Are Bayesian networks typically faithful?

Philip Boeken* Patrick Forré† Joris M. Mooij†

March 11, 2026

Abstract

Faithfulness is a common assumption in causal inference, often motivated by the fact that the faithful parameters of linear Gaussian and discrete Bayesian networks are typical, and the folklore belief that this should also hold for other classes of Bayesian networks. We address this open question by showing that among all Bayesian networks over a given DAG, the faithful Bayesian networks are indeed 'typical': they constitute a dense, open set with respect to the total variation metric. This does not directly imply that faithfulness is typical in restricted classes of Bayesian networks that are often considered in statistical applications. To this end we consider the class of Bayesian networks parametrised by conditional exponential families, for which we show that under regularity conditions, the faithful parameters constitute a dense and open set, the unfaithful parameters have Lebesgue measure zero, and the induced faithful distributions are open and dense in the weak topology. This extends the existing results for linear Gaussian and discrete Bayesian networks. We also show for nonparametric classes of Bayesian networks with uniformly equicontinuous and uniformly bounded conditional densities that the faithful Bayesian networks are open and dense in the weak topology. All these results also hold for Bayesian networks with latent variables, if faithfulness is only required to hold with respect to the latent projection. Finally, for the considered conditional exponential family parametrisations and nonparametric conditional density models, the topological properties of conditional independence imply the existence of a consistent conditional independence test. Together with the topological properties of faithfulness, this implies that sound constraint-based causal discovery algorithms like *PC* and *FCI* are consistent on an open and dense – and hence 'typical' – set of Bayesian networks.

1 Introduction

Given a Bayesian network over a DAG G with variables V and a finite sample from its distribution $\mathbb{P}(X_V)$, the task of *causal discovery* algorithms is to infer the graph G from the data. *Constraint-based* causal discovery methods do so by testing for conditional independencies $X_A \perp\!\!\!\perp_{\mathbb{P}} X_B \,|\, X_C$ for various choices of $A, B, C \subseteq V$, and use this information to reconstruct G, up to certain equivalences. A core assumption of many constraint-based causal discovery algorithms is that a correctly inferred set of conditional independencies in $\mathbb{P}(X_V)$ characterises the corresponding set of d-separations in G: for all subsets of vertices $A, B, C \subseteq V$ we have

$$A \perp^d_G B \,|\, C \iff X_A \perp\!\!\!\perp_{\mathbb{P}} X_B \,|\, X_C.$$

Bayesian networks for which this condition holds are called *faithful* (Pearl, 1988; Spirtes et al., 1993). The implication from left to right holds for all Bayesian networks, and is called the *global Markov property* (Verma and Pearl, 1990). The implication from right to left does not always hold: there exist Bayesian networks which have conditional independencies that are not due to a corresponding d-separation in the graph — instead, they might be due to cancelling paths, deterministic variables, or deterministic relations (see Example 1 below).

In absence of any knowledge of the graph G, faithfulness is an untestable assumption (Zhang and Spirtes, 2008). In practice, this assumption is often motivated by theoretical results that for certain

*Department of Mathematics, VU Amsterdam, p.a.boeken@vu.nl
†Korteweg-de Vries Institute for Mathematics, University of Amsterdam

parametric models, the faithful distributions are 'typical'. For a given DAG G, Spirtes et al. (1993) and Meek (1995) consider specific parametrisations $\Theta_{\mathcal{N}}$ and $\Theta_{\mathcal{D}}$ of linear Gaussian and discrete Bayesian networks respectively (which are subsets of $\mathbb{R}^d$ for appropriate $d \in \mathbb{N}$, see Examples 4 and 5 below in Section 4) and show that drawing the parameter values at random will give a faithful Bayesian network with probability one:

Theorem 1 (Spirtes et al., 1993)**.** *With respect to Lebesgue measure over $\Theta_{\mathcal{N}}$, the subset of parameter values whose distribution is unfaithful to G is measure-zero.*

Theorem 2 (Meek, 1995)**.** *With respect to Lebesgue measure over $\Theta_{\mathcal{D}}$, the subset of parameter values whose distribution is unfaithful to G is measure-zero.*

To our knowledge, no such results are available for other parametric or nonparametric classes of distributions. In this work we prove such a result: without restriction to any parametric or nonparametric class of distributions, the faithful distributions are typical. As there is no canonical analogue of the Lebesgue measure for the nonparametric space of Bayesian networks, we don't consider the measure-theoretic notion of typicality, but instead consider a topological notion. Our most general nonparametric result, Theorem 5, is as follows:

> **Among all distributions that are Markov with respect to a given DAG, the faithful distributions constitute a dense, open set.**

As a consequence, the set of faithful distributions is non-empty, and unfaithful distributions are *nowhere dense* (defined below) and are thus 'atypical'. This topological property is with respect to the total variation metric on the joint distribution $\mathbb{P}(X_V)$ over all vertices V of the Bayesian network. This result holds for any choice of *standard Borel* outcome spaces; it holds in particular for continuous variables $X_V \in \mathbb{R}^{|V|}$, discrete variables $X_V \in \mathbb{Z}^{|V|}$, and mixed data. Formally, a Bayesian network is a DAG G with for every vertex v a Markov kernel $\mathbb{P}(X_v \mid X_{\mathrm{pa}(v)})$, where $\mathrm{pa}(v)$ denote the parents of v in G. The above result is about the observational distributions, but not about the Bayesian networks themselves, which are tuples of Markov kernels (one for each $v \in V$). To this end, we introduce a metric d_{TV}° on the space of Bayesian networks — corresponding to total variation convergence of the Markov kernels, uniformly in the conditioning variables — for which we show in Theorem 6 that the faithful Bayesian networks are open and dense.

These claims do not automatically generalise to other topologies than the total variation topology or the topology induced by d_{TV}°: open sets are maintained under refinements, and dense sets are maintained under coarsening, but nothing can be said in general about open and dense sets. Besides the total variation topology, the weak topology is of particular interest since it is tightly connected to statistical testability (Dembo and Peres, 1994; Ermakov, 2017; Genin and Kelly, 2017; Boeken et al., 2026). Since faithfulness is not open in the weak topology, some regularity conditions are necessary to obtain the analogue of the above result in the weak topology.

In practice, one often imposes regularity assumptions on the distribution to facilitate statistical inference. To this end, we consider two subclasses of Bayesian networks. First, we consider the class of Bayesian networks parametrised by conditional exponential families. Under regularity conditions, we obtain in Theorem 8 that if there exists a faithful parameter, then the faithful parameters constitute a dense and open set, and the unfaithful parameters have Lebesgue measure zero. We further show in Theorem 9 that the induced set of faithful observational distributions is open and dense with respect to the weak topology. Second, we consider nonparametric models of Bayesian networks with uniformly equicontinuous and uniformly bounded conditional densities. In this class we also obtain that if there exists a faithful model, then the faithful models constitute a dense and open set with respect to d_{TV}° (Theorem 4). Since for this model class the weak topology and the total variation topology coincide, convergence in this metric corresponds to weak convergence of the Markov kernels, uniformly in the conditioning variable. We also show that the induced set of faithful observational distributions is open and dense with respect to the weak topology (Theorem 11).

To relate these results to constraint-based causal discovery, we show that for the subclasses of Bayesian networks with conditional exponential families or nonparametric conditional densities, it

follows from results by Genin and Kelly (2017) and Lauritzen (2024) that there exists a consistent conditional independence test, and hence that any sound constraint-based causal discovery algorithm using such a test is consistent on an open and dense set of Bayesian networks (Theorem 13). This holds for settings where the samples spaces are separable complete metric spaces.

There exist multiple mathematical notions of 'atypicality'. Given a set M, 'small' subsets of M are characterised by so-called σ-ideals: collections of subsets of M containing $\emptyset$, that are closed under taking subsets and countable unions. The family of Lebesgue measure zero sets is a σ-ideal, and so is the family of meager sets:

Definition 1. A set $I \subseteq M$ is *dense* in another set $U \subseteq M$ if every point in U is in I or is a limit point of I. The set I is *nowhere dense* if there is no non-empty open subset of M in which I is dense, and it is *meager* if it is a countable union of nowhere dense sets.

For example, the set of integers $\mathbb{Z}$ is nowhere dense in $\mathbb{R}$, and the rationals $\mathbb{Q}$ are meager in $\mathbb{R}$. The boundary of every open or closed set is nowhere dense, and subsets of nowhere dense sets are nowhere dense. Complements of dense sets are not necessarily nowhere dense or meager, but complements of *dense and open* sets are nowhere dense. Comeager sets (complements of meager sets) are commonly referred to as *typical* (Kechris, 1995). Our results imply that unfaithful distributions and parameters are nowhere dense, which is an even a stronger notion of atypicality.

In causality, the σ-ideal of meager sets is considered by Ibeling and Icard (2021), who show that discrete causal models for which *Pearl's Causal Hierarchy* collapses[1] are meager, which is a topological analogue of a Lebesgue measure-zero result from Bareinboim et al. (2022). Lin and Zhang (2020) prove open- and denseness of faithful parameters of discrete Bayesian networks.

The outline of this paper is as follows. In Section 2 we provide some technical prerequisites about Bayesian networks, the total variation metric, the bounded-Lipschitz metric and the weak topology. In Section 3 we prove for unconstrained Bayesian networks that faithful distributions are dense and open in the total variation metric, and that the faithful Bayesian networks are open and dense in a newly introduced metric d_{TV}°. In Section 4 we focus on conditional exponential family parametrisations of Bayesian networks, where we show that the faithful parameters are open and dense in the Euclidean parameter space — generalising the results of Spirtes et al. and Meek for linear Gaussian and discrete Bayesian networks — and that the induced faithful distributions are open and dense in the weak topology. In Section 5 we prove for classes of distributions with uniformly equicontinuous and uniformly bounded densities that faithful Bayesian networks are open and dense with respect to d_{TV}°, and that faithful distributions are open and dense with respect to the weak topology. In Section 6 we extend our results to Bayesian networks with latent variables. Finally, in Section 7 we discuss the relative value of these results with their various notions of typicality. We explicitly show how topological properties of conditional independence and faithfulness imply the existence of consistent conditional independence tests, and with that a constraint-based causal discovery algorithm that is consistent on an open and dense set of Bayesian networks.

2 Technical prerequisites

A *directed acyclic graph* (DAG) is a tuple $G = (V, E)$ with V a finite set of vertices and $E \subset V \times V$ a set of directed edges such that there are no directed cycles. Given such a finite index set V, let $\mathcal{X}_V = \prod_{v \in V} \mathcal{X}_v$ be a product of separable complete metric spaces, each equipped with the Borel σ-algebra $\mathcal{B}(\mathcal{X}_v)$ (which are *standard Borel spaces*), and let $\mathcal{P}(\mathcal{X}_V)$ be the set of probability measures on $\mathcal{X}_V$. Random variables will be denoted with X_V, and their values with x_V. For $A, B \subseteq V$, a *Markov kernel* $\mathbb{P}(X_B \mid X_A)$ is a measurable map $\mathcal{X}_A \to \mathcal{P}(\mathcal{X}_B)$, where $\mathcal{P}(\mathcal{X}_B)$ is equipped with the smallest σ-algebra that makes for all $D \in \mathcal{B}(\mathcal{X}_B)$ the evaluation map $\mathrm{ev}_D : \mathcal{P}(\mathcal{X}_B) \to [0,1], \mathbb{P} \mapsto \mathbb{P}(X_B \in D)$

[1] A structural causal model 'collapses' when all counterfactual (interventional) queries are identifiable from interventional (observational) distributions.

measurable. For Markov kernels $\mathbb{P}(X_A \,|\, X_B), \mathbb{P}(X_B \,|\, X_C)$, their *product* is defined as the Markov kernel

$$\mathbb{P}(X_A \,|\, X_B) \otimes \mathbb{P}(X_B \,|\, X_C) : \mathcal{X}_C \to \mathcal{P}(\mathcal{X}_{A \cup B}), \quad x_C \mapsto \left(D \mapsto \int_D \mathrm{d}\mathbb{P}(x_A \,|\, x_B) \mathrm{d}\mathbb{P}(x_B \,|\, x_C) \right)$$

where $D \in \mathcal{B}(\mathcal{X}_{A \cup B})$. Since $\mathcal{X}_V$ is standard Borel, there exists for any joint distribution $\mathbb{P}(X_A, X_B)$ (where $A, B \subseteq V$) a Markov kernel (often referred to as *conditional distribution*) $\mathbb{P}(X_B \,|\, X_A)$ such that $\mathbb{P}(X_A, X_B) = \mathbb{P}(X_B \,|\, X_A) \otimes \mathbb{P}(X_A)$ (Bogachev, 2007, Corollary 10.4.15). Given distribution $\mathbb{P} \in \mathcal{P}(\mathcal{X}_V)$ and sets $A, B, C \subseteq V$, we say that X_A is *conditionally independent* of X_B given X_C, written $X_A \perp\!\!\!\perp_{\mathbb{P}} X_B \,|\, X_C$, if for all $E_A \in \mathcal{B}(\mathcal{X}_A)$ and $E_B \in \mathcal{B}(\mathcal{X}_B)$ we have $\mathbb{P}(X_A \in E_A, X_B \in E_B \,|\, X_C) = \mathbb{P}(X_A \in E_A \,|\, X_C)\mathbb{P}(X_B \in E_B \,|\, X_C)$ a.s. If $\mathbb{P}(X_A, X_B, X_C)$ has a density $p(x_A, x_B, x_C)$, we have $X_A \perp\!\!\!\perp_{\mathbb{P}} X_B \,|\, X_C$ if and only if $p(x_A, x_B \,|\, x_C) = p(x_A \,|\, x_C)p(x_B \,|\, x_C)$ a.e. If the density is continuous, this is equivalent to the factorisation holding everywhere in the support of $\mathbb{P}(X_A, X_B, X_C)$.

