfulness implies $X_i \not\perp\!\!\!\perp_{\mathbb{P}} X_j \mid X_S$ for every such S. Then K must also be a collider in γ in H, since otherwise there would exist some S containing $C_1, \dots, C_l, k$ such that i is m-separated from j given S in H^*. □

We proceed to proving the theorem.

Proof of Theorem 1. (a) is implied by Lemma 4(a).

Since restricted faithfulness implies adjacency faithfulness, $\text{skel}(G) = \text{skel}(G^*)$. It remains to show that G and G^* have the same v-structures, and that if γ is a discriminating path for k in both G and G^*, then k is a collider on γ in G iff it is a collider on γ in G^*.

Equality of skeletons together with Lemma 4(b) and (c) imply that G and H have the same v-structures. If $\gamma := \langle i, C_1, \dots, C_l, k, j\rangle$ is a discriminating path in both G^* and G, then Lemma 4(d) implies that k is a collider in γ in G^* iff k is a collider in γ in G. □

D Proof of Proposition 2

Proof. It is sufficient to show this for $G = G^*$, since Markov equivalence implies that $\mathcal{I}(G) = \mathcal{I}(G^*)$. Suppose $G = (V, D, B)$. Let $\pi = \text{po}(G)$. We have already shown that G_π is an IMAP. Therefore, it is sufficient to show the converse, i.e., that if $X_i \perp\!\!\!\perp_{\mathbb{P}} X_j \mid S$ then $i \perp\!\!\!\perp_{G_\pi} j \mid S$.

By Theorem 4.2 of Richardson and Spirtes (2002), for any $i, j \in V$ adjacent, $i \not\perp\!\!\!\perp_{G_\pi} j \mid \text{an}^*_{G_\pi}(i,j)$. The faithfulness condition would then imply that $X_i \not\perp\!\!\!\perp_{\mathbb{P}} X_j | X_{\text{pre}^*_\pi(i,j)}$. □

E Conjecture Simulations

In figure 1, we display a scatter plot of the number of edges of the graphs that we tested our algorithm on, without failure. The plot includes over 200,000 points, corresponding to 200,000 generated graphs of various parameters. For each of these, graphs, we have tested the oracle version of our algorithm, i.e., $\mathcal{I}(\mathbb{P}) = \mathcal{I}(G^*)$, and it converged to a graph in the Markov equivalence class of the true graph. We have not found a single counterexample to the conjecture thus far.

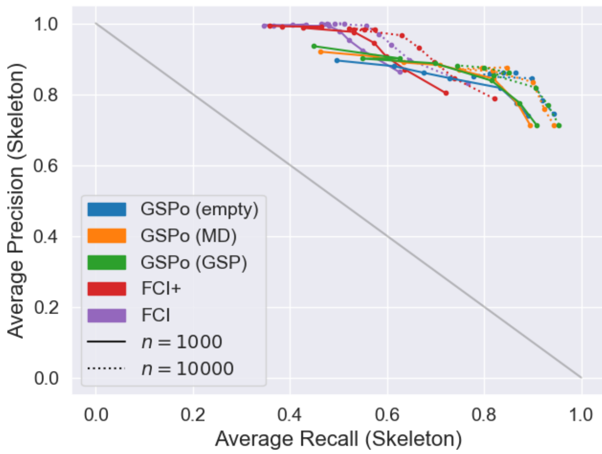

Figure 2: Average performance over 100 MAGs for each algorithm, when $p = 50$, $K = 12$, and $s = 3$. Each variant of GSPo was run on 8 α values from 10^{-10} to .7, and each variant of FCI was run on 7 α values from 10^{-20} to .5

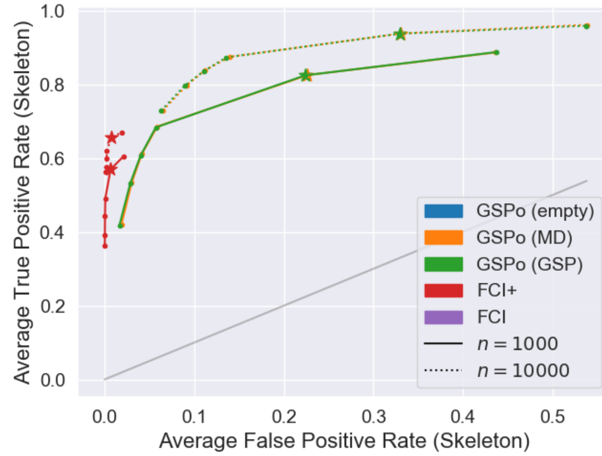

Figure 3: Average performance over 100 MAGs for each algorithm, when $p = 50$, $K = 12$, and $s = 3$. Each variant of GSPo was run on 8 α values from 10^{-10} to .7, and each variant of FCI was run on 7 α values from 10^{-20} to .5

F Additional Simulations

In this section, we followed the same procedure for DMAG sampling procedure as described in Section 5. Fig. 2 gives the precision-recall curve for the same settings as in Fig. 6a in Section 5.

In Figure 3, we use $p = 50$ nodes, $K = 12$ latent variables, and $s = 3$ expected neighbors per node in the DAG before marginalization. For 100 graphs, we find that this results in MAGs with an average of 43% bidirected edges, ranging from 14% to 71% bidirected edges, and an average of 5 neighbors per node in the MAGs. Due to the slow runtime of FCI, GSPo with empty initialization, and FCI+ with high α values, our comparison between the algorithms for larger graphs is limited, and mainly serves to demonstrate that GSPo has similar performance on larger graphs for the same range of α values.

