Minimal I-MAP MCMC for Scalable Structure Discovery in Causal DAG Models

Raj Agrawal [1 2 3] **Tamara Broderick** [1 2] **Caroline Uhler** [2 3]

Abstract

Learning a Bayesian network (BN) from data can be useful for decision-making or discovering causal relationships. However, traditional methods often fail in modern applications, which exhibit a larger number of observed variables than data points. The resulting uncertainty about the underlying network as well as the desire to incorporate prior information recommend a Bayesian approach to learning the BN, but the highly combinatorial structure of BNs poses a striking challenge for inference. The current state-of-the-art methods such as order MCMC are faster than previous methods but prevent the use of many natural structural priors and still have running time exponential in the maximum indegree of the true directed acyclic graph (DAG) of the BN. We here propose an alternative posterior approximation based on the observation that, if we incorporate empirical conditional independence tests, we can focus on a high-probability DAG associated with each order of the vertices. We show that our method allows the desired flexibility in prior specification, removes timing dependence on the maximum indegree, and yields provably good posterior approximations; in addition, we show that it achieves superior accuracy, scalability, and sampler mixing on several datasets.

1. Introduction

Bayesian networks (BNs)—or probabilistic graphical models based on directed acyclic graphs (DAGs)—form a powerful framework for representing complex dependencies among random variables. Learning BNs among observed variables from data has proven useful in decision tasks—such as credit assessment or automated medical diagnosis—as well as discovery tasks—such as learning gene regulatory networks (Spirtes et al., 2000; Friedman et al., 2000; Pearl, 2009; Robins et al., 2000; Khashei & Mirahmadi, 2015). When the number of data points is much larger than the number of observed variables, a point estimate of the BN can be found using constraint-based or greedy search methods (Spirtes et al., 2000; Chickering, 2002; Tsamardinos et al., 2006). However, in many applications, the number of observed variables is *larger* than the number of data points. In this case, many BNs may agree with the observed data. A Bayesian approach offers a natural weighting scheme across BNs via the Bayesian posterior distribution. This weighting propagates coherently to point estimates and uncertainty quantification for structural features of the BN (such as the presence of certain directed edges). Moreover, a Bayesian approach allows the incorporation of prior information, which is common in applications of interest (Mukherjee & Speed, 2008).

Unfortunately, due to the combinatorial explosion of the number of DAGs, exact posterior computation is intractable for graphs with more than thirty nodes (Koivisto & Sood, 2004; Tian & He, 2009). This motivated Madigan & York (1995) to propose *structure MCMC*, an approximate method using Markov chain Monte Carlo (MCMC). To overcome the slow mixing of structure MCMC and its variants (Grzegorczyk & Husmeier, 2008), Friedman & Koller (2003) introduced *order MCMC*. This algorithm achieves significantly faster mixing by running a Markov chain not over DAGs, but in the reduced space of permutations (i.e., orders) of the vertices of the DAG. However, order MCMC requires a particular (often undesirable) form for the prior on DAGs, and its iterations suffer from exponential time and memory complexity in the maximum indegree of the true DAG. Heuristic fixes for scalability exist (Friedman & Koller, 2003), but their statistical costs are unclear.

In this paper, we propose a new method to leverage the improved mixing of MCMC moves in the permutation space; in addition, our approach comes with theoretical guarantees on approximation quality and allows more realistic DAG

[1]Computer Science and Artificial Intelligence Laboratory [2]Institute for Data, Systems and Society [3]Laboratory for Information and Decision Systems, Massachusetts Institute of Technology. Correspondence to: Raj Agrawal <r.agrawal@csail.mit.edu, r.agrawal.raj@gmail.com>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

priors. The key new ingredient is an observation by Verma & Pearl (1992) that has been used for causal inference in the frequentist setting (Mohammadi et al., 2018; Raskutti & Uhler, 2013); namely, if we have access to conditional independence (CI) tests, we can associate each permutation with a unique DAG known as the *minimal I-MAP* (independence map). This is the sparsest DAG that is consistent with a given permutation and Markov to a given set of CI relations. We prove that the vast majority of posterior mass is concentrated on the corresponding reduced space of DAGs, and we call our method *minimal I-MAP MCMC*.

We start in Section 2 by reviewing BNs and Bayesian learning of BNs. We show how to reduce to the space of minimal I-MAPS in Section 3 and theoretically bound the posterior approximation error induced by this reduction. In Section 4, we show by an *empirical Bayes* argument that sufficiently accurate CI tests allow using what amounts to our original prior and likelihood on DAGs but, crucially, restricted to the space of minimal I-MAPs. Thus, we demonstrate that our method allows arbitrary prior structural information. In Section 5, we present an MCMC approach for sampling according to this minimal I-MAP model and provide intuition for why it exhibits good mixing properties. Moreover, we prove that, for p the number of observed variables and k the maximum indegree of the true DAG, our method takes $\mathcal{O}(p^2)$ memory and $\mathcal{O}(p^4)$ time per MCMC iteration (vs. $\mathcal{O}(p^{k+1})$ time and memory for order MCMC). In Section 6 we empirically compare our model to order MCMC and *partition MCMC* (Kuipers & Moffa, 2017), the state-of-the-art version of structure MCMC. In experiments we observe $\mathcal{O}(p^3)$ time scaling for our method, and we demonstrate better mixing and ROC performance for our method on several datasets.

2. Preliminaries and Related Work

2.1. Bayesian Networks

Let $G = ([p], A)$ be a *directed acyclic graph* (DAG) consisting of a collection of vertices $[p] := \{1, \dots, p\}$ and a collection of arrows (i.e., directed edges) A, where $(i, j) \in A$ represents the arrow $i \to j$. A DAG induces a *partial order* $\precsim$ on the vertices $[p]$ through $(i, j) \in A$ if and only if $i \precsim j$. Let S_p be the symmetric group of order p. A *topological order* of a DAG is a permutation $\pi \in S_p$ such that for every edge $(i, j) \in A$, $i \precsim j$ in π; thus it is a *total order* that extends (i.e., is consistent with) the partial order of the DAG, also known as a *linear extension* of the partial order.