For any vertex $v \in V$, its parents in G are given by the set $\mathrm{pa}(v) := \{w \in V : w \to v \in V\}$, and its ancestors are given by $\{w \in V : \exists \text{ a path } (w \to \ldots \to v) \in V\}$. A *Bayesian network*[2] consists of a DAG G and a tuple of Markov kernels $(\mathbb{P}(X_v \,|\, X_{\mathrm{pa}(v)}))_{v \in V}$. The joint distribution $\mathbb{P}(X_V) = \bigotimes_{v \in V} \mathbb{P}(X_v \,|\, X_{\mathrm{pa}(v)})$ is referred to as the *observational distribution*. Given DAG G with path $\pi = a \ast\!\!-\!\!\ast \ldots \ast\!\!-\!\!\ast b$, a *collider* is a vertex v with $\ldots \to v \leftarrow \ldots$ in π, where $\ast\!\!-\!\!\ast$ is a placeholder for either $\to$ or $\leftarrow$. For sets of vertices $A, B, C \subseteq V$ we say that A and B are *d-separated* given C, written $A \perp^d_G B \,|\, C$, if for every path $\pi = a \ast\!\!-\!\!\ast \ldots \ast\!\!-\!\!\ast b$ between $a \in A$ and $b \in B$, there is a collider on π that is not an ancestor of C, or there is a non-collider on π in C. The sets A and B are *d-connected* given C if they are not d-separated, written $A \not\perp^d_G B \,|\, C$.

Definition 2. Given a DAG G and distribution $\mathbb{P}$, we say that $\mathbb{P}$ is *Markov with respect to* G if for all $A, B, C \subseteq V$ we have

$$A \perp^d_G B \,|\, C \implies X_A \perp\!\!\!\perp_{\mathbb{P}} X_B \,|\, X_C. \tag{1}$$

For a pair $(G, \mathbb{P})$, this is also referred to as the *global Markov property*.

Theorem 3 (Verma and Pearl, 1990)**.** *The global Markov property holds for all Bayesian networks.*

For a general Bayesian network, the set of conditional independencies in its observational distribution $\mathbb{P}$ does not characterise the set of d-separations in G: we might have a d-connection $A \not\perp^d_G B \,|\, C$ but still have a conditional independence $X_A \perp\!\!\!\perp_{\mathbb{P}} X_B \,|\, X_C$. A Bayesian network is called *faithful* if these cases are excluded:

Definition 3. Given a DAG G and distribution $\mathbb{P}$, we say that $\mathbb{P}$ is *faithful with respect to* G if for all $A, B, C \subseteq V$ we have

$$A \not\perp^d_G B \,|\, C \implies X_A \not\perp\!\!\!\perp_{\mathbb{P}} X_B \,|\, X_C.$$

A Bayesian network is faithful if its observational distribution is faithful with respect to its graph.

Example 1. The following Bayesian networks are unfaithful. The corresponding graphs G^a, G^b and G^c are depicted in Figure 1.

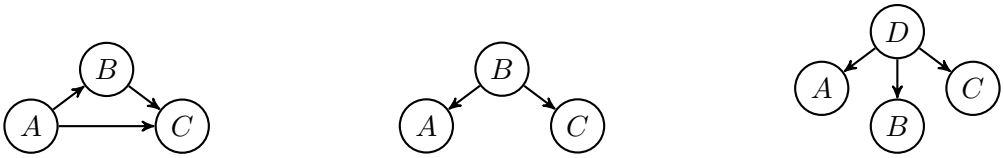
a) Cancelling paths: let $\mathbb{P}(X_A)$ be any distribution and let $\mathbb{P}(X_B \,|\, X_A) = \mathcal{N}(\beta_{AB} X_A, \sigma_B^2)$ and $\mathbb{P}(X_C \,|\, X_A, X_B) = \mathcal{N}(\beta_{AC} X_A + \beta_{BC} X_B, \sigma_C^2)$ for given variances $\sigma_A^2, \sigma_B^2, \sigma_C^2 > 0$ and coefficients $\beta_{AC}, \beta_{AB}, \beta_{BC} \in \mathbb{R}$ with $\beta_{AC} = -\beta_{AB}\beta_{BC}$. Then $A \not\perp^d_{G^a} C$ and $X_A \perp\!\!\!\perp X_C$.[3]

b) Deterministic variables: let $\mathbb{P}(X_A \,|\, X_B)$ and $\mathbb{P}(X_C \,|\, X_B)$ be Markov kernels and let $\mathbb{P}(X_B) = \delta_{x_B}$ for some $x_B \in \mathcal{X}_B$, so X_B deterministically has the value x_B. Then we have $A \not\perp^d_{G^b} C$ and $X_A \perp\!\!\!\perp X_C$.

[2] In statistical literature, Bayesian networks are often defined as joint distributions $\mathbb{P}(X_V)$, from which one must deduce the conditional distributions, which are not uniquely defined. For *causal* modelling, it is more suitable to model the Bayesian network as a set of Markov kernels, which uniquely specify the effects of interventions. Our results about the typicality of faithfulness cover both viewpoints: they are shown in the space of observational distributions, and in the space of Bayesian networks.

[3] A realistic example is when opening up a window (A) signals the thermostat to turn up the heating (B), so the inflow of cold air is perfectly offset by the heating, causing a net zero effect on room temperature (C).

c) Deterministic relations: let $\mathbb{P}(X_A \,|\, X_D)$ and $\mathbb{P}(X_C \,|\, X_D)$ be Markov kernels and $\mathbb{P}(X_D)$ any distribution and let $\mathbb{P}(X_B \,|\, X_D) = \delta_{X_D}$, so we deterministically set $X_B = X_D$. Then we have $A \not\perp\!\!\!\perp^d_{G^c} C \,|\, B$ and $X_A \perp\!\!\!\perp X_C \,|\, X_B$.[4]



(a) DAG G^a (b) DAG G^b (c) DAG G^c

Figure 1: DAGs of the Bayesian networks that are given in Example 1.

An important step in our proof of the typicality of faithful distributions, is that conditional independence is a topologically closed property, which means that it is preserved by taking limits. Whether this holds depends on the particular choice of the topology on $\mathcal{P}(\mathcal{X}_V)$. A well-known topology is the one related to weak convergence: given probability measures $\mathbb{P}, \mathbb{P}_1, \mathbb{P}_2, ... \in \mathcal{P}(\mathcal{X}_V)$ we say that $\mathbb{P}_n$ *converges weakly to* $\mathbb{P}$ (also known as *convergence in distribution*) if $\mathbb{E}_{\mathbb{P}_n}[f] \to \mathbb{E}_{\mathbb{P}}[f]$ for all continuous functions $f : \mathcal{X}_V \to [-1, 1]$, denoted by $\mathbb{P}_n \xrightarrow{w} \mathbb{P}$. This topology is metrised by the *bounded-Lipschitz* metric $d_{BL}(\mathbb{P}, \mathbb{Q}) := \sup_f \left| \int f \mathrm{d}\mathbb{P} - \int f \mathrm{d}\mathbb{Q} \right|$, where the supremum is taken over all functions $f : \mathcal{X}_v \to [-1, 1]$ with $\mathrm{Lip}(f) \leq 1$, where $\mathrm{Lip}(f)$ denotes the Lipschitz constant of f (Bogachev, 2007, Theorem 8.3.2). However, weak convergence does not necessarily preserve conditional independence: for a weakly convergent sequence $\mathbb{P}_n \to \mathbb{P}$ with $X_A \perp\!\!\!\perp_{\mathbb{P}_n} X_B \,|\, X_C$ for all $n \in \mathbb{N}$, we might have $X_A \not\perp\!\!\!\perp_{\mathbb{P}} X_B \,|\, X_C$ in the limit; see e.g. Lauritzen (1996), pp. 38-39 for an example. If the sample space $\mathcal{X}_C$ is uncountable, then the set $\{\mathbb{P} : X_A \perp\!\!\!\perp_{\mathbb{P}} X_B \,|\, X_C\}$ is even dense in the weak topology on $\mathcal{P}(\mathcal{X}_{A \cup B \cup C})$ (Boeken et al., 2026).

Due to this fact, we also consider the total variation topology on $\mathcal{P}(\mathcal{X}_V)$, induced by the *total variation metric* $d_{TV}(\mathbb{P}, \mathbb{Q}) := \sup_{A \in \mathcal{B}(\mathcal{X}_V)} |\mathbb{P}(A) - \mathbb{Q}(A)|$. This can equivalently be written as $d_{TV}(\mathbb{P}, \mathbb{Q}) = \sup_f \left| \int f \mathrm{d}\mathbb{P} - \int f \mathrm{d}\mathbb{Q} \right|$ where the supremum is taken over all measurable functions $f : \mathcal{X}_V \to [-1, 1]$, so it is generally stronger than weak convergence. Convergence in this metric is denoted by $\mathbb{P}_n \xrightarrow{tv} \mathbb{P}$. By Lauritzen (2024) we have that conditional independence is closed in total variation:

Theorem 4 (Lauritzen, 2024)**.** *Given probability measures $\mathbb{P}, \mathbb{P}_1, \mathbb{P}_2, ... \in \mathcal{P}(\mathcal{X}_V)$ such that $\mathbb{P}_n \xrightarrow{tv} \mathbb{P}$, if we have $X_A \perp\!\!\!\perp_{\mathbb{P}_n} X_B \,|\, X_C$ for all $n \in \mathbb{N}$, then also $X_A \perp\!\!\!\perp_{\mathbb{P}} X_B \,|\, X_C$.*

3 Unconstrained Bayesian networks

In this section, we consider the typicality of faithfulness in the unconstrained class of Bayesian networks; that is, Bayesian networks without any assumptions on the Markov kernels. Faithfulness is a property of the observational distribution of a Bayesian network, so we first consider in Section 3.1 the typicality of faithful observational distributions in the space of all distributions that are Markov with respect to the given DAG, equipped with the total variation metric. In Section 3.2 we consider the typicality of faithfulness in the space of Bayesian networks (which are Markov with respect to the given DAG G) equipped with a newly introduced metric.

[4]For Bayesian networks with known deterministic variables or relations, Geiger et al. (1990) introduced the *D-separation* criterion. This takes into account part of the determinism to deduce additional conditional independencies that are not implied by the *d*-separation criterion.

3.1 Typicality of faithfulness in the space of observational distributions

Given a DAG $G = (V, E)$, we consider the following sets of Markov, faithful, and unfaithful distributions relative to G:

$$M_G := \left\{ \mathbb{P} \in \mathcal{P}(\mathcal{X}_V) : A \perp_G^d B \,|\, C \implies X_A \perp\!\!\!\perp_{\mathbb{P}} X_B \,|\, X_C \text{ for all } A, B, C \subseteq V \right\}$$
$$F_G := \left\{ \mathbb{P} \in M_G : A \not\perp_G^d B \,|\, C \implies X_A \not\perp\!\!\!\perp_{\mathbb{P}} X_B \,|\, X_C \text{ for all } A, B, C \subseteq V \right\}$$
$$U_G := M_G \setminus F_G.$$

We will derive properties of F_G and U_G as subsets of the metric space (M_G, d_{TV}); in later sections we will add regularity conditions on M_G, F_G and U_G, and consider other topologies. First, if we let $I_{A,B|C} = \{\mathbb{P} \in \mathcal{P}(\mathcal{X}_V) : X_A \perp\!\!\!\perp_{\mathbb{P}} X_B \,|\, X_C\}$, note that we can write

$$M_G = \mathcal{P}(\mathcal{X}_V) \cap \bigcap_{A \perp_G^d B \,|\, C} I_{A,B|C}, \qquad F_G = M_G \cap \bigcap_{A \not\perp_G^d B \,|\, C} (M_G \setminus I_{A,B|C}), \qquad U_G = M_G \setminus F_G.$$

From Theorem 4 it is immediate that M_G is a closed subspace of $\mathcal{P}(\mathcal{X}_V)$, and that F_G is open in M_G. For our most general nonparametric result, it remains to show that F_G is dense. The following result states that the set of distributions that are Markov *and* have a particular conditional dependence is dense in total variation. The proof refers to technical lemmas that are provided in Section 3.1.1.