In Figure 4, we use the same set of DMAGs as used in 6c, in particular, $p = 10, 20, 30, 40, 50$, $K = 3$, and $s = 3$, but report the average computation time instead of the median computation time. We can observe that GSPo with the empty initialization and FCI both have much higher average computation times than median computation times, indicating that they are more susceptible to outlier instances from our sampled MAGs.

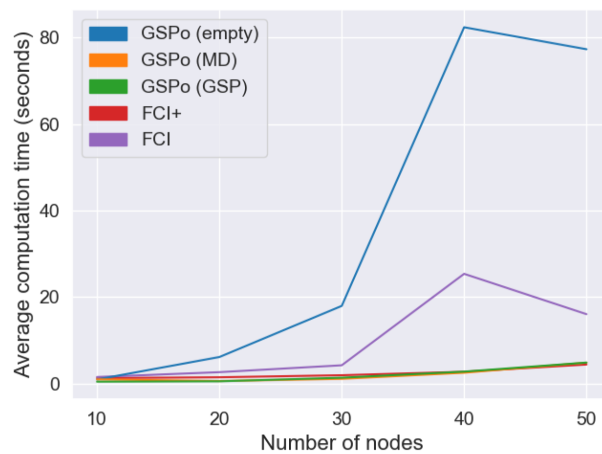

Figure 4: Average runtime over 100 MAGs for $p = 10$, 20, 30, 40, 50, $K = 3$, and $s = 3$. Each variant of GSPo and FCI+ were run with $\alpha = .1$, while FCI was run with $\alpha = 10^{-3}$ due to the extremely long runtime of higher α values.

References

Serge Bouc. The poset of posets. *arXiv preprint arXiv:1311.2219*, 2013.

David Maxwell Chickering. A transformational characterization of equivalent Bayesian network structures. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 87–98. Morgan Kaufmann Publishers Inc., 1995.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–554, 2002.

Tom Claassen, Joris Mooij, and Tom Heskes. Learning sparse causal models is not NP-hard. *arXiv preprint arXiv:1309.6824*, 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*, pages 294–321, 2012.

Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3-4):601–620, 2000.

Steven B Gillispie and Michael D Perlman. Enumerating Markov equivalence classes of acyclic digraph models. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 171–177. Morgan Kaufmann Publishers Inc., 2001.

David Heckerman, Abe Mamdani, and Michael P Wellman. Real-world applications of Bayesian networks. *Communications of the ACM*, 38(3):24–26, 1995.

Steffen L Lauritzen. *Graphical Models*, volume 17. Clarendon Press, 1996.

F. Mohammadi, C. Uhler, C. Wang, and J. Yu. Generalized permutohedra from probabilistic graphical models. *SIAM Journal on Discrete Mathematics*, 32:64–93, 2018.

Preetam Nandy, Alain Hauser, Marloes H Maathuis, et al. High-dimensional consistency in score-based and hybrid structure learning. *The Annals of Statistics*, 46(6A):3151–3183, 2018.

Christopher Nowzohour, Marloes H Maathuis, Robin J Evans, Peter Bühlmann, et al. Distributional equivalence and structure learning for bow-free acyclic path diagrams. *Electronic Journal of Statistics*, 11 (2):5342–5374, 2017.

Joseph Ramsey, Jiji Zhang, and Peter L Spirtes. Adjacency-faithfulness and conservative causal inference. *arXiv preprint arXiv:1206.6843*, 2012.

Garvesh Raskutti and Caroline Uhler. Learning directed acyclic graph models based on sparsest permutations. *Stat*, 7(1):e183, 2018.

Thomas Richardson. Markov properties for acyclic directed mixed graphs. Technical report, Technical Report, 1999.

Thomas Richardson and Peter Spirtes. Ancestral graph Markov models. *The Annals of Statistics*, 30 (4):962–1030, 2002.

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000.

Ilya Shpitser, Thomas S Richardson, James M Robins, and Robin Evans. Parameter and structure learning in nested markov models. *arXiv preprint arXiv:1207.5058*, 2012.

Liam Solus, Yuhao Wang, Lenka Matejovicova, and Caroline Uhler. Consistency guarantees for permutation-based causal inference algorithms. *arXiv preprint arXiv:1702.03530*, 2017.

Peter Spirtes and Thomas Richardson. A polynomial time algorithm for determining dag equivalence in the presence of latent variables and selection bias. In *Proceedings of the 6th International Workshop on Artificial Intelligence and Statistics*, pages 489–500, 1996.

Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, Prediction, and Search.* MIT press, 2000.

Richard P Stanley. Enumerative combinatorics volume 1 second edition. *Cambridge studies in Advanced Mathematics*, 2011.

Konstantinos Tsirlis, Vincenzo Lagani, Sofia Triantafillou, and Ioannis Tsamardinos. On scoring maximal ancestral graphs with the max–min hill climbing algorithm. *International Journal of Approximate Reasoning*, 102:74–85, 2018.

Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, Bin Yu, et al. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, 41 (2):436–463, 2013.

Sara Van de Geer, Peter Bühlmann, et al. ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *The Annals of Statistics*, 41(2):536–567, 2013.

Jiji Zhang and Peter Spirtes. Strong faithfulness and uniform consistency in causal inference. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 632–639. Morgan Kaufmann Publishers Inc., 2002.

Jiji Zhang and Peter Spirtes. A transformational characterization of Markov equivalence for directed acyclic graphs with latent variables. *arXiv preprint arXiv:1207.1419*, 2012.