A *Bayesian network* is specified by a DAG G and a corresponding set of edge weights $\theta \in \mathbb{R}^{|A|}$. Each node in G is associated with a random variable X_i. Under the *Markov Assumption*, which we assume throughout, each variable X_i is conditionally independent of its nondescendants given its parents, i.e., the joint distribution factors as $\prod_{i=1}^{p} \mathbb{P}\big(X_i \mid \mathsf{Pa}_G(X_i)\big)$, where $\mathsf{Pa}_G(X_i)$ denotes the parents of node X_i (Spirtes et al., 2000, Chapter 4). This factorization implies a set of *conditional independence* (CI) relations that can be read off from the DAG G by *d-separation*. The *faithfulness assumption* states that the only CI relations realized by $\mathbb{P}$ are those implied by d-separation in G (Spirtes et al., 2000, Chapter 4). DAGs that share the same d-separation statements make up the *Markov equivalence class* of a DAG (Lauritzen, 1996, Chapter 3). The Markov equivalence class of a DAG can be uniquely represented by its *CP-DAG*, which places arrows on those edges consistent across the equivalence class (Andersson et al., 1997). The arrows of the CP-DAG are called *compelled edges* and represent direct causal effects (Andersson et al., 1997).

2.2. Bayesian Inference for DAG models

In many applications, the goal is to recover a function $f(G)$ of the underlying causal DAG G given n i.i.d. samples on the nodes, which we denote by $D = \{X_{mi} : m \in [n], i \in [p]\}$. For example, we might ask whether a directed edge (i, j) is in A, or we might wish to discover which nodes are in the Markov blanket of a node i. In applications where n is large relative to p, a point estimate of G—and thereby of $f(G)$—suffices from both a practical and theoretical perspective (Chickering, 2002). However, in many applications of modern interest, n is small relative to p. In this case there may be many DAGs that agree with the observed data and it is then desirable to infer a distribution across DAGs instead of outputting just one DAG. Taking a Bayesian approach we can define a *prior* $\mathbb{P}(G)$ on the space of DAGs, which can encode prior structural knowledge about the underlying DAG—as well as desirable properties such as sparsity. The *likelihood* $\mathbb{P}(D \mid G)$ is obtained by marginalizing out θ:

$$\begin{aligned}\mathbb{P}(D \mid G) &= \int_\theta \mathbb{P}(D, \theta \mid G)\, d\theta \\ &= \int_\theta \mathbb{P}(D \mid \theta, G)\mathbb{P}(\theta \mid G)\, d\theta\end{aligned}$$

and can be tractably computed for certain classes of distributions (Geiger & Heckerman, 1999; Kuipers et al., 2014). Applying Bayes theorem yields the *posterior distribution* $\mathbb{P}(G \mid D) \propto \mathbb{P}(D \mid G)\mathbb{P}(G)$, which describes the state of knowledge about G after observing the data D. From the posterior one can then compute $\mathbb{E}_{\mathbb{P}(G|D)} f(G)$, the posterior mean of the function of interest. Note that in the common setting where f takes the form of an indicator function, this quantity is simply a posterior probability.

Unfortunately, computing the normalizing constant of the posterior distribution is intractable already for moderately sized graphs, since it requires evaluating a sum over the space of all DAGs on p vertices (Koivisto & Sood, 2004). To sample from the posterior without computing the normal-

izing constant, Madigan & York (1995) proposed *structure MCMC*, which constructs a Markov chain on the space of DAGs with stationary distribution equal to the exact posterior. T samples $\{G_t\}$ from such a Markov chain can then be used to approximate the posterior mean of the function of interest, namely $\mathbb{E}_{\mathbb{P}(G|D)} f(G) \approx T^{-1} \sum_{t=1}^{T} f(G_t)$.

Problematically, the posterior over DAGs is known to exhibit many local maxima, so structure MCMC exhibits poor mixing even on moderately sized problems (Friedman & Koller, 2003; Ellis & Wong, 2008). To overcome this limitation, Friedman & Koller (2003) proposed *order MCMC*, which constructs a Markov chain on the space of permutations, where the moves are transpositions. The posterior over orders is smoother than the posterior over DAGs, since the likelihood corresponding to each order is a sum over many DAGs, and increased smoothness usually leads to better mixing behavior. However, strong modularity assumptions are needed to make computing the likelihood tractable. Even under these assumptions, there remains a considerable computational cost: namely, let k be the maximum indegree of the underlying DAG, then the likelihood can be computed in $\mathcal{O}(p^{k+1})$ time and memory (Friedman & Koller, 2003). Hence, in practice k can be at most 3 or 4 for this method to scale to large networks. The Monte Carlo estimate $\frac{1}{T} \sum_{i=1}^{T} f(G_{\pi_t})$, where G_{π_t} is drawn from $\mathbb{P}(G \mid \pi_t, D)$ and π_t is sampled from the posterior over permutations $\mathbb{P}(\pi \mid D)$, is then used to approximate the posterior mean of the function of interest. Friedman & Koller (2003) obtain a practical MCMC sampler when the prior over permutations is uniform, but such a model introduces significant bias toward DAGs that are consistent with more permutations (Ellis & Wong, 2008). Correcting for this bias by re-weighting each sampled DAG by the inverse number of its linear extensions can be done, but it is $\#P$ in general (Ellis & Wong, 2008).

A recent extension of order MCMC is *partial order MCMC* (Niinimaki et al., 2016). This method works on the reduced space of partial orders, thereby leading to improved mixing as compared to order MCMC, but with a similar runtime. Kuipers & Moffa (2017) further introduced a related method known as *partition MCMC*, which avoids the bias of order MCMC by working on the larger space of node partitions consisting of permutations and corresponding *partition* elements. Although partition MCMC generally mixes more slowly than order MCMC, it was empirically found to mix more quickly than structure MCMC (Kuipers & Moffa, 2017).

3. Reduction to the Space of Minimal I-MAPs

To overcome the computational bottleneck of order MCMC and at the same time avoid the slow mixing of structure MCMC, we propose to restrict our focus to a carefully chosen, reduced space of DAGs that is in near one-to-one correspondence with the space of permutations. We construct this subspace of DAGs from the CI relations that hold in the data D. In Appendix A we review a CI testing framework based on partial correlations for the Gaussian setting.