Lemma 1. *For every $\mathbb{P} \in M_G$ and every $A, B, C \subseteq V$ such that $A \not\perp_G^d B \,|\, C$, there is a sequence $\mathbb{P}_1, \mathbb{P}_2, \ldots \in M_G$ such that $X_A \not\perp\!\!\!\perp_{\mathbb{P}_n} X_B \,|\, X_C$ for all $n \in \mathbb{N}$ and $\mathbb{P}_n \overset{tv}{\to} \mathbb{P}$.*

Proof. Let $\mathbb{P} \in M_G$ be given be such that $X_A \perp\!\!\!\perp_{\mathbb{P}} X_B \,|\, X_C$ — otherwise the result holds trivially. By Lemma 2, there exists a $\mathbb{P}_1$ that is Markov and has $X_A \not\perp\!\!\!\perp_{\mathbb{P}_1} X_B \,|\, X_C$. The interpolation $(\mathbb{P}_\lambda)_{\lambda \in (0,1)}$ between $\mathbb{P}$ and $\mathbb{P}_1$ from Definition 4 lies in M_G, and has that $X_A \not\perp\!\!\!\perp_{\mathbb{P}_\lambda} X_B \,|\, X_C$ for all positive λ below some $\lambda^* \in (0, 1)$ (Lemma 3), which converges in total variation to $\mathbb{P}$ as $\lambda \to 0$ (Lemma 4). One obtains a suitable sequence by setting $\mathbb{P}_n := \mathbb{P}_{\lambda^*/2n}$. ■

In other words, the set $\{\mathbb{P} \in M_G : X_A \not\perp\!\!\!\perp_{\mathbb{P}} X_B \,|\, X_C\}$ is dense in M_G. As a corollary that might be of independent interest, we have that conditional dependence is dense in total variation.

Corollary 1. *The set $\{\mathbb{P} \in \mathcal{P}(\mathcal{X}_V) : X_A \not\perp\!\!\!\perp_{\mathbb{P}} X_B \,|\, X_C\}$ is dense in $(\mathcal{P}(\mathcal{X}_V), d_{TV})$.*

Proof. Let G be a fully connected DAG with vertices V, then $M_G = \mathcal{P}(\mathcal{X}_V)$ and the result follows from Lemma 1. ■

Our first result concerning the faithfulness of nonparametric Bayesian networks is as follows.

Theorem 5. *Given a DAG G, the set of faithful distributions F_G is non-empty, open and dense, and the unfaithful distributions U_G are nowhere dense in (M_G, d_{TV}).*

Proof. By Theorem 4 and Lemma 1 we have for any given $A, B, C \subseteq V$ with $A \not\perp_G^d B \,|\, C$ that $M_G \setminus I_{A,B|C}$ is dense and open in M_G. Hence, F_G is a dense open set as it is a finite intersection of dense open sets. Since M_G is non-empty (take for example a product of independent distributions), the dense set F_G is non-empty as well, proving the existence of a faithful distribution. Finally, U_G is the complement of a dense open set, hence nowhere dense. ■

To conclude, unfaithful distributions are 'atypical': there is no non-empty open set of distributions that are Markov with respect to G, in which any faithful distribution in this set can be approximated by unfaithful ones. This loosely says that there is no 'cluster' of unfaithful distributions.

3.1.1 Conditional dependence is dense in total variation

In this section, we fill in the details of the proof of Theorem 5.

Lemma 2. *For any DAG G, standard Borel space $\mathcal{X}_V$ and subsets $A, B, C \subseteq V$ such that $A \not\perp\!\!\!\perp^d_G B \,|\, C$, there exists a distribution $\mathbb{P} \in M_G$ with the conditional dependence $X_A \not\perp\!\!\!\perp_{\mathbb{P}} X_B \,|\, X_C$.*

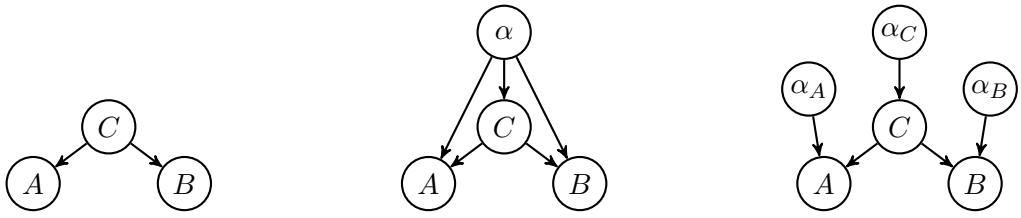
Proof. For each $v \in V$ pick an injective $f_v : \{0,1\} \to \mathcal{X}_v$ and note that sets $\{f_v(0)\}$ and $\{f_v(1)\}$ are measurable since $\mathcal{X}_v$ is standard Borel. We will construct a binary distribution on the image of f_V that has the required dependence. Note that without loss of generality we can assume that A and B are singletons: any $\mathbb{P}(X_V)$ with $X_A \not\perp\!\!\!\perp_{\mathbb{P}} X_B \,|\, X_C$ also has $X_{A'} \not\perp\!\!\!\perp_{\mathbb{P}} X_{B'} \,|\, X_C$ for supersets $A \subset A'$ and $B \subset B'$. Also, the given d-connection implies $A, B \notin C$. If we have $A = B$, for all $v \in V$ set $\mathbb{P}(X_v = f_v(0)) = p$ and $\mathbb{P}(X_v = f_v(1)) = 1 - p$ for some $p \in (0,1)$ and let $\mathbb{P}(X_V) = \bigotimes_{v \in V} \mathbb{P}(X_v)$. Then $\mathbb{P}(X_V)$ is Markov and $X_A \not\perp\!\!\!\perp_{\mathbb{P}} X_B \,|\, X_C$. If $A \neq B$, then by Meek (1998) Lemma 3,[5] there exists a distribution $\tilde{\mathbb{P}}$ on $\{0,1\}^{|V|}$ that is Markov with respect to G and which has the conditional dependence $X_A \not\perp\!\!\!\perp_{\tilde{\mathbb{P}}} X_B \,|\, X_C$, so there are $\tilde{x}_A, \tilde{x}_B, \tilde{x}_C \in \{0,1\}$ with $\tilde{\mathbb{P}}(\tilde{x}_C) > 0$ such that $\tilde{\mathbb{P}}(\tilde{x}_A, \tilde{x}_B \,|\, \tilde{x}_C) \neq \tilde{\mathbb{P}}(\tilde{x}_A \,|\, \tilde{x}_C)\tilde{\mathbb{P}}(\tilde{x}_B \,|\, \tilde{x}_C)$. Define the pushforward $\mathbb{P}(X_V) := \tilde{\mathbb{P}} \circ f_V^{-1}$, which has

$$\begin{aligned}\mathbb{P}\big(X_A = f_A(\tilde{x}_A), X_B = f_B(\tilde{x}_B) \,|\, X_C = f_C(\tilde{x}_C)\big) &= \tilde{\mathbb{P}}\big(\tilde{x}_A, \tilde{x}_B \,|\, \tilde{x}_C\big) \\ &\neq \tilde{\mathbb{P}}\big(\tilde{x}_A \,|\, \tilde{x}_C\big)\tilde{\mathbb{P}}\big(\tilde{x}_B \,|\, \tilde{x}_C\big) = \mathbb{P}\big(X_A = f_A(\tilde{x}_A) \,|\, X_C = f_C(\tilde{x}_C)\big)\mathbb{P}\big(X_B = f_B(\tilde{x}_B) \,|\, X_C = f_C(\tilde{x}_C)\big)\end{aligned}$$

so indeed $X_A \not\perp\!\!\!\perp_{\mathbb{P}} X_B \,|\, X_C$. By a similar reasoning, for any $A, B, C \subseteq V$ a conditional independence $X_A \perp\!\!\!\perp_{\tilde{\mathbb{P}}} X_B \,|\, X_C$ implies $X_A \perp\!\!\!\perp_{\mathbb{P}} X_B \,|\, X_C$, and thus $\mathbb{P} \in M_G$. ∎

Next, we aim to construct an interpolation of any two given $\mathbb{P}_0, \mathbb{P}_1 \in M_G$, within M_G. Naively taking a mixture of the observational distributions does not give a distribution that is Markov with respect to G, as is shown in the following example.

Example 2. Let $(\mathbb{P}_i(X_A \,|\, X_C), \mathbb{P}_i(X_B \,|\, X_C), \mathbb{P}_i(X_C))$ for $i \in \{0,1\}$ be Bayesian networks with DAG G as depicted in Figure 2a, which both have $X_A \perp\!\!\!\perp X_B \,|\, X_C$. A mixture of the observational distributions $\tilde{\mathbb{P}}_\lambda(X_A, X_B, X_C) := (1-\lambda)\mathbb{P}_0(X_A, X_B, X_C) + \lambda\mathbb{P}_1(X_A, X_B, X_C)$ would correspond to the $(A \cup B \cup C)$-marginal of the Bayesian network $(\mathbb{P}_\alpha(X_A \,|\, X_C), \mathbb{P}_\alpha(X_B \,|\, X_C), \mathbb{P}_\alpha(X_C), \mathbb{P}(\alpha))$ with $\alpha \sim \text{Bernoulli}(\lambda)$. Its graph is depicted in Figure 2b, from which we see that $\tilde{\mathbb{P}}_\lambda$ need not be Markov with respect to G, as we might have $X_A \not\perp\!\!\!\perp_{\mathbb{P}_\lambda} X_B \,|\, X_C$. Instead, taking a mixture of the Markov kernels of the Bayesian networks gives $(\mathbb{P}_{\alpha_A}(X_A \,|\, X_C), \mathbb{P}_{\alpha_B}(X_B \,|\, X_C), \mathbb{P}_{\alpha_C}(X_C), \mathbb{P}(\alpha_A), \mathbb{P}(\alpha_B), \mathbb{P}(\alpha_C))$ with $\alpha_A, \alpha_B, \alpha_C \sim \text{Bernoulli}(\lambda)$ i.i.d., whose $(A \cup B \cup C)$-marginal $\mathbb{P}_\lambda(X_A, X_B, X_C)$ (see Definition 4 below) is Markov with respect to G (see Figure 2c).



(a) DAG G (b) Non-Markov mixture (c) Markov mixture

Figure 2: Graphs relating to different mixtures of Bayesian networks with graph G.

The issue that is detailed in the previous example is resolved in Definition 4.

Definition 4. Given a DAG G and two distributions $\mathbb{P}_0, \mathbb{P}_1 \in M_G$ define the interpolation

$$\mathbb{P}_\lambda(X_V) := \bigotimes_{v \in V} \left((1-\lambda)\mathbb{P}_0(X_v \,|\, X_{\text{pa}(v)}) + \lambda\mathbb{P}_1(X_v \,|\, X_{\text{pa}(v)})\right).$$

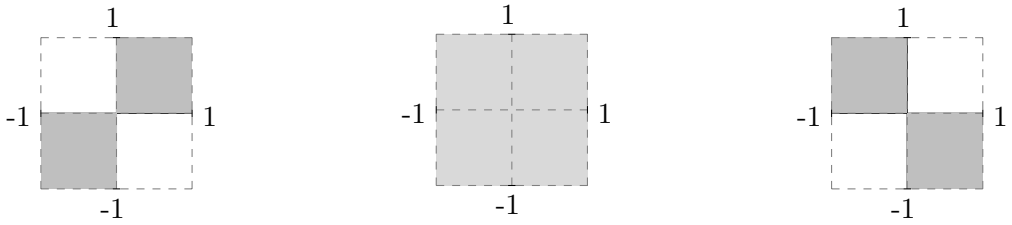
[5] Meek (1995) proves this result assuming weak transitivity of binary distributions, which does not hold in general. Meek (1998) provides a correct proof based on *marginal* weak transitivity.

When $\mathbb{P}_0$ and $\mathbb{P}_1$ are the observational distributions of Bayesian networks $m_0 = (\mathbb{P}_0(X_v \,|\, X_{\mathrm{pa}(v)}))_{v\in V}$ and $m_1 = (\mathbb{P}_1(X_v \,|\, X_{\mathrm{pa}(v)}))_{v\in V}$, this essentially amounts to considering the observational distribution $\mathbb{P}_\lambda$ of the mixture $(1-\lambda)m_0 + \lambda m_1$ of the Bayesian networks in the sense of Figure 2c — this will explicitly be deployed in Sections 3.2 and 5.1. It is therefore immediate that $\mathbb{P}_\lambda \in M_G$ for all $\lambda \in [0,1]$. If $\mathbb{P}_0$ and $\mathbb{P}_1$ have densities p_0 and p_1 with respect to some measure $\mathbb{Q}$, then $\mathbb{P}_\lambda$ has a density p_λ given by the expansion

$$\begin{aligned} p_\lambda(x_V) &= \prod_{v\in V} \big((1-\lambda)p_0(x_v \,|\, x_{\mathrm{pa}(v)}) + \lambda p_1(x_v \,|\, x_{\mathrm{pa}(v)})\big) \\ &= \sum_{\alpha\in\{0,1\}^d} (1-\lambda)^{d-|\alpha|}\lambda^{|\alpha|} p_{\alpha_d}(x_{v_d} \,|\, x_{\mathrm{pa}(v_d)}) ... p_{\alpha_1}(x_{v_1}) \end{aligned} \tag{2}$$

where $d = |V|$ and $(v_1, ..., v_d)$ is a topological ordering of G. Our goal is to show that if we have conditional dependence $X_A \not\perp\!\!\!\perp_{\mathbb{P}_1} X_B \,|\, X_C$ in $\mathbb{P}_1$, then it is maintained in the interpolation $\mathbb{P}_\lambda$ as λ approaches 0. This is not immediate, as shown in the following example.

Example 3. Consider a Bayesian network with variables X, Y taking values in the interval $[-1,1]$ and graph $X \to Y$. Let $\mathbb{P}_0(X,Y)$ be a uniform distribution on $(0,1)\times(0,1) \cup (-1,0)\times(-1,0)$ and $\mathbb{P}_1$ a uniform distribution on $(-1,0)\times(0,1) \cup (0,1)\times(-1,0)$. The interpolation $\mathbb{P}_\lambda$ has a uniform distribution on $(-1,1)^2$ for $\lambda = 1/2$, and thus an independence $X \perp\!\!\!\perp Y$. This is graphically depicted in Figure 3.



(a) $\mathbb{P}_0 : X \not\perp\!\!\!\perp Y$ (b) $\mathbb{P}_{\frac{1}{2}} : X \perp\!\!\!\perp Y$ (c) $\mathbb{P}_1 : X \not\perp\!\!\!\perp Y$

Figure 3: Mixtures of dependent variables can become independent.

Nevertheless, given $\mathbb{P}_0$ and $\mathbb{P}_1$, the dependence is maintained on an interval $(0, \lambda^*) \subset (0,1)$, as shown by the following result.

Lemma 3. *Given distributions $\mathbb{P}_0, \mathbb{P}_1 \in M_G$ with dependence $X_A \not\perp\!\!\!\perp_{\mathbb{P}_1} X_B \,|\, X_C$, there exists for the interpolation $\mathbb{P}_\lambda$ from Definition 4 a $\lambda^* \in (0,1)$ such that $X_A \not\perp\!\!\!\perp_{\mathbb{P}_\lambda} X_B \,|\, X_C$ for all $\lambda \in (0, \lambda^*)$.*

Proof. Define $\mathbb{Q} := \mathbb{P}_0 + \mathbb{P}_1$, and let p_0, p_1, p_λ be densities of $\mathbb{P}_0, \mathbb{P}_1$ and $\mathbb{P}_\lambda$ with respect to $\mathbb{Q}$. Note that conditional dependence $X_A \not\perp\!\!\!\perp_{\mathbb{P}_1} X_B \,|\, X_C$ is equivalent to the existence of measurable E_A and E_B such that $E'_C := \{x_C : \mathbb{P}_1(X_A \in E_A, X_B \in E_B \,|\, x_C) \neq \mathbb{P}_1(X_A \in E_A \,|\, x_C)\mathbb{P}_1(X_B \in E_B \,|\, x_C)\}$ has $\mathbb{P}_1(E'_C) > 0$, hence there is a non-empty measurable subset $E_C \subseteq E'_C$ such that $p_1(x_C) > 0$ and $\mathbb{P}_1(E_C) > 0$ for all $x_C \in E_C$. For $x_C \in E_C$ we have the equivalence

$$\begin{aligned} &\mathbb{P}_1(X_A \in E_A, X_B \in E_B \,|\, X_C = x_C) \neq \mathbb{P}_1(X_A \in E_A \,|\, X_C = x_C)\mathbb{P}_1(X_B \in E_B \,|\, X_C = x_C) \\ \iff &\int_{E_A\times E_B} p_1(x_A, x_B \,|\, x_C)\mathrm{d}\mathbb{Q}(x_A, x_B) \neq \int_{E_A} p_1(x_A \,|\, x_C)\mathrm{d}\mathbb{Q}(x_A) \int_{E_B} p_1(x_B \,|\, x_C)\mathrm{d}\mathbb{Q}(x_B) \\ \iff &\int_{E_A\times E_B} p_1(x_A, x_B, x_C)p_1(x_C)\mathrm{d}\mathbb{Q}(x_A, x_B) \neq \int_{E_A} p_1(x_A, x_C)\mathrm{d}\mathbb{Q}(x_A) \int_{E_B} p_1(x_B, x_C)\mathrm{d}\mathbb{Q}(x_B). \end{aligned}$$

Define

$$\begin{aligned} q(\lambda, x_C) := &\int_{E_A\times E_B} p_\lambda(x_A, x_B, x_C)p_\lambda(x_C)\mathrm{d}\mathbb{Q}(x_A, x_B) \\ &- \int_{E_A} p_\lambda(x_A, x_C)\mathrm{d}\mathbb{Q}(x_A) \int_{E_B} p_\lambda(x_B, x_C)\mathrm{d}\mathbb{Q}(x_B), \end{aligned}$$

for which we have $q(1, x_C) \neq 0$ for all $x_C \in E_C$. From (2) we see that $q(\lambda, x_C)$ is a non-zero polynomial in λ for every $x_C \in E_C$, and so $q(\lambda, x_C) \neq 0$ for all $\lambda \in (0, \lambda^*(x_C))$ with $\lambda^*(x_C) := \min\{\{1\} \cup R(x_C)\}$ where $R(x_C) := \{\lambda \in (0, 1] : q(\lambda, x_C) = 0\}$. Since $q(\lambda, x_C)$ is a *Carathéodory function* (continuous in λ and measurable in x_C), we have by Aliprantis and Border (2006), Corollary 18.8, that the correspondence $\bar{R}(x_C) := \{\lambda \in [0, 1] : q(\lambda, x_C) = 0\}$ is *weakly measurable* (i.e. $\{x_C : \bar{R}(x_C) \cap A \neq \emptyset\} \in \mathcal{B}(\mathcal{X}_C)$ for all open A). Since for each open A we have $R(x_C) \cap A = \bar{R}(x_C) \cap (A \setminus \{0\})$, the correspondence $R(x_C)$ is weakly measurable as well. It then follows from Aliprantis and Border (2006), Theorem 18.19 that the selection $\lambda^*(x_C)$ is measurable. Our goal is to show that there is a $\lambda^* \in (0, 1)$ (independent of x_C) and a set $E_C^* \in \mathcal{B}(\mathcal{X}_C)$ with $\mathbb{P}_\lambda(E_C^*) > 0$ and $q(\lambda, x_C) \neq 0$ for all $\lambda \in (0, \lambda^*)$ and all $x_C \in E_C^*$, which would imply that $X_A \not\perp\!\!\!\perp_{\mathbb{P}_\lambda} X_B \,|\, X_C$ for all $\lambda \in (0, \lambda^*)$.[6] Define $E_C^n := \{x_C \in E_C : \lambda^*(x_C) > 1/n\}$, then $E_C^1 \subseteq E_C^2 \subseteq ... \subseteq E_C$ with $\lim_n \mathbb{P}_1(E_C^n) = \mathbb{P}_1(E_C) > 0$, so there exists an N such that $\mathbb{P}_1(E_C^n) > 0$ for all $n \geq N$. Setting $\lambda^* := 1/N$ and $E_C^* := E_C^N$ we get $q(\lambda, x_C) \neq 0$ for all $\lambda \in (0, \lambda^*)$ for all $x_C \in E_C^*$. Since $\mathbb{P}_1$ is absolutely continuous with respect to $\mathbb{P}_\lambda$ for all $\lambda \in (0, 1)$ we also have $\mathbb{P}_\lambda(E_C^*) > 0$, implying that $X_A \not\perp\!\!\!\perp_{\mathbb{P}_\lambda} X_B \,|\, X_C$ for all $\lambda \in (0, \lambda^*)$, which is the desired result. ■

Lemma 4. *Given two distributions $\mathbb{P}_0, \mathbb{P}_1 \in M_G$, we have for the interpolation $\mathbb{P}_\lambda$ from Definition 4 that $\mathbb{P}_\lambda \overset{tv}{\to} \mathbb{P}_0$ as $\lambda \to 0$.*

Proof. Define $\mathbb{Q} := \mathbb{P}_0 + \mathbb{P}_1$, and let p_0, p_1, p_λ be densities of $\mathbb{P}_0, \mathbb{P}_1$ and $\mathbb{P}_\lambda$ with respect to $\mathbb{Q}$. From (2) we get the expression

$$p_\lambda(x_V) = (1-\lambda)^d p_0(x_V) + \sum_{\substack{\alpha \in \{0,1\}^d \\ |\alpha| > 0}} (1-\lambda)^{d-|\alpha|} \lambda^{|\alpha|} p_{\alpha_d}(x_{v_d} \,|\, x_{\mathrm{pa}(v_d)}) ... p_{\alpha_1}(x_{v_1})$$

so we have pointwise convergence $p_\lambda(x_V) \to p_0(x_V)$ as $\lambda \to 0$. By Scheffé (1947) we conclude that $\mathbb{P}_\lambda \overset{tv}{\to} \mathbb{P}_0$. ■

3.2 Typicality of faithfulness in the space of Bayesian networks

In this section we extend Theorem 5 from the space of observational distributions to the space of Bayesian networks, defined as follows:

Definition 5. Given a DAG G with finite index set V and standard Borel space $\mathcal{X}_v$ for every $v \in V$, *the space of Bayesian networks with graph G* is defined as

$$\mathrm{BN}_G := \prod_{v \in V} \left\{ \mathbb{P}(X_v \,|\, X_{\mathrm{pa}(v)}) : \mathcal{X}_{\mathrm{pa}(v)} \to \mathcal{P}(\mathcal{X}_v) \text{ measurable} \right\}.$$

The faithfulness of a Bayesian network is a property of its observational distribution $\mathbb{P} \in M_G$. To formalise the relation between the Bayesian network and the observational distribution we introduce the *distribution* map, defined as

$$D : \mathrm{BN}_G \to M_G, \quad (\mathbb{P}(X_v \,|\, X_{\mathrm{pa}(v)}))_{v \in V} \mapsto \bigotimes_{v \in V} \mathbb{P}(X_v \,|\, X_{\mathrm{pa}(v)}).$$

We are interested in whether the faithful Bayesian networks $D^{-1}(F_G)$ are typical in BN_G. To get a well-defined notion of typicality we require a topology on BN_G, which we introduce via the following metric.

Definition 6. For $m, m' \in \mathrm{BN}_G$, consider the following metric:

$$d_{TV}^\circ(m, m') := \sum_{v \in V} \sup_{x_{\mathrm{pa}(v)}} d_{TV}(\mathbb{P}_m(X_v \,|\, x_{\mathrm{pa}(v)}), \mathbb{P}_{m'}(X_v \,|\, x_{\mathrm{pa}(v)})).$$

One readily verifies that this is indeed a metric.

[6] It would be more straightforward to show that $\mathbb{P}_\lambda(X_A \in E_A, X_B \in E_B \,|\, X_C \in E_C) \neq \mathbb{P}_\lambda(X_A \in E_A \,|\, X_C \in E_C)\mathbb{P}_\lambda(X_B \in E_B \,|\, X_C \in E_C)$ for some E_A, E_B, E_C with $\mathbb{P}_\lambda(E_C) > 0$, but this does not imply the conditional dependence $X_A \not\perp\!\!\!\perp_{\mathbb{P}_\lambda} X_B \,|\, X_C$, hence the conditioning on the individual $x_C \in E_C$. See also Neykov et al. (2021), p.3.

This metric measures the worst-case total variation distance between corresponding Markov kernels, summed over all vertices. Convergence $d^{\circ}_{TV}(m_n, m) \to 0$ means that all conditional distributions $\mathbb{P}_{m_n}(X_v \,|\, x_{\mathrm{pa}(v)})$ converge uniformly in $x_{\mathrm{pa}(v)}$ to $\mathbb{P}_m(X_v \,|\, x_{\mathrm{pa}(v)})$ in total variation. Note that this is a strong notion of closeness: it requires agreement of the Markov kernels for *all* parent values, including those that might have zero probability under the observational distribution. Consequently, our metric distinguishes Bayesian networks that have the same observational distribution but differ on a measure-zero set of parent values. This metric is natural for Bayesian networks as causal models, where by intervening the Markov kernels represent mechanisms which could be evaluated on *all* parent values — not just those occurring under a specific distribution of the parent variables.