Given a CI test, let $\hat{\mathcal{O}}^{(n)}_{i,j|S}(D, \alpha)$ be 1 if the corresponding CI test at level α based on the n data points in D was rejected—i.e., $X_i \not\perp\!\!\!\perp X_j \mid X_S$—and 0 otherwise. Let $\hat{\mathcal{O}}_n(D, \alpha)$ denote the collection of CI test outcomes across all triples (i, j, S). Given $\hat{\mathcal{O}}_n(D, \alpha)$ we associate to each permutation $\pi \in S_p$ its *minimal I-MAP* $\hat{G}_\pi$: a DAG with vertices $[p]$ and arrows $\pi(i) \to \pi(j)$ if and only if $\hat{\mathcal{O}}^{(n)}_{i,j|S}(D, \alpha) = 1$ with $i < j$ and $S = \{\pi(1) \cdots \pi(j-1)\} \setminus \{\pi(i)\}$. In light of Occam's razor it is natural to consider this mapping, since removing any edge in $\hat{G}_\pi$ induces a CI relation that is not in $\hat{\mathcal{O}}_n(D, \alpha)$ (Spirtes et al., 2000; Raskutti & Uhler, 2013).

Let $\hat{\mathcal{G}} := \{\hat{G}_\pi \mid \pi \in S_p\}$. Then any posterior $\mathbb{P}(\pi \mid D)$ defined by a likelihood and prior on S_p induces a distribution on the space of all DAGs, denoted by $\mathcal{G}$, namely

$$\hat{\mathbb{P}}(G \mid D) := \sum_{\pi \in S_p} \mathbb{1}\{G = \hat{G}_\pi\} \mathbb{P}(\pi \mid D). \tag{1}$$

This distribution places mass only on $\hat{\mathcal{G}}$ and weights each minimal I-MAP according to the posterior probability of sampling a permutation that is consistent with it.

In the remainder of this section we introduce our main result showing that the posterior mean of a function based on the posterior $\mathbb{P}(G \mid D)$ defined by a likelihood and prior on $\mathcal{G}$ is well approximated by the posterior mean of the function based on the distribution $\hat{\mathbb{P}}(G \mid D)$ that has support only on $\hat{\mathcal{G}}$. Before stating our main result, we introduce the assumptions required for this result.

Assumptions 3.1. *Let (G^*, θ^*) define the true but unknown Bayesian network. Let $\mathcal{O}^*$ be the equivalent of $\hat{\mathcal{O}}_n(D, \alpha)$ based on the true but unknown joint distribution on G^*. For each $\pi \in S_p$ let G^*_π denote the minimal I-MAP with respect to $\mathcal{O}^*$. We make the following assumptions:*

(a) $X_i \mid (G^*, \theta^*)$ is multivariate Gaussian.

(b) Let $\rho^*_{i,j|S}$ be the partial correlation derived from the Bayesian network (G^*, θ^*) for the triple (i, j, S) and let $Q^* := \sup_{(i,j,S)} \{|\rho^*_{i,j|S}|\}$. Then there exists $q^* < 1$ such that $\mathbb{P}(Q^* < q^*) = 1$.

(c) Let $R^* := \inf_{(i,j,S)} \{|\rho^*_{i,j|S}| : \rho^*_{i,j|S} \neq 0\}$. Then there exists $r^* > 0$ such that $\mathbb{P}(R^* > r^*) = 1$.

(d) $\hat{G}_\pi$ is a sufficient statistic for $\mathbb{P}(G \mid \pi, D)$, i.e., $\mathbb{P}(G \mid \pi, D) = \mathbb{P}(G \mid \hat{G}_\pi)$.

(e) Let A_π denote the event that $\{\hat{G}_\pi = G^*_\pi\}$. Then $\mathbb{P}(A_\pi \mid \hat{G}_\pi) = \mathbb{P}(A_\pi)$.

(f) There exists some $M < \infty$ such that $\max_G |f(G)| \leq M$.

An in-depth discussion of these assumptions is provided in Appendix C. Assumption 3.1(c) can be regarded as the Bayesian analogue of the *strong-faithfulness assumption*, which is known to be restrictive (Uhler et al., 2013) but is a standard assumption in causal inference for obtaining theoretical guarantees (Kalisch & Buhlmann, 2007; Zhang & Spirtes, 2012). Practitioners often choose f to be an indicator function (e.g. the presence of a directed edge), so Assumption 3.1(f) is typically satisfied in practice.

We now state our main result that motivates constructing a Markov chain on the reduced DAG space $\hat{\mathcal{G}}$ instead of $\mathcal{G}$.

Theorem 3.2. *Under Assumptions 3.1(a)-(f) it holds that*

$$\left|\mathbb{E}_{\mathbb{P}(G|D)} f(G) - \mathbb{E}_{\hat{\mathbb{P}}(G|D)} f(G)\right| \leq 2f(n,p),$$

where $f(n,p) = C_1 M p^2 (n-p) \exp\{-C_2 (r^*)^2 (n-p)\}$.

Theorem 3.2 is proven in Appendix D.2. The main ingredient of the proof is the following lemma that bounds the probability of the events A_π^C for all π.

Lemma 3.3. *Under Assumptions 3.1 (a)-(c) there exist constants C_1, C_2 that depend only on q^* such that*

$$\mathbb{P}(G_\pi^* \neq \hat{G}_\pi) \leq f(n,p),$$

for all $\pi \in S_p$, where $f(n,p)$ is as in Theorem 3.2 and $\hat{G}_\pi$ is constructed using Fisher's z-transform to do CI testing at level $\alpha = 2\big(1 - \Phi\big(\frac{\sqrt{n} r^*}{2}\big)\big)$.

From Theorem 3.2 and Equation (1), it follows that

$$\begin{aligned} \mathbb{E}_{\mathbb{P}(G|D)}[f(G)] &\approx \mathbb{E}_{\hat{\mathbb{P}}(G|D)}[f(G)] \\ &= \sum_{\pi \in S_p} f(\hat{G}_\pi) \mathbb{P}(\pi \mid D). \end{aligned} \tag{2}$$

Hence, using the near one-to-one mapping between S_p and $\hat{\mathcal{G}}$ to associate to each permutation a particular DAG, we can show that the posterior mean $\mathbb{E}_{\mathbb{P}(G|D)} f(G)$ can be well approximated by sampling from a posterior over permutations. This is of particular interest given the observation by Friedman & Koller (2003) that a posterior over permutations is generally smoother than a posterior over DAGs and hence more conducive to fast mixing in MCMC methods.