Lemma 5. *The distribution map $D : (\mathrm{BN}_G, d^{\circ}_{TV}) \to (M_G, d_{TV})$ is continuous.*

Proof. Let $m_0, m_1 \in \mathrm{BN}_G$ and write $\mathbb{P}_0 := D(m_0)$ and $\mathbb{P}_1 := D(m_1)$. Let $v_1, ..., v_d$ be a reverse topological ordering of G (which has $\mathrm{pa}(v_k) \subseteq v_{k+1}, ..., v_d$), and let $\mathbb{Q}_k := \bigotimes_{i=1}^{k} \mathbb{P}_1(X_{v_i} \,|\, X_{\mathrm{pa}(v_i)}) \otimes \bigotimes_{i=k+1}^{d} \mathbb{P}_0(X_{v_i} \,|\, X_{\mathrm{pa}(v_i)})$, then $\mathbb{Q}_0 = \mathbb{P}_0$ and $\mathbb{Q}_d = \mathbb{P}_1$, so we have $d_{TV}(\mathbb{P}_0, \mathbb{P}_1) \leq \sum_{k=1}^{d} d_{TV}(\mathbb{Q}_{k-1}, \mathbb{Q}_k)$ by the triangle inequality. Given a measurable function $f : \mathcal{X}_V \to [-1, 1]$, define the functions

$$g^k_f(x_{v_k}, ..., x_{v_d}) := \int f(x_{v_1}, ..., x_{v_d}) \mathrm{d} \bigotimes_{i=1}^{k-1} \mathbb{P}_1(X_{v_i} \,|\, X_{\mathrm{pa}(v_i)})$$

$$h^k_f(x_{v_k}, x_{\mathrm{pa}(v_k)}) := \int g^k_f(x_{v_k}, x_{\mathrm{pa}(v_k)}, ...) \mathrm{d}\mathbb{P}_0(X_{\{v_{k+1}, ..., v_d\} \setminus \mathrm{pa}(v_k)}),$$

which are bounded and measurable. Note that $\mathbb{Q}_{k-1}$ and $\mathbb{Q}_k$ only differ at the Markov kernel for v_k, having $\mathbb{P}_0(X_{v_k} \,|\, X_{\mathrm{pa}(v_k)})$ and $\mathbb{P}_1(X_{v_k} \,|\, X_{\mathrm{pa}(v_k)})$ at that position respectively. By taking the supremum over all measurable $f : \mathcal{X}_V \to [-1, 1]$ we then bound $d_{TV}(\mathbb{Q}_{k-1}, \mathbb{Q}_k)$ as

$$\begin{aligned}
&\sup_f \left| \int f d\mathbb{Q}_{k-1} - \int f \mathrm{d}\mathbb{Q}_k \right| \\
&= \sup_f \left| \int \left(\int g^k_f \mathrm{d}\mathbb{P}_0(X_{v_k} \,|\, X_{\mathrm{pa}(v_k)}) - \int g^k_f \mathrm{d}\mathbb{P}_1(X_{v_k} \,|\, X_{\mathrm{pa}(v_k)}) \right) \mathrm{d}\mathbb{P}_0(X_{v_{k+1}}, ..., X_{v_d}) \right| \\
&\leq \sup_{x_{\mathrm{pa}(v_k)}} \sup_f \left| \int \left(\int g^k_f \mathrm{d}\mathbb{P}_0(X_{v_k} \,|\, x_{\mathrm{pa}(v_k)}) - \int g^k_f \mathrm{d}\mathbb{P}_1(X_{v_k} \,|\, x_{\mathrm{pa}(v_k)}) \right) \mathrm{d}\mathbb{P}_0(X_{\{v_{k+1}, ..., v_d\} \setminus \mathrm{pa}(v_k)}) \right| \\
&= \sup_{x_{\mathrm{pa}(v_k)}} \sup_f \left| \int h^k_f(x_{v_k}, x_{\mathrm{pa}(v_k)}) \mathrm{d}\mathbb{P}_0(X_{v_k} \,|\, x_{\mathrm{pa}(v_k)}) - \int h^k_f(x_{v_k}, x_{\mathrm{pa}(v_k)}) \mathrm{d}\mathbb{P}_1(X_{v_k} \,|\, x_{\mathrm{pa}(v_k)}) \right| \\
&\leq \sup_{x_{\mathrm{pa}(v_k)}} d_{TV}(\mathbb{P}_0(X_{v_k} \,|\, x_{\mathrm{pa}(v_k)}), \mathbb{P}_1(X_{v_k} \,|\, x_{\mathrm{pa}(v_k)})),
\end{aligned}$$

using Fubini's Theorem. This gives $d_{TV}(\mathbb{P}_0(X_V), \mathbb{P}_1(X_V)) \leq d^{\circ}_{TV}(m_0, m_1)$. ∎

Contrasting with the typicality of faithful distributions of Bayesian networks, we now show that Bayesian networks themselves are typically faithful.

Theorem 6. *The set of faithful Bayesian networks $D^{-1}(F_G)$ is open and dense in $(\mathrm{BN}_G, d^{\circ}_{TV})$.*

Proof. From Lemma 5 and Theorem 5 it follows that the set of faithful Bayesian networks $D^{-1}(F_G)$ is open and non-empty. Let $m_0 \in \mathrm{BN}_G$, $m_1 \in D^{-1}(F_G)$ and $m_\lambda := (1 - \lambda)m_0 + \lambda m_1$, then

$$\begin{aligned}
d^{\circ}_{TV}(m_\lambda, m_0) &= \sum_{v \in V} \sup_{x_{\mathrm{pa}(v)}} d_{TV}(\mathbb{P}_\lambda(X_v \,|\, x_{\mathrm{pa}(v)}), \mathbb{P}_0(X_v \,|\, x_{\mathrm{pa}(v)})) \\
&= \lambda \sum_{v \in V} \sup_{x_{\mathrm{pa}(v)}} d_{TV}(\mathbb{P}_1(X_v \,|\, x_{\mathrm{pa}(v)}), \mathbb{P}_0(X_v \,|\, x_{\mathrm{pa}(v)})) \\
&= \lambda d^{\circ}_{TV}(m_1, m_0) \to 0
\end{aligned}$$

as $\lambda \to 0$. By repeated application of Lemma 3 for every d-connection in G, there is a $\lambda^* \in (0, 1)$ such that $m_\lambda \in F_G$ for all $\lambda \in (0, \lambda^*)$, so $D^{-1}(F_G)$ is dense in BN_G. ∎

4 Conditional exponential family parametrisations

The preceding section raises the question whether the topological typicality of faithful Bayesian networks also holds for specific parametrisations of Bayesian networks. In this section we answer this question in the affirmative, by extending the results of Spirtes et al. (1993) and Meek (1995) to sufficiently regular conditional exponential family parametrisations of Bayesian networks.

Formally, a *parametrisation* of a Bayesian network with graph G is a set $\Theta \subseteq \mathbb{R}^d$ for some $d \in \mathbb{N}$ and a map

$$\varphi : \Theta \to \mathrm{BN}_G, \quad \theta \mapsto (\mathbb{P}_\theta(X_v \,|\, X_{\mathrm{pa}(v)}))_{v \in V}.$$

The corresponding map from the parameter values to the observational distribution is defined as

$$T : \Theta \to M_G, \quad T := D \circ \varphi.$$

We consider the question whether the set of *faithful parameters* $T^{-1}(F_G)$ is typical in Θ with respect to the Euclidean topology, in particular for the following class of conditional exponential family Bayesian networks, inspired by Feigin (1981).

Definition 7. Given a DAG G with vertices V and sample space $\mathcal{X}_v \subseteq \mathbb{R}$ for each $v \in V$, a *conditional exponential family Bayesian network* has for each $v \in V$ a conditional density

$$p_{\theta_v}(x_v \,|\, x_{\mathrm{pa}(v)}) = b_v(x_v, x_{\mathrm{pa}(v)}) e^{\eta_v(\theta_v)^\top t_v(x_v, x_{\mathrm{pa}(v)}) - A_v(\eta_v(\theta_v), x_{\mathrm{pa}(v)})}$$

with respect to a given locally finite dominating measure μ_v on $\mathcal{X}_v$, with parameter space $\Theta_v \subseteq \mathbb{R}^{d_v}$ for some $d_v \in \mathbb{N}$, a function $b_v : \mathcal{X}_v \times \mathcal{X}_{\mathrm{pa}(v)} \to [0, \infty)$, *natural parameter* $\eta_v : \Theta_v \to \mathbb{R}^{k_v}$ for some $k_v \in \mathbb{N}$ and *sufficient statistic* $t_v : \mathcal{X}_v \times \mathcal{X}_{\mathrm{pa}(v)} \to \mathbb{R}^{k_v}$ such that $A_v(\eta_v(\theta_v), x_{\mathrm{pa}(v)}) < \infty$ for all $\theta_v \in \Theta_v$ and $x_{\mathrm{pa}(v)} \in \mathcal{X}_{\mathrm{pa}(v)}$, where $A_v(\eta_v, x_{\mathrm{pa}(v)}) := \log \int b_v(x_v, x_{\mathrm{pa}(v)}) e^{\eta_v^\top t_v(x_v, x_{\mathrm{pa}(v)})} \mathrm{d}\mu_v(x_v)$.

This gives rise to a joint density $p_\theta(x_V) = \prod_{v \in V} p_{\theta_v}(x_v \,|\, x_{\mathrm{pa}(v)})$ of the distribution $\mathbb{P}_\theta(X_V) := T(\theta) \in M_G$ and the joint parameter space $\Theta := \prod_{v \in V} \Theta_v$. This model class allows the modelling of mixed data types, see e.g. Yang et al. (2014).

Example 4. For linear Gaussian Bayesian networks, Spirtes et al. (1993) parametrise for each $v \in V$ the conditional distribution $\mathbb{P}_\theta(X_v \,|\, X_{\mathrm{pa}(v)} = x_{\mathrm{pa}(v)}) = \mathcal{N}(\beta_v^\top x_{\mathrm{pa}(v)}, \sigma_v^2)$ by a linear coefficient β_v and a strictly positive variance σ_v^2. This gives the parameter space

$$\Theta_{\mathcal{N}} := \prod_{v \in V} \left\{ (\beta_v, \sigma_v^2) \in \mathbb{R}^{|\,\mathrm{pa}(v)|} \times \mathbb{R}_{>0} \right\},$$

so when writing $\theta_v = (\beta_v, \sigma_v^2)$ it has sufficient statistic $t_v(x_v, x_{\mathrm{pa}(v)})^\top = (x_v x_{\mathrm{pa}(v)}^\top, x_v^2)$, natural parameter $\eta_v(\theta_v)^\top = (\beta_v^\top / \sigma_v^2, -1/(2\sigma_v^2))$, $b_v = 1$ and as dominating measure μ_v the Lebesgue measure.

Example 5. For discrete distributions with finite state space, Meek (1995) considers for each conditional distribution $\mathbb{P}_\theta(X_v \,|\, X_{\mathrm{pa}(v)} = x_{\mathrm{pa}(v)})$ a parameter $\theta_{v, x_{\mathrm{pa}(v)}}$ in the $|\mathcal{X}_v| - 1$-dimensional simplex $\Delta^{|\mathcal{X}_v|}$. This gives the parameter space

$$\Theta_{\mathcal{D}} := \prod_{v \in V} \prod_{x_{\mathrm{pa}(v)} \in \mathcal{X}_{\mathrm{pa}(v)}} \left\{ \theta_{v, x_{\mathrm{pa}(v)}} \in \Delta^{|\mathcal{X}_v|} \right\}.$$

The sufficient statistic is the vector $t_v(x_v, x_{\mathrm{pa}(v)})$ of length $|\mathcal{X}_{\mathrm{pa}(v)}| \times |\mathcal{X}_v|$ with entry 1 at the $(x_{\mathrm{pa}(v)}, x_v)$ position and zeros elsewhere, the natural parameter $\eta_v(\theta_v)$ is given by the vector with entry $\log(\theta_{v, x_{\mathrm{pa}(v)}, x_v})$ for every $(x_{\mathrm{pa}(v)}, x_v)$ pair, $b_v = 1$, and as dominating measure μ_v the counting measure.

4.1 Typicality of faithfulness in the parameter space

To obtain that faithful parameters are typical we require that the conditional independence constraints of faithfulness violations are highly restrictive. In Theorem 5 (in particular, Lemma 3) we leveraged that conditional independence is a polynomial constraint in the interpolation parameter, but this approach is not feasible for conditional exponential families since these are not closed under taking mixtures. Instead, we ensure that each marginal density $p_\theta(x_A)$ is an analytic function in the parameter θ, and hence that conditional independence is the zero set of an analytic function, which is atypical in the parameter space.

Theorem 7. *Given DAG G with vertices V, let μ_v be a σ-finite measure for each $v \in V$ and let $\varphi : \Theta \to \mathrm{BN}_G$ be a Bayesian network parametrisation with Θ open and connected, such that for each $A \subseteq V$ the marginal distribution $\mathbb{P}_\theta(X_A)$ has a density $p_\theta(x_A)$ with respect to $\mu_A := \bigotimes_{v\in A} \mu_v$ which is μ_A-a.s. continuous in x_A and analytic in θ, and such that its support does not depend on θ. If there is at least one faithful parameter in Θ, then the set of faithful parameters is open and dense in Θ, and the unfaithful parameters have Lebesgue measure zero.*

Proof. Let $A, B, C \subseteq V$ such that $A \not\perp\!\!\!\perp^d_G B \,|\, C$, and let $\theta_1 \in \Theta$ be a parameter such that $X_A \not\perp\!\!\!\perp_{\mathbb{P}_{\theta_1}} X_B \,|\, X_C$, which exists by assumption. Let x_A, x_B, x_C lie in the support of $\mathbb{P}_{\theta_1}$ such that $p_1(x_A, x_B, x_C)p_1(x_C) - p_1(x_A, x_C)p_1(x_B, x_C) \neq 0$ and define

$$q(\theta) := p_\theta(x_A, x_B, x_C)p_\theta(x_C) - p_\theta(x_A, x_C)p_\theta(x_B, x_C).$$

By assumption the support of $\mathbb{P}_\theta$ does not vary in θ, so x_A, x_B and x_C are in the support of $\mathbb{P}_\theta$ for any $\theta \in \Theta$. By continuity of the densities we have $q(\theta_1) \neq 0$ and $q(\theta) = 0$ for all θ such that $X_A \perp\!\!\!\perp_{\mathbb{P}_\theta} X_B \,|\, X_C$, i.e. $\{\theta \in \Theta : X_A \perp\!\!\!\perp_{\mathbb{P}_\theta} X_B \,|\, X_C\} \subseteq \{\theta \in \Theta : q(\theta) = 0\}$. It follows from the identity theorem[7] that the zero set of a nonconstant real analytic function on an open and connected domain is nowhere dense (if it were dense on an open set, then by continuity the function would be 0 on that open set, and hence zero on the whole domain), and has Lebesgue measure zero (Mityagin, 2020). Hence, $T^{-1}(U_G) = \bigcup_{A \not\perp\!\!\!\perp^d_G B \,|\, C} \{\theta \in \Theta : X_A \perp\!\!\!\perp_{\mathbb{P}_\theta} X_B \,|\, X_C\}$ is nowhere dense and has Lebesgue measure zero. Finally, since the unfaithful parameters are nowhere dense, the set of faithful parameters is dense. Since $\theta \mapsto p_\theta(x_V)$ is continuous, it follows from Scheffé's Theorem that $\theta \mapsto \mathbb{P}_\theta(X_V)$ is continuous with respect to the total variation metric, so by Theorem 4 the set of faithful parameters is open. ∎

Remark 1. The proof of Theorem 7 only requires that for every d-connection $A \not\perp\!\!\!\perp^d_G B \,|\, C$ in the graph there is a parameter $\theta \in \Theta$ such that $X_A \not\perp\!\!\!\perp_{\mathbb{P}_\theta} X_B \,|\, X_C$. Given the required analyticity, it follows from the preceding theorem that this is equivalent to the existence of a parameter that is faithful. For a specific parametrisation, the former condition might be easier to prove than the latter — this strategy has also been employed in the original proofs of Theorems 1 and 2.

The following result shows that for conditional exponential family parametrisations whose joint distribution lies in a regular exponential family, the analyticity property of the marginal $p_\theta(x_A)$ is ensured if the parametrisation has analytic natural parameters η_v. An exponential family is *regular* if the density $p_\theta(x_V)$ (with respect to μ_V) is proportional to $\tilde{b}(x_V)e^{\tilde{\eta}(\theta)^\top \tilde{t}(x_V)}$ for some non-negative function $\tilde{b}$ and vector-valued functions $\tilde{\eta}$ and $\tilde{t}$ such that the parametrisation is *minimal* (i.e., the components of $\tilde{\eta}$ are affinely independent and the components of $\tilde{t}$ are μ-a.s. affinely independent), *full* (the natural parameters satisfy $\tilde{\eta}(\Theta) = \{\gamma : \int \tilde{b}(x_V)e^{\gamma^\top \tilde{t}(x_V)} \mathrm{d}\mu(x_V) < \infty\}$) and $\tilde{\eta}(\Theta)$ is open — see Barndorff-Nielsen, 2014 (page 116).

Theorem 8. *Given DAG G with vertices V, consider a conditional exponential family parametrisation of the Bayesian network such that for each $v \in V$ the set Θ_v is open and connected and the natural parameter $\theta \mapsto \eta_v(\theta)$ is analytic with open image $\eta_v(\Theta_v)$, and the joint density $p_\theta(x_V)$ lies in a regular exponential family such that for every $A \subseteq V$ the marginal density $p_\theta(x_A)$ is μ_A-a.s. continuous. If there is at least one faithful parameter in Θ, then the set of faithful parameters is open and dense, and the unfaithful parameters have Lebesgue measure zero.*

[7] Every real analytic function on a connected open domain in $\mathbb{R}^n$ can be extended to a holomorphic function on a connected open domain in $\mathbb{C}^n$, for which the identity theorem of holomorphic functions applies (Gunning and Rossi, 1965). This identity theorem for functions of real variables is obtained by restricting again to $\mathbb{R}^n$.

Proof. The map $\gamma \mapsto e^{A_v(\gamma, x_{\mathrm{pa}(v)})} = \int_{\mathcal{X}_v} b_v(x_v, x_{\mathrm{pa}(v)}) e^{\eta_v^\top t_v(x_v, x_{\mathrm{pa}(v)})} \mathrm{d}\mu_v(x_v)$ can alternatively be written as the complex Fourier-Laplace transform of the push-forward measure $d\nu := d(t_v)_*(b_v \mu_v)$ given by $\gamma \mapsto \int e^{\gamma^\top y} \mathrm{d}\nu(y)$, defined on $\Theta_v' := \{\gamma + i\zeta : \gamma \in \eta_v(\Theta_v), \zeta \in \mathbb{R}^{k_v}\}$. Since the range of η_v is open, it follows from Barndorff-Nielsen (2014), Theorem 7.2 that this map is holomorphic on Θ_v'.[8] As a composition of analytic functions, $p_v(x_v \,|\, x_{\mathrm{pa}(v)})$ is analytic in θ. As a product of analytic functions, the joint density $p_\theta(x_V) = \tilde{b}(x_V) e^{\tilde{\eta}(\theta)^\top \tilde{t}(x_V) - \tilde{A}(\tilde{\eta}(\theta))}$ is analytic in θ as well. By minimality of the exponential family $p_\theta(x_V)$ there are $x_V^0, x_V^1, ..., x_V^k$ such that all vectors $u_i := \tilde{t}(x_V^i) - \tilde{t}(x_V^0)$ are linearly independent. We can write $U^\top \tilde{\eta}(\theta) = g(\theta)$ with $U := (u_1, ..., u_k)$ invertible and $g_i(\theta) := \log(p_\theta(x_V^i)) - \log(p_\theta(x_V^0)) + C_i$ analytic for some constant C_i, hence $\tilde{\eta}(\theta)$ is analytic as well, as a product of analytic functions. Similar as before, it follows from Barndorff-Nielsen (2014), Theorem 7.2 that for every $I \subseteq V$, the map $\eta \mapsto \int_{\mathcal{X}_I} \tilde{b}(x_V) e^{\eta^\top \tilde{t}(x_V)} \mathrm{d}\mu(x_I)$ is holomorphic on a domain which by regularity of the exponential family contains $\tilde{\eta}(\Theta)$. This implies analyticity of the unnormalised density $\tilde{p}_\theta(x_A) = \int_{\mathcal{X}_{V \setminus A}} \tilde{b}(x_V) e^{\tilde{\eta}(\theta)^\top \tilde{t}(x_V)} \mathrm{d}\mu(x_{V \setminus A})$ and the log partition function $\tilde{A}(\eta) = \log \int_{\mathcal{X}_V} \tilde{b}(x_V) e^{\tilde{\eta}(\theta)^\top \tilde{t}(x_V)} \mathrm{d}\mu(x_V)$, hence also of the density $p_\theta(x_A)$. The final claim follows from Theorem 7. ∎

Remark 2. When Θ is the closure of an open convex set, a statement similar to Theorem 8 holds. Since the boundary of an open convex set is nowhere dense and has Lebesgue measure zero (Lang, 1986), we obtain that if there is a faithful parameter in $\mathrm{int}(\Theta)$, the unfaithful parameters are nowhere dense and have Lebesgue measure zero in $\mathrm{int}(\Theta)$, hence also in Θ.

For conditional exponential family parametrisations, the joint distribution does not automatically lie in a regular exponential family. For example, when $Y|X = x \sim \mathrm{Gamma}(\theta_1 x, 1)$ and $X \sim \mathrm{Exp}(\theta_2)$, the joint density is proportional to $\exp(-\theta_2 x + (\theta_1 x - 1)\log(y) - y - \log(\Gamma(\theta_1 x)))$, and this exponent cannot be written as the product of a function of the parameters and a function of the data, so it's not in an exponential family. Positive examples are the discrete and Gaussian Bayesian networks, or the network specified by $Y|X = x \sim \mathrm{Exp}(\theta_1 x)$ and $X \sim \mathrm{Exp}(\theta_2)$.

Given that Spirtes et al. (1993) and Meek (1995) have proven for every DAG G the existence of faithful parameters in $\Theta_{\mathcal{N}}$ and in the interior of $\Theta_{\mathcal{D}}$, and given that these models induce joint distributions which are minimal exponential families, we obtain Theorems 1 and 2 and their topological analogues as corollaries of Theorem 8 and Remark 2:

Corollary 2. *The set of faithful parameters $\{\theta \in \Theta_{\mathcal{N}} : T(\theta) \in F_G\}$ is open and dense and the set of unfaithful parameters $\{\theta \in \Theta_{\mathcal{N}} : T(\theta) \in U_G\}$ has Lebesgue measure zero.*

Corollary 3. *The set of faithful parameters $\{\theta \in \Theta_{\mathcal{D}} : T(\theta) \in F_G\}$ is open and dense[9] and the set of unfaithful parameters $\{\theta \in \Theta_{\mathcal{D}} : T(\theta) \in U_G\}$ has Lebesgue measure zero.*

4.2 Typicality of faithfulness in the space of observational distributions

Similar to Section 3.1, we also investigate the typicality of faithful observational distributions. This is considered with respect to the set of distributions induced by a given conditional exponential family parametrisation $(\Theta, b, \eta, t, \mu)$, denoted by $M_G^{\exp} := T(\Theta)$. The faithful distributions are denoted by $F_G^{\exp} := T(\Theta) \cap F_G$. For this set of observational distributions, the total variation topology coincides with the weak topology, so in contrast with the result in Section 3.1, we now obtain that faithfulness is open and dense in the weak topology.

Theorem 9. *Consider a conditional exponential family Bayesian network parametrisation satisfying the assumptions of Theorem 8, then the total variation topology and the weak topology coincide on $M_G^{\exp}$. If there is at least one faithful parameter in Θ, then the set of faithful distributions $F_G^{\exp}$ is open and dense in $M_G^{\exp}$, in either topology.*

[8] Formally Barndorff-Nielsen (2014) Theorem 7.2 assumes ν to be a probability measure, but the result nevertheless holds when ν is σ-finite.

[9] That this set is open and dense has also been shown by Lin and Zhang (2020). From Remark 2 we can merely conclude that the unfaithful parameters are nowhere dense, but since by Scheffé's theorem the map T is continuous, the set of unfaithful parameters is *closed* and nowhere dense, hence the faithful parameters are open and dense.

Proof. Let $\tilde{\eta}(\theta)$ denote the natural parameter of the minimal parametrisation of the joint density $p_\theta(x_V)$, and let $\tilde{t}$ denote the sufficient statistic. The μ_V-a.e. continuity of $p_\theta(x_V)$ in x_V implies that $\tilde{t}$ is μ_V-a.e. continuous as well, analogous to why analyticity of $\tilde{\eta}(\theta)$ follows from the analyticity of $p_\theta(x_V)$ due to the minimality of the exponential family in Theorem 8. Hence, by Barndorff-Nielsen (2014), Theorem 8.3, we have that $\tilde{\eta}(\Theta)$ and $M_G^{\exp}$ are homeomorphic, where $M_G^{\exp}$ is equipped with the weak topology. Note that Barndorff-Nielsen actually shows that the convergence $\eta_n \to \eta$ in $\tilde{\eta}(\Theta)$ implies $\mathbb{P}_{\eta_n} \xrightarrow{tv} \mathbb{P}_\eta$, so the weak topology and the total variation topology coincide on $M_G^{\exp}$.