4. Bayesian Inference on Minimal I-MAPs

Our original Bayesian generative model consisted of a prior $\mathbb{P}(G)$ and a likelihood $\mathbb{P}(D|G)$. In some sense, π may be thought of an auxiliary random variable that aids our reduction to the minimal I-MAP space. But inventing a prior and likelihood for π in order to arrive at the posterior $\mathbb{P}(\pi \mid D)$ in Equation (2) may be conceptually difficult. In particular, it is natural to imagine we might have prior and modeling information for G rather than π in applications. And S_p does not induce a partition in $\mathcal{G}$ (Ellis & Wong, 2008); see also Appendix F. In this section, we demonstrate that, when the available CI information is sufficiently reliable, a good approximation to $\mathbb{E}_{\mathbb{P}(G|D)}[f(G)]$ can be obtained as follows.

$$\mathbb{E}_{\mathbb{P}(G|D)}[f(G)] \approx \sum_{\hat{G} \in \hat{\mathcal{G}}} f(\hat{G}) \mathbb{P}(\hat{G}|D), \quad \text{where} \tag{3}$$
$$\mathbb{P}(\hat{G}|D) \propto \mathbb{P}(D|G = \hat{G}) \mathbb{P}(G = \hat{G})$$

and the final two terms are the original likelihood $\mathbb{P}(D \mid G)$ and prior $\mathbb{P}(G)$ restricted to the minimal I-MAP space. This formula is intuitively appealing; it effectively says that we can obtain a good approximation of the desired posterior expectation by simply restricting our original model to the minimal I-MAP space.

To show this, we start from Equation (2) and let $\hat{\mathcal{O}}_n := \hat{\mathcal{O}}_n(D, \alpha)$ for brevity. Note that $\mathbb{P}(\pi|D) = \mathbb{P}(\pi|D, \hat{\mathcal{O}}_n)$ since $\hat{\mathcal{O}}_n$ is a function of D. Then, by Bayes theorem,

$$\mathbb{P}(\pi \mid D) \propto \mathbb{P}(D \mid \pi, \hat{\mathcal{O}}_n) \mathbb{P}(\pi \mid \hat{\mathcal{O}}_n). \tag{4}$$

Conditioning on a statistic of the data, namely $\hat{\mathcal{O}}_n$ here, before applying Bayes theorem may be thought of as an *empirical Bayes* procedure (Darnieder, 2011).

We examine each of the two factors on the righthand side of Equation (4) in turn. Recall that $A_\pi := \{\hat{G}_\pi = G_\pi^*\}$ is the event that we make no CI errors. First, note that

$$\begin{aligned} \mathbb{P}(D \mid \pi, \hat{\mathcal{O}}_n) &= \sum_{G \in \mathcal{G}} \mathbb{P}(D \mid \pi, \hat{\mathcal{O}}_n, G) \mathbb{P}(G \mid \pi, \mathcal{O}_n) \\ &= \sum_{G \in \mathcal{G}} \mathbb{P}(D \mid G) \mathbb{P}(G \mid \hat{G}_\pi). \end{aligned} \tag{5}$$

Moreover, note that

$$\begin{aligned} &\mathbb{P}(G \mid \hat{G}_\pi) \\ &= \mathbb{P}(G \mid \hat{G}_\pi, A_\pi) \mathbb{P}(A_\pi | \hat{G}_\pi) + \mathbb{P}(G \mid \hat{G}_\pi, A_\pi^C) \mathbb{P}(A_\pi^C | \hat{G}_\pi) \\ &= \mathbb{P}(G \mid \hat{G}_\pi, A_\pi) \mathbb{P}(A_\pi) + \mathbb{P}(G \mid \hat{G}_\pi, A_\pi^C) \mathbb{P}(A_\pi^C) \end{aligned}$$

By Assumption 3.1(e), $\mathbb{P}(A_\pi | \hat{G}_\pi) = \mathbb{P}(A_\pi)$. By Lemma 3.3, $\mathbb{P}(A_\pi)$ approaches 1 exponentially fast in n, and so $\mathbb{P}(A_\pi^C)$ approaches zero exponentially fast in n. Observing that $\mathbb{P}(G | \hat{G}_\pi, A_\pi) = \mathbb{1}\{G = \hat{G}_\pi\}$ and that $\mathbb{P}(G | \hat{G}_\pi, A_\pi^C)$ is bounded by one, we find $\mathbb{P}(G | \hat{G}_\pi) \approx \mathbb{1}\{G = \hat{G}_\pi\}$ for a sufficiently accurate CI test. Therefore, substituting back into Equation (5), we find that

$$\mathbb{P}(D \mid \pi, \hat{\mathcal{O}}_n) \approx \mathbb{P}(D \mid G = \hat{G}_\pi),$$

the likelihood restricted to the space of minimal I-MAPs.

A similar argument, detailed in Appendix E, yields that the second term in Equation (4) is approximately equal to the prior restricted to the space of minimal I-MAPs:

$$\mathbb{P}(\pi \mid \hat{\mathcal{O}}_n) \approx \mathbb{P}(G = \hat{G}_\pi).$$

Finally, if we let $\mathbb{P}(\hat{G}_\pi|D)$ represent the distribution over $\hat{G}_\pi$ proportional to the likelihood $\mathbb{P}(D|G = \hat{G}_\pi)$ times the prior $\mathbb{P}(G = \hat{G}_\pi)$, we can replace Equation (2) with Equation (3) at the beginning of this section, as was our goal. In the next section we develop a Markov Chain Monte Carlo sampler with the desired stationary distribution, $\mathbb{P}(\hat{G}_\pi|D)$.

5. Minimal I-MAP MCMC

In this section we develop a Markov Chain Monte Carlo sampler, which we call *minimal I-MAP MCMC*, to generate approximate samples from the target distribution, $\mathbb{P}(\hat{G}_\pi|D)$. We show that unlike structure MCMC our approach is amenable to fast mixing. Furthermore, we show that minimal I-MAP MCMC overcomes the computational limitations of order MCMC, since its complexity does not depend on the maximum indegree of the underlying DAG G^*.

Our minimal I-MAP MCMC algorithm is detailed in Algorithm 1 for the Gaussian setting. Algorithm 2, denoted as *update minimal I-MAP (UMI)*, is used as a step in Algorithm 1 and describes how to compute a minimal I-MAP $\hat{G}_\tau$ from a minimal I-MAP $\hat{G}_\pi$ when π and τ differ by an adjacent transposition without recomputing all edges; see also Solus et al. (2017). We prove the following proposition about the correctness of our sampler in Appendix D.3.

Proposition 5.1. *In the Gaussian setting, the transitions in Algorithm 1 define an ergodic, aperiodic Markov chain on $\hat{\mathcal{G}}$ with stationary distribution $\mathbb{P}(\hat{G}_\pi|D)$.*

Note that minimal I-MAP MCMC can easily be extended to the non-Gaussian setting by replacing the CI tests based on partial correlations by CI tests based on mutual information. However, for non-Gaussian data our theoretical guarantees do not necessarily hold.

In Section 6 we show empirically that minimal I-MAP MCMC mixes faster than other MCMC samplers. The following example provides intuition for this behavior.

Example 5.1. Suppose the true DAG G^* is the star graph with arrows $2 \to 1, 3 \to 1, \dots p \to 1$. For the sake of simplicity, suppose $\mathcal{O}_n(D, \alpha) = \mathcal{O}^*$. Then for the permutation $\tau = (13 \cdots p2)$ the corresponding minimal I-MAP $\hat{G}_\tau$ equals the fully connected graph. However, a single transposition from τ yields the permutation $\pi = (23 \cdots p1)$, which is consistent with the DAG G^*. Hence minimal I-MAP MCMC can move in a single step from the fully connected graph to the correct DAG, while structure MCMC, which updates one edge at a time, would require many steps and could get stuck along the way.

While this example is clearly idealized, it captures the intuition that traversing the space of minimal I-MAPs via transpositions allows the sampler to make large jumps in DAG space, which allows it to escape local maxima faster and hence mix faster than structure MCMC. In the following result we characterize the memory and time complexity of minimal I-MAP MCMC, showing that unlike order MCMC it does not depend on the maximum indegree of the true DAG G^*. The proof is given in Appendix D.5.

Proposition 5.2. *Let κ be the thinning rate of the Markov chain and T the number of iterations. Consider minimal I-MAP MCMC (Algorithm 1) with a proposal distribution that puts mass only on adjacent transpositions, i.e.*

$$q(\pi_t \to \pi_{t+1}) = \begin{cases} s & \text{if } \pi_t = \pi_{t+1} \\ \frac{1-s}{p} & \text{if } I(\pi_t, \pi_{t+1}) = 1 \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < s < 1$ and $I(\cdot, \cdot) = 1$ if the permutations differ by a single adjacent transposition. This algorithm takes $\mathcal{O}(\kappa T p^2)$ memory and has average time complexity of $\mathcal{O}(Tp^4 + p^5)$. Note that a transposition between the first and last element of a permutation is still considered an adjacent transposition in our definition.

Using a proposal that considers only adjacent transpositions leads to a considerable speed up. In particular, if we consider any possible transition, updating $\hat{G}_{\pi_t}$ to $\hat{G}_{\pi_{t+1}}$ requires $\mathcal{O}(p^2)$ CI tests in general. But the cost is reduced to $\Theta(p)$ CI tests for adjacent transpositions that do not swap the first and last elements. Since performing a CI test based on partial correlations takes $\mathcal{O}(p^3)$ time (Viertl, 2011), this yields a total speed up of a factor of p at each step. We should note that Algorithm 1 can be sped up by considering only adjacent transpositions that are connected by an edge; i.e., in minimal I-MAP space $\hat{\mathcal{G}}$ these adjacent transpositions would correspond to considering only *covered edge* flips (Spirtes et al., 2000; Solus et al., 2017).

We now comment on why our method does not face the computational intractability of order MCMC. Working in the space of minimal I-MAPS parametrized by permutations is similar in spirit to order MCMC, but our approximation of the posterior (that is, the approximation we make even before applying MCMC) allows us to avoid the poor scaling of order MCMC. In particular, the intractability of order MCMC arises due to the focus on an exact likelihood; acquiring this likelihood requires summing over $\mathcal{O}(p^{k+1})$ parent sets in order to sum over the full space of DAGs. In our case, we instead exploit the fact that the likelihood concentrates around a single DAG $\hat{G}_\pi$ once we condition on $\hat{\mathcal{O}}_n(D, \alpha)$.

Algorithm 1 Minimal I-MAP MCMC

Input: Data D, number of iterations T, significance level α, initial permutation π_0, sparsity strength γ, thinning rate κ
Output: $\hat{G}_{\pi_1}, \cdots, \hat{G}_{\pi_{\lceil \kappa T \rceil}}$
Construct $\hat{G}_{\pi_0}$ from $\hat{\mathcal{O}}_n(D, \alpha)$ via Fisher's z-transform
for $i = 1$ **to** T **do**
  Sample $\pi_i \sim q(\pi_{i-1} \to \pi_i)$
  $\hat{G}_{\pi_i} = \text{UMI}(\pi_i, \pi_{i-1}, \hat{G}_{\pi_{i-1}}, \alpha, D)$ (Algorithm 2, Appendix B)
  $p_{i-1} = \log \mathbb{P}(D \mid \hat{G}_{\pi_{i-1}}) \mathbb{P}(\hat{G}_{\pi_{i-1}})$
  $p_i = \log \mathbb{P}(D \mid \hat{G}_{\pi_i}) \mathbb{P}(\hat{G}_{\pi_i})$
  $s_i = \min\{1, \exp(p_i - p_{i-1})\}$
  $z_i \sim \text{Bernoulli}(s_i)$
  if $z_i = 0$ **then**
    $\pi_i = \pi_{i-1}$ and $\hat{G}_{\pi_i} = \hat{G}_{\pi_{i-1}}$ (chain does not move)
  end if
  if T is divisible by $\lceil \frac{1}{\kappa} \rceil$ **then**
    Store $\hat{G}_{\pi_i}$
  end if
end for