If there is a faithful parameter, then by Theorem 8 the faithful parameters are dense in Θ. From the proof of Theorem 8 we have that the natural parameter $\theta \mapsto \tilde{\eta}(\theta)$ is analytic and hence continuous. The faithful parameters are dense in $\tilde{\eta}(\Theta)$, hence also in $M_G^{\exp}$, that is, $F_G^{\exp}$ is dense in $M_G^{\exp}$. By Theorem 4 conditional independence is closed in $M_G^{\exp}$, and hence $F_G^{\exp}$ is open in $M_G^{\exp}$. ■

5 Nonparametric conditional density models

Due to their parametric nature, exponential families might not be flexible enough for certain statistical applications. In this section, we consider distributions on the separable complete metric spaces $\mathcal{X}_V = \prod_{v\in V} \mathcal{X}_v$ that have uniformly continuous and uniformly bounded (conditional) densities with respect to a given locally finite measure $\mu_V := \bigotimes_{v\in V} \mu_v$. A function $f : \mathcal{X}_V \to \mathbb{R}$ is uniformly continuous if it admits a *modulus of continuity*, i.e. an increasing function $\omega : [0,\infty) \to [0,\infty)$ with $\lim_{t\to 0} \omega(t) = \omega(0) = 0$ such that $|f(x) - f(x')| \leq \omega(d(x,x'))$ for all x, y, where d denotes the metric on $\mathcal{X}_V$. Throughout we exclude moduli of continuity with $\lim_{t\downarrow 0} \omega(t)/t = 0$, which only admit constant functions. Classes of uniformly continuous functions which all admit the same modulus of continuity are called *uniformly equicontinuous*.

Definition 8. Given a DAG G with vertices V, a measure μ_v for each $v \in V$, modulus of continuity ω and $K > 0$, the class of *equicontinuous and bounded Bayesian networks* $\mathrm{BN}_G^{\mathrm{eqb}}$ *induced by the tuple* (K, ω, μ) has for each $v \in V$ a conditional density $(x_v, x_{\mathrm{pa}(v)}) \mapsto p(x_v \,|\, x_{\mathrm{pa}(v)})$ with respect to μ_v which is bounded by K and admits modulus of continuity ω.

Given a tuple (K, ω, μ), let $\mathcal{P}^{\mathrm{eqb}}(\mathcal{X}_V) \subseteq \mathcal{P}(\mathcal{X}_V)$ be the set of probability measures which have a density with respect to μ which are bounded by K and admit the modulus of continuity ω. The following is essentially a reformulation of Boos (1985), Lemma 1, extending the original result to more general sample spaces.

Lemma 6. *The weak topology and total variation topology coincide on* $\mathcal{P}^{\mathrm{eqb}}(\mathcal{X}_V)$*, which is closed in* $\mathcal{P}(\mathcal{X}_V)$*.*[10]

Proof. Let $\mathbb{P}_n \xrightarrow{w} \mathbb{P}$ weakly with $\mathbb{P}_n \in \mathcal{P}^{\mathrm{eqb}}(\mathcal{X}_V)$ for all $n \in \mathbb{N}$. By Ascoli's theorem (Munkres, 2014, Theorem 47.1), the class of uniformly bounded and uniformly equicontinuous densities is relatively compact in the topology of uniform convergence on compacta. In particular, for any subsequence n' there is a further subsequence n'' and a $p^* : \mathcal{X}_V \to [0,\infty)$ such that $p_{n''} \to p^*$ uniformly on compacta. This implies that p^* has modulus of continuity ω and is uniformly bounded by K. By Fatou's Lemma we have $\int p^* \mathrm{d}\mu \leq \lim_{n''} \int p_{n''} \mathrm{d}\mu \leq 1$. Since $\mathbb{P}_{n''} \xrightarrow{w} \mathbb{P}$, the sequence $(\mathbb{P}_{n''})$ is uniformly tight by Prokhorov's theorem (Bogachev, 2007, Theorem 8.6.4). For any $\varepsilon > 0$, let K_ε be a compact set with $\mathbb{P}_{n''}(K_\varepsilon^c) \leq \varepsilon$ for all n'' in the further subsequence. Since μ is locally finite, we have $\mu(K_\varepsilon) < \infty$. Since $p_{n''} \to p^*$ uniformly on K_ε and $\mu(K_\varepsilon) < \infty$ we have $\int p^* d\mu \geq \int_{K_\varepsilon} p^* d\mu = \lim_{n''} \int_{K_\varepsilon} p_{n''} d\mu \geq 1 - \varepsilon$, so $\int p^* d\mu \geq 1$. Hence p^* integrates to 1, and so $\mathbb{P}^* \in \mathcal{P}^{\mathrm{eqb}}(\mathcal{X}_V)$. By Scheffé (1947), convergence of the densities implies weak convergence $\mathbb{P}_{n''} \xrightarrow{w} \mathbb{P}^*$. The weak convergence $\mathbb{P}_n \xrightarrow{w} \mathbb{P}$ also implies convergence of the subsequence $\mathbb{P}_{n''} \xrightarrow{w} \mathbb{P}$, and thus $p = p^*$ μ-a.e. The μ-a.e. convergence $p_{n''} \to p$ implies convergence $p_n \to p$ as well (otherwise there exists a subsequence n' with $|p_{n'} - p| > \varepsilon$ on some set with positive μ-measure, which contradicts the existence of a convergent further subsequence), which implies total variation convergence $\mathbb{P}_n \xrightarrow{tv} \mathbb{P}$, again by Scheffé's Theorem. ■

[10] This also holds if the sample spaces are separable metric spaces and the dominating measure μ_V is Radon.

For a specific class of equicontinuous and bounded Bayesian networks $\mathrm{BN}_G^{\mathrm{eqb}}$, the induced set of observational distributions is given by $M_G^{\mathrm{eqb}} := D(\mathrm{BN}_G^{\mathrm{eqb}})$. For each model in $\mathrm{BN}_G^{\mathrm{eqb}}$, the corresponding joint density $p(x_V)$ in M_G^{eqb} admits the modulus of continuity $\omega' := |V|K\omega$ and bound $K' := 1 \vee K^{|V|}$, so M_G^{eqb} (with parameters (K, ω, μ)) is contained in $\mathcal{P}^{\mathrm{eqb}}(\mathcal{X}_V)$ (with parameters (K', ω', μ)). Hence, the weak topology and total variation topology also coincide on M_G^{eqb}.

5.1 Typicality of faithfulness in the space of Bayesian networks

Similar as in Section 3.2, we prove the typicality of the faithful Bayesian networks. First, we obtain the result that if there is at least one faithful Bayesian network in $\mathrm{BN}_G^{\mathrm{eqb}}$, then faithfulness is typical with respect to the metric d_{TV}°, which is closely related to the weak topology for this model class. For each $v \in V$ and $x_{\mathrm{pa}(v)}$ the conditional $\mathbb{P}(X_v \,|\, x_{\mathrm{pa}(v)})$ lies in $\mathcal{P}^{\mathrm{eqb}}(\mathcal{X}_v)$ (with parameters (K, ω, μ_v)). By Lemma 6, the total variation and weak topologies coincide on each of these kernel spaces, so convergence in d_{TV}° corresponds to weak convergence of each Markov kernel uniformly over its conditioning variable.

Theorem 10. *Given a DAG G and class $\mathrm{BN}_G^{\mathrm{eqb}}$ induced by a tuple (K, ω, μ), if there is at least one faithful model in $\mathrm{BN}_G^{\mathrm{eqb}}$, then the set of faithful Bayesian networks is open and dense in $(\mathrm{BN}_G^{\mathrm{eqb}}, d_{TV}^{\circ})$.*

Proof. It immediately follows from Theorem 4 and Lemma 5 that the set of faithful Bayesian networks in $\mathrm{BN}_G^{\mathrm{eqb}}$ is open with respect to d_{TV}°. One readily verifies that $\mathrm{BN}_G^{\mathrm{eqb}}$ is closed under taking mixtures, so similar as in Definition 4, we can take between any unfaithful model m_0 and faithful model m_1 the mixture $m_\lambda := (1-\lambda)m_0 + \lambda m_1$ for which we have $m_\lambda \in \mathrm{BN}_G^{\mathrm{eqb}}$. By applying Lemma 3 (with μ_V as dominating measure instead of $\mathbb{Q} = \mathbb{P}_0 + \mathbb{P}_1$ to ensure that p_0, p_1 and hence p_λ are uniformly equicontinuous and uniformly bounded) to each d-connection in G we have that the tail of the sequence $m_{1/n}$ is faithful, and by a similar derivation as in Theorem 6 it converges to m_0, hence the faithful Bayesian networks are dense. ■

For a specific model class, the question remains whether it contains a faithful Bayesian network. For real sample spaces and densities with respect to Lebesgue measure, this is confirmed by leveraging Corollary 2.

Lemma 7. *Given a DAG G, let $\mathcal{X}_V = \mathbb{R}^{|V|}$, and let $\mathrm{BN}_G^{\mathrm{eqb}}$ be induced by a tuple (K, ω, μ) with μ the Lebesgue measure. There is a faithful model $m \in \mathrm{BN}_G^{\mathrm{eqb}}$.*

Proof. By Corollary 2, the faithful parameters are dense in $\Theta_{\mathcal{N}}$. As function of $(x_v, x_{\mathrm{pa}(v)})$, the conditional density $p_\theta(x_v \,|\, x_{\mathrm{pa}(v)})$ has supremum $S(\theta) := 1/\sqrt{2\pi\sigma_v^2}$ and Lipschitz constant $L(\theta) := \sqrt{1 + \|\beta_v\|_2^2}/(\sqrt{2\pi e}\sigma_v^2)$, which are both continuous functions onto $(0, \infty)$, and hence the images of the faithful parameters under these two mappings are dense in $(0, \infty)$.

Let $A := \lim_{t \downarrow 0} \omega(t)/t$, and let δ be the smallest strictly positive number such that $\omega(\delta) = \delta A/2$, and set $\delta = \infty$ if this condition is not met. Since ω is non-decreasing, any $A/2$-Lipschitz conditional density bounded by $K \wedge \omega(\delta)$ admits the modulus ω and is bounded by K. Since the images of the mappings L and S are dense in $(0, \infty)$, there is a faithful parameter that satisfies $L(\theta) \leq A/2$ and $S(\theta) \leq K \wedge \omega(\delta)$. ■

Corollary 4. *Under the assumptions of Lemma 7, the set of faithful Bayesian networks is open and dense in $(\mathrm{BN}_G^{\mathrm{eqb}}, d_{TV}^{\circ})$.*

5.2 Typicality of faithfulness in the space of observational distributions

Finally, we also show that the observational distributions induced by the model class $\mathrm{BN}_G^{\mathrm{eqb}}$ are typically faithful. To that end, recall the definition of the set of observational distributions M_G^{eqb} of $\mathrm{BN}_G^{\mathrm{eqb}}$ induced by a triple (K, ω, μ), and define the faithful distributions as $F_G^{\mathrm{eqb}} := M_G^{\mathrm{eqb}} \cap F_G$.

Theorem 11. *Given a DAG G and class $\mathrm{BN}_G^{\mathrm{eqb}}$ induced by a tuple (K, ω, μ), if there is at least one faithful model in $\mathrm{BN}_G^{\mathrm{eqb}}$, then the set of faithful distributions F_G^{eqb} is open and dense in M_G^{eqb}, both in the weak topology and in the total variation metric.*

Proof. If there is a faithful model in $\mathrm{BN}_G^{\mathrm{eqb}}$, then the set of faithful models is dense in $(\mathrm{BN}_G^{\mathrm{eqb}}, d_{TV}^{\circ})$. By Lemma 5 the distribution map $D : (\mathrm{BN}_G^{\mathrm{eqb}}, d_{TV}^{\circ}) \to (M_G^{\mathrm{eqb}}, d_{TV})$ is continuous, so then the set F_G^{eqb} is dense in M_G^{eqb}. It is closed by Theorem 4. By Lemma 6 the weak topology and the total variation topology coincide on M_G^{eqb}. ■

6 Bayesian networks with latent variables

The assumption that all variables in the Bayesian network must be observed is often too restrictive in practice. When certain variables remain unobserved, a suitable modelling class is that of Bayesian networks with observed variables V and latent variables W.

Given a DAG G over $V \cup W$, the *latent projection* of G onto V is the *Acyclic Directed Mixed Graph* (ADMG) G_V with vertices V, directed edges $a \to b$ if there is a path $a \to w_1 \to ... \to w_n \to b$ in G with $w_i \in W$ for all $i = 1, ..., n$ (if any), and bi-directed edges $a \leftrightarrow b$ if there is a bifurcation $a \leftarrow w_1 \leftarrow ... \leftarrow w_k \to ... \to w_n \to b$ in G with $w_i \in W$ for all $i = 1, ..., n$ (Verma, 1993). An example of a DAG G and its latent projection G_V is given in Figure 4.