6. Experiments

In this section we empirically compare minimal I-MAP MCMC to order and partition MCMC. We chose partition MCMC since it does not have the bias of order MCMC and empirically has faster mixing than structure MCMC (Kuipers & Moffa, 2017). We use the max-min-hill-climbing (MMHC) algorithm (Tsamardinos et al., 2006) in conjunction with the nonparametric DAG bootstrap approach (Friedman et al., 1999) as an additional baseline for comparison. For each dataset, we ran the Markov chains for 10^5 iterations, including a burn-in of 2×10^4 iterations, and thinned the remaining iterations by a factor of 100. Seeded runs correspond to starting the Markov chain at the permutation/DAG obtained using MMHC. We also considered "cheat" runs that start at the true permutation with the intuition that we expect high scores on the true generating model. In terms of software, we used the code provided by Kuipers & Moffa (2017) to run partition and order MCMC. We used the method and software of Kangas et al. (2016) for counting linear extensions for bias correction, and we implemented minimal I-MAP MCMC using the R-package `bnlearn`.

6.1. Prior and Likelihood

As in many applications, a prior that induces sparsity in the underlying structure is desirable for interpretability and computation. Further, note that the true DAG G^* is equal to the sparsest minimal I-MAP G^*_π over all permutations S_p based on CI relations $\mathcal{O}^*$ (Verma & Pearl, 1992; Solus et al., 2017; Raskutti & Uhler, 2013); thus, on minimal I-MAP space, a sparsity prior is natural. To achieve this end, we choose a prior of the form

$$\mathbb{P}(G) = \mathbb{P}^*(G) \exp\big(-\gamma \|G\|\big),$$

where $\mathbb{P}^*(G)$ can include any structural information known about the DAG. Except where explicitly mentioned in what follows, we use this prior with $\mathbb{P}^*(G)$ uniform over DAGs. We note that, unlike our method or partition MCMC, order MCMC uses a uniform prior over permutations; the induced prior over DAGs as a result of such a prior is $\mathbb{P}_{\mathsf{order}}(G) = |\#\mathsf{linext}(G)|\mathbb{P}(G)$, where $|\#\mathsf{linext}(G)|$ denotes the number of linear extensions of G (Ellis & Wong, 2008). Finally, each method assumes the data follow a multivariate Gaussian distribution with a Wishart prior on the network parameters. This assumption allows computation of $\mathbb{P}(D|G)$ via the *BGe* score (Geiger & Heckerman, 1999; Kuipers et al., 2014).

6.2. Mixing and Convergence

We consider three different datasets. The first two were obtained by simulating data from a network consisting of $p = 30$ nodes with $n = 100$ and $n = 1000$ observations respectively. The data were generated according to a linear structural equation model with additive Gaussian noise, where the edge weights on the underlying DAG G^* were sampled uniformly from $[-1, -.25] \cup [.25, 1]$ as in (Solus et al., 2017). The third dataset is from the *Dream4* in-silico network challenge (Schaffter et al., 2011) on gene regulation. In particular, we examine the *multifactorial* dataset consisting of ten nodes and ten observations.

In Figure 1 we analyze the mixing performance of the different methods. The convergence of different runs to the same score neighborhood can be taken as an indication of adequate mixing. Figure 1 suggests that for the $n = 100$ and the Dream4 dataset all three methods have mixed well while for the dataset with $n = 1000$ samples there is evidence of poor mixing in all methods since the posterior landscape is more peaky due to increased sample size. However, the score plots are markedly worse for order and partition MCMC.

Since the end goal is often to obtain robust estimates of particular feature probabilities, in Figure 2 we analyze the correlation between the approximated posterior probabilities of directed features with respect to different seeded runs. Figure 2 shows that the correlation is higher across all models for the dataset with $n = 100$ samples, which is to be expected, since the chains seem to have mixed given the analysis in the score plots in Figure 1. Conversely, for the $n = 1000$ dataset, Figure 2 shows that order and partition MCMC yield vastly different posterior probabilities across different runs, while minimal I-MAP MCMC maintains high correlation, thus suggesting again superior mixing.

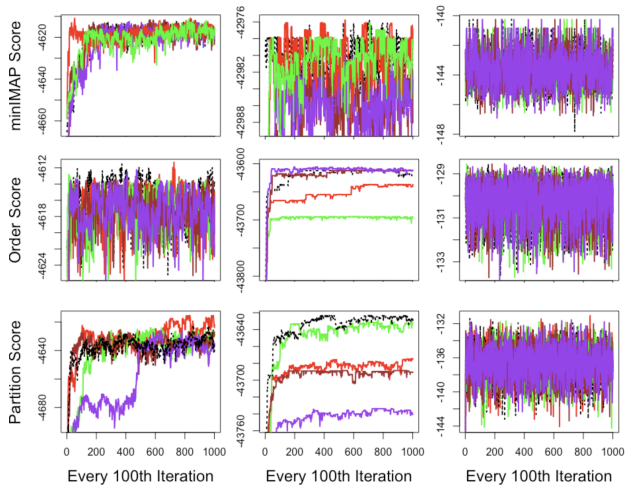

Figure 1. From left to right, the columns represent the $n = 100$, $n = 1000$, and Dream4 datasets, respectively. From top to bottom, the rows correspond to minimal I-MAP (minIMAP), order, and partition MCMC. The black dotted line corresponds to runs seeded with the true permutation. The purple and brown lines correspond to runs seeded with a random permutation and the red and green curves represent runs seeded with MMHC.

6.3. ROC Performance

As described in Section 2, the bias of order MCMC can be removed by dividing the functional of interest $f(G_t)$ by the number of linear extensions of G_t, where G_t is a DAG sampled during the Monte-Carlo Step. We denote this by *full bias correction* (*FBC*). Although this leads to an unbiased estimator for order MCMC, there is a bias-variance trade-off. If a sampled DAG has few linear extensions, this DAG will be given more weight in the Monte Carlo step, thereby increasing the variance. Therefore, we also consider a *partial bias correction* (*PBC*), where the weights are truncated and re-normalized to belong to the 25th and 75th quartile of the inverse linear extension counts of the sampled DAGs. Finally, we denote *no bias correction* by *NBC*.

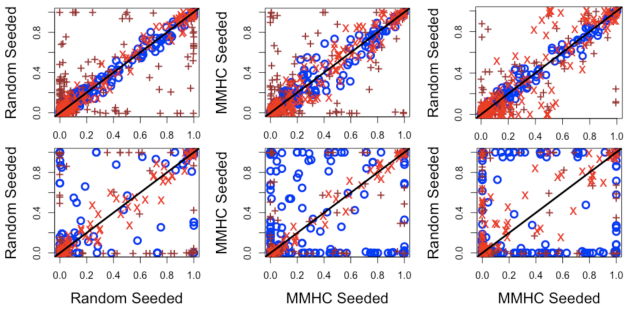

Figure 2. The top row and bottom row correspond to the $n = 100$ and $n = 1000$ datasets respectively. Each point represents the posterior probability of a directed feature estimated by different seeded runs of MCMC. We consider all possible combinations of random and MMHC seeded runs for completeness. Red (x), blue (o), and brown (+) correspond to minIMAP, order, and partition MCMC respectively.

Table 1. AUROC results by dataset and method. NBC, PBC, and FBC stand for no, partial, and full bias correction. The columns represent AUROC values for undirected and compelled features respectively.

METHOD	N=100		N=1000		DREAM4	
MINIMAP	.946	**.695**	**1.00**	**.958**	.574	.556
ORDER-NBC	**.957**	.675	.949	.395	**.599**	.600
ORDER-PBC	.956	.677	.949	.393	.579	.444
ORDER-FBC	.952	**.695**	.950	.395	.563	.489
PARTITION	.857	.660	.890	.674	.497	**.733**
MMHC-BOOT	.842	.693	.892	.668	.552	.533

In Table 1 we report the area under the ROC curves (AUROC) for detecting directed and undirected features for the different methods. For order MCMC, we see a marginal performance boost after bias correction on the simulated datasets, but worse performance on Dream4. For the $n = 100$ and $n = 1000$ datasets, the Bayesian models perform better than the MMHC bootstrap. While Table 1 shows that MMHC achieves the highest AUROC performance on the Dream4 dataset, the corresponding ROC plot provided in Figure 4 in Appendix H shows that minimal I-MAP MCMC and order MCMC compare favorably to MMHC when the true negative rate (TNR), which equals one minus the false positive rate (FPR), is greater than 0.4. This range for the TNR is the relevant regime for biological applications, where it is often more important to control for Type I errors (i.e. incorrectly specifying causal relationships between nodes).

The second column of Table 1 for each dataset shows AUROC performance on the compelled edges and Figure 4 in Appendix H contains the corresponding ROC plots. Recovering compelled edges is important because these are the only causal effects that are identifiable from observational data alone (Pearl, 2009). Table 1 shows that minimal I-MAP MCMC achieves the best performance in terms of recovering compelled edges on the $n = 1000$ dataset and is marginally better than the other methods on the $n = 100$ dataset.

6.4. Time and Memory Complexity

Since partition MCMC has a similar time and memory complexity as order MCMC, we focus on comparing minimal I-MAP MCMC to order MCMC in these regards. Recall that p denotes the number of nodes and k denotes the maximum indegree of the underlying DAG G^*. To control for different implementations, we computed the average iteration times relative to the average iteration time for $p = 25$ nodes. The average iteration times do not include the time it takes to

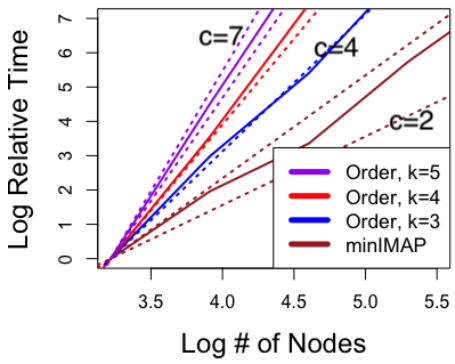

Figure 3. Average iteration times for different sized networks. The times are relative to the average iteration time for $p = 25$ nodes; c denotes the slope of the dotted lines and estimates the computational complexity $\mathcal{O}(p^c)$.

cache all the scores in order MCMC and the time it takes to construct $\hat{G}_{\pi_0}$ for initiating the Markov chain. Figure 3 shows that order MCMC scales similarly to its predicted theoretical complexity of $\mathcal{O}(p^{k+1})$. For minimal I-MAP MCMC, we provided a bound of $\mathcal{O}(p^4)$ in Proposition 5.2. Figure 3 suggests that the complexity scales by a factor of p better than the bound we obtained, namely $\mathcal{O}(p^3)$. Finally, we note that order MCMC runs out of memory quickly when either k or p grows. As a specific example, for only $p = 80$ nodes and $k = 5$, order MCMC takes over 40 GB of space while minimal I-MAP MCMC takes around 1 MB.

6.5. Incorporating Priors

Unlike minimal I-MAP MCMC, both partition and order MCMC require that the prior over DAGs factorizes as $\mathbb{P}(G) = \prod_{i=1}^{p} \rho(X_i, \mathsf{Pa}_G(X_i))$ which is defined as *structure modularity* by Friedman & Koller (2003). $\mathbb{P}(G)$ is used in order MCMC to specify the conditional distribution $\mathbb{P}(G \mid \pi) = I(G \precsim \pi)\mathbb{P}(G)$, which is needed to calculate the likelihood $\mathbb{P}(D \mid \pi)$ in order MCMC; see also Appendix F. The assumption of structure modularity is a practical limitation. In biological applications, for example, prior information often comes in the form of path information between classes of vertices, which is not structure modular in general. In the following, we illustrate this point using the biological network studied by Mukherjee & Speed (2008) (reproduced in Figure 5 in Appendix H). In this application, we have prior knowledge on both orders and paths. In particular, we expect ligands to come before receptors, and receptors before cytosolic proteins. In addition, we expect to see paths from ligands to receptors and paths from receptors to cytosolic proteins (Mukherjee & Speed, 2008). Such path information cannot be used by order and partition MCMC since this information is not structure modular. To test if path knowledge leads to better inference, we compared the ROC plots (Figure 5, Appendix H) and AUROC (Table 2) for directed edge recovery for the different methods. Table 2 shows that the path prior leads to a boost in AUROC performance of minimal I-MAP MCMC by $1 - 2\%$ percent, thereby suggesting that structure modularity can be limiting for certain applications. The specific form of the path and order prior are provided in Appendix G.