(a) DAG G (b) Latent projection G_V

Figure 4: DAG G and latent projection G_V onto $V := \{A, B, C\}$.

The definition of d-separation for ADMGs (also known as *m-separation* (Richardson, 2003)) employs an extended notion of a collider: given ADMG G_V with path $\pi = a \ast\!\!\ast ... \ast\!\!\ast b$, a *collider* is a vertex v with $\to v \leftarrow$, $\leftrightarrow v \leftarrow$, $\to v \leftrightarrow$ or $\leftrightarrow v \leftrightarrow$ in π. As for DAGs, sets of vertices A and B are *d-separated* given C in ADMG G, written $A \perp_G^d B \,|\, C$, if for every path $\pi = a \ast\!\!\ast ... \ast\!\!\ast b$ between $a \in A$ and $b \in B$, there is a collider in π that is not an ancestor of C, or there is a non-collider in π in C. The independence models of G and G_V with respect to V are equal: for any $A, B, C \subseteq V$ we have $A \perp_G^d B \,|\, C$ if and only if $A \perp_{G_V}^d B \,|\, C$ (Verma, 1993). As a corollary the Markov property (1) also holds for the latent projection G_V of Bayesian networks with latent variables.

The question that we consider is whether (parameters of) Bayesian networks with latent variables are typically faithful to their latent projection. Write U_{G_V}, F_{G_V} for the distributions over $\mathcal{X}_{V \cup W}$ that are unfaithful and faithful with respect to the ADMG G_V respectively. The core observation for extending results of Sections 3, 4 and 5 from DAGs to ADMGs is the following:

Lemma 8. *Given DAG G with vertices $V \cup W$ and its latent projection G_V onto V, any distribution over $V \cup W$ that is unfaithful with respect to G_V is also unfaithful with respect to G.*

Proof. The latent projection preserves d-separations, so the result follows immediately from the expression for the set of unfaithful distributions:

$$\bigcup_{A \not\perp_G^d B \,|\, C} \{\mathbb{P} \in M_G : X_A \perp\!\!\!\perp_{\mathbb{P}} X_B \,|\, X_C\}.$$

For U_{G_V} the union ranges over subsets $A, B, C \subseteq V$ and for U_G the union ranges over $A, B, C \subseteq V \cup W$ (with a d-connection), hence we get $U_{G_V} \subseteq U_G$. ■

Now, the preceding results are straightforwardly extended to conclude for unconstrained Bayesian networks with latent variables, conditional exponential family Bayesian networks with latent variables, and equicontinuous and bounded Bayesian networks with latent variables that faithfulness (which is only required to hold with respect to the latent projection) is open and dense. The extensions of the results for typicality in the space of observational distributions (Theorems 5, 9 and 11) immediately follow from Lemma 8. Considering the distribution map $D : \mathrm{BN}_G \to M_G$, the inclusion $D^{-1}(U_{G_V}) \subseteq D^{-1}(U_G)$

(which follows from Lemma 8) implies that $D^{-1}(F_{G_V}) \supseteq D^{-1}(F_G)$ is dense. Openness follows from the same argument as in the original theorems: U_{G_V} is a finite union of conditional independence sets which are closed in the total variation topology by Theorem 4, and D is continuous. This gives the extensions of the results for the Bayesian networks themselves (Theorems 6 and 10 and Corollary 4). The extensions of the results for Euclidean parameters (Theorem 8 and Corollaries 2 and 3) follow by the same reasoning applied to $T = D \circ \varphi$.

7 Discussion

In this work, we have established several results concerning the typicality of the faithfulness property, showing in various settings that Bayesian networks are indeed typically faithful in both a topological and a measure-theoretic sense. We have shown:

- For unconstrained nonparametric Bayesian networks:
  - the faithful distributions are open and dense with respect to the total variation metric (Thm. 5);
  - the faithful Bayesian networks are open and dense with respect to the metric d_{TV}° (Thm. 6).
- For sufficiently regular conditional exponential family Bayesian networks:
  - if there exists a faithful parameter, then the faithful parameters are open and dense with respect to the Euclidean topology and the unfaithful parameters have Lebesgue measure zero (Thm. 8);
  - if there exists a faithful parameter, then the faithful distributions are open and dense with respect to the weak topology and the total variation metric (Thm. 9);
  - for linear Gaussian and discrete Bayesian networks, there is a faithful parameter, so faithfulness is typical in these ways (Cor. 2 and 3).
- For equicontinuous and bounded Bayesian networks:
  - if there exists a faithful model, then the faithful models are open and dense with respect to the metric d_{TV}° (Thm. 10);
  - if there exists a faithful model, then the faithful distributions are open and dense in the weak topology and the total variation metric (Thm. 11);
  - for real sample spaces and densities with respect to Lebesgue measure, there exists a faithful parameter, so faithful Bayesian networks c.q. distributions are open and dense (Cor. 4).

This collection of results naturally raises the question: what is the relative value of these results, with their different notions of typicality? The two main notions that we use — topological and measure-theoretic — do not necessarily coincide. For example, the Smith-Volterra-Cantor set is a nowhere dense subset of $[0, 1]$ that has Lebesgue measure $1/2$. In general, *every* subset of $\mathbb{R}$ is the disjoint union of a meager set and a Lebesgue null set (Oxtoby, 1980, Theorem 1.6): a set that is small in one sense may be large in the other sense.

The measure-theoretic approach, which identifies 'atypical' sets with those of (Lebesgue) measure zero, is for example useful when one samples a Bayesian network, e.g. for performing simulations. If the unfaithful parameters have measure zero, the probability of encountering one when sampling from a distribution with a density is zero. Note that this σ-ideal depends on the choice of σ-algebra and measure. A restriction is that this approach is challenging in the nonparametric case, as no canonical analogue of Lebesgue measure exists in infinite-dimensional spaces. For infinite-dimensional locally compact topological groups the Haar measure is a natural choice, but the space of (observational distributions of) Bayesian networks is generally not locally compact with respect to the weak or total variation topology. Hunt et al. (1992) introduce the notion of *shy sets* as analogue of Lebesgue-null sets

in arbitrary linear metric spaces. The space of probability distributions is not linear so this concept is not applicable in our case.

The topological approach identifies 'typical' sets with those that are either open and dense, or complements of nowhere dense sets, or complements of meager sets. Clearly, such notions of typicality depend on the choice of the topology. The total variation topology is convenient, since here, conditional independence is a closed property. The weak topology is a natural choice as it is closely related to testability (Dembo and Peres, 1994; Genin and Kelly, 2017; Boeken et al., 2026). However, obtaining results in the weak topology requires finding regularity conditions such that conditional independence is closed, which does not hold in general. Finding such regularity conditions is challenging. One approach is to find regularity conditions such that the weak topology and total variation topology coincide. To this end, the employed conditions by Barndorff-Nielsen (2014) and Boos (1985) for exponential families and equicontinuous and bounded densities are convenient. Boeken et al. (2026) deviate from this approach of equating the total variation topology and weak topology, as they provide alternative uniform continuity conditions on the Markov kernels $x_B \mapsto \mathbb{P}(X_A \,|\, x_B)$, for which they directly show that conditional independence is closed in the weak topology. However, these regularity conditions fail to combine nicely with the interpolation of Bayesian networks that we consider in our proof technique for showing that faithful Bayesian networks are dense, hence we do not follow that approach in the current work. We have introduced the metric d_{TV}° on the space of Bayesian networks, which measures the worst-case distance between corresponding Markov kernels. This is a natural choice when viewing Bayesian networks as causal models, where the Markov kernels represent mechanisms that are defined for all parent values. Alternative metrics could be considered, for instance by replacing the supremum over parent values with an average weighted by some reference measure, which would yield a weaker topology and potentially stronger typicality results. However, such a metric would depend on the choice of reference measure, and would not distinguish between Bayesian networks that differ only on parent values with zero reference measure — an undesirable property for causal models.

7.1 Implications for constraint-based causal discovery

The topological properties of conditional independence have two important implications for constraint-based causal discovery:

- when conditional independence is closed in the weak topology there exists a consistent test, which can be used to make any sound constraint-based causal discovery algorithm consistent, and
- the set of faithful Bayesian networks is open and dense, so any causal discovery algorithm that is consistent under the faithfulness assumption is has a topologically large 'domain of consistency'.

For the first statement, consider any given set of probability measures $\mathcal{P}$ on $\mathcal{X}_A \times \mathcal{X}_B \times \mathcal{X}_C$, and let $H_0 = \{\mathbb{P} \in \mathcal{P} : X_A \perp\!\!\!\perp_{\mathbb{P}} X_B \,|\, X_C\}$ and $H_1 = \{\mathbb{P} \in \mathcal{P} : X_A \not\perp\!\!\!\perp_{\mathbb{P}} X_B \,|\, X_C\}$. Under various regularity conditions on $\mathcal{P}$, Dembo and Peres (1994); Ermakov (2017); Genin and Kelly (2017) and Boeken et al. (2026) give topological characterisations for the (uniformly) consistent testability of any pair of statistical hypotheses. Importantly, Genin and Kelly (2017) (Theorem 4.1) show for distributions that have a density with respect to a common dominating measure, that if a null-hypothesis H_0 is closed in the weak topology, then there exists a consistent test, i.e. for every $n \in \mathbb{N}$ there exists a map $\varphi_n : \mathcal{X}_V^n \to \{0, 1\}$ such that $\lim_{n\to\infty} \mathbb{P}^n(\varphi_n = i) = 1$ if and only if $\mathbb{P} \in H_i$.[11] For classes of distributions where the weak topology and the total variation topology coincide, it follows from Theorem 4 that conditional independence is consistently testable.

Since causal discovery algorithms are agnostic of the underlying causal graph, we need to consider the space of Bayesian networks and induced observational distributions of Bayesian networks with all

[11] Actually, Genin and Kelly (2017) show more: that this test is consistent and has Type-1 error control. We ignore this property because it is unclear how Type-1 error control translates to error control for causal discovery algorithms. This suggests that we are restricting the model classes more than necessary. Indeed, the existence of a merely consistent test (without requiring Type-1 error control) is equivalent to the null and alternative hypotheses being F_σ in the weak topology (Genin and Kelly, 2017, Theorem 4.3). This can be shown to hold for conditional (in)dependence in countable unions of the classes considered in Theorem 12, and for which analogues of Theorem 13 and Corollary 5 hold with *comeager* domains of consistency.

possible causal graphs over a fixed set of variables. To formalise this, let V be a finite index set, let $\mathcal{X}_V = \prod_{v \in V} \mathcal{X}_v$ be a product of separable complete metric spaces, let μ be a dominating measure μ on $\mathcal{X}_V$, and finally let $\mathrm{BN}^{\mathrm{eqb}} := \bigcup_G \{G\} \times \mathrm{BN}_G^{\mathrm{eqb}}$ be the set of all Bayesian networks, where the union is taken over all DAGs G with vertices V. Similar to Lin and Zhang (2020), equip this space with the metric $\delta + d_{TV}^{\circ}$, where δ is the discrete metric on the space of DAGs with vertices V. This way we can extend the result about consistent testability without relying on a specific graph G.

Theorem 12. *Conditional independence is consistently testable in the set of observational distributions induced by* $\mathrm{BN}^{\mathrm{eqb}}$.

Since the weak topology and total variation topology coincide for the conditional exponential family Bayesian networks of Theorem 9, a similar result can be derived for this model class.

For the second statement, it follows from Theorem 10 that the set of faithful Bayesian networks is open and dense in $\mathrm{BN}^{\mathrm{eqb}}$. Hence, any constraint-based causal discovery algorithm that is sound under the assumption of faithfulness (meaning that it arrives at the correct conclusion when using a conditional independence oracle) — like the *PC-* and *FCI-algorithms* (Spirtes et al., 1993) — is consistent on a topologically large set among all possible Bayesian networks.

Theorem 13. *Given a set of vertices V, if for each DAG G with vertices V there exists a faithful Bayesian network in $\mathrm{BN}_G^{\mathrm{eqb}}$, then every causal discovery algorithm that is sound under the assumption of faithfulness is consistent on an open and dense domain in* $\mathrm{BN}^{\mathrm{eqb}}$.

Obtaining these results for arbitrary Bayesian networks without any regularity assumptions on the distributions is impossible. Although conditional independence is closed with respect to the total variation metric, this does not imply that it is consistently testable. Indeed, Shah and Peters (2020); Neykov et al. (2021) and Boeken et al. (2026) prove that without imposing regularity conditions, conditional independence is not consistently testable.

Faithfulness is sufficient, but not *necessary