Table 2. AUROC results for directed edge recovery in the protein network in Figure 5.

METHOD	AUROC
MINIMAP W/ PATH AND ORDER PRIOR	**.929**
MINIMAP W/ ORDER PRIOR	.917
ORDER W/ ORDER PRIOR	.874
PARTITION W/ ORDER PRIOR	.912
MMHC-BOOT	.909

7. Concluding Remarks

In this paper, we introduced minimal I-MAP MCMC, a new Bayesian approach to structure recovery in causal DAG models. Our algorithm works on the data-driven space of minimal I-MAPs with theoretical guarantees on posterior approximation quality. We showed that unlike order or partition MCMC the complexity of an iteration in minimal I-MAP MCMC does not depend on the maximum indegree of the true underlying DAG. This theoretical result was confirmed in our empirical study. In addition, our empirical study showed that minimal I-MAP MCMC achieves similar or faster mixing than other state-of-the-art methods for Bayesian structure recovery.

While we have focused on the Gaussian setting, it would be interesting in future work to extend the theoretical analysis to other distributions, in particular the discrete setting. Finally, it would be interesting to explore the performance of minimal I-MAP MCMC for obtaining MAP estimates or as a new DAG scoring criterion. In particular, the scoring criterion of the *greedy SP (GSP)* algorithm (Solus et al., 2017) is equivalent to our DAG score (i.e., unnormalized posterior probability) when $\gamma \to \infty$ in the prior in Section 6.1 and the search space is restricted to $\hat{\mathcal{G}}$. In this case, the likelihood term has no influence in picking the minimal I-MAP from $\hat{\mathcal{G}}$. We might therefore find improved performance in terms of structure recovery over the GSP algorithm by incorporating the likelihood term by setting $\gamma < \infty$.

Acknowledgements

Raj Agrawal was supported by an MIT Aziz Ashar Presidential Fellowship. Tamara Broderick was supported in part by an ARO Young Investigator Program Award, ONR

grant N00014-17-1-2072, and a Google Faculty Research Award. Caroline Uhler was supported in part by NSF (DMS-1651995), ONR (N00014-17-1-2147), and a Sloan Fellowship.

References

Andersson, S. A., Madigan, D., and Perlman, M. D. A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25:505–541, 1997.

Chickering, D. M. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3: 507–554, 2002.

Darnieder, W. *Bayesian Methods for Data-Dependent Priors*. PhD thesis, Ohio State University, 2011.

Ellis, B. and Wong, W. H. Learning causal Bayesian network structures from experimental data. *Journal of the American Statistical Association*, 103:778–789, 2008.

Fisher, R. A. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10:507–521, 1915.

Friedman, N. and Koller, D. Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, 50:95–125, 2003.

Friedman, N., Goldszmidt, M., and Wyner, A. J. Data analysis with Bayesian networks: A bootstrap approach. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 1999.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. Using Bayesian networks to analyze expression data. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*, 2000.

Geiger, D. and Heckerman, D. Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 1999.

Grzegorczyk, M. and Husmeier, D. Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. *Machine Learning*, 71:265–305, 2008.

Kalisch, M. and Buhlmann, P. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8:613–636, 2007.

Kangas, K., Hankala, T., Niinimaki, T., and Koivisto, M. Counting linear extensions of sparse posets. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 2016.

Khashei, M. and Mirahmadi, A. A soft intelligent risk evaluation model for credit scoring classification. *International Journal of Financial Studies*, 3:411–422, 2015.

Koivisto, M. and Sood, K. Exact Bayesian structure discovery in Bayesian networks. *Journal of Machine Learning Research*, 5:549–573, 2004.

Kuipers, J. and Moffa, G. Partition MCMC for inference on acyclic digraphs. *Journal of the American Statistical Association*, 112:282–299, 2017.

Kuipers, J., Moffa, G., and Heckerman, D. Addendum on the scoring of Gaussian directed acyclic graphical models. *The Annals of Statistics*, 42:1689–1691, 2014.

Lauritzen, S. *Graphical Models*. Oxford University Press, 1996.

Madigan, D. and York, J. Bayesian graphical models for discrete data. *International Statistical Review*, 63:215–232, 1995.

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

Mukherjee, S. and Speed, T. P. Network inference using informative priors. *Proceedings of the National Academy of Sciences*, 105:14313–14318, 2008.

Niinimaki, T., Parviainen, P., and Koivisto, M. Structure discovery in Bayesian networks by sampling partial orders. *Journal of Machine Learning Research*, 17:2002–2048, 2016.

Pearl, J. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009.

Raskutti, G. and Uhler, C. Learning directed acyclic graphs based on sparsest permutations. 2013.

Robins, J., Hernan, M. A., and Brumback, B. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11:550–60, 2000.

Schaffter, T., Marbach, D., and Floreano, D. *Bioinformatics*, 27:2263–2270, 2011.

Solus, L., Wang, Y., Matejovicova, L., and Uhler, C. Consistency guarantees for permutation-based causal inference algorithms. *arXiv:1702.03530*, 2017.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

Tian, J. and He, R. Computing posterior probabilities of structural features in Bayesian networks. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 2009.

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65:31–78, 2006.

Uhler, C., Raskutti, G., Bühlmann, P., and Yu, B. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, 41:436–463, 2013.

Verma, T. and Pearl, J. An algorithm for deciding if a set of observed independencies has a causal explanation. In *Proceedings of the Eighth Conference on Uncertainty in Artificial Intelligence*, 1992.

Viertl, R. *Probability and Bayesian Statistics*. Springer US, 2011.

Zhang, J. and Spirtes, P. Strong faithfulness and uniform consistency in causal inference. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, 2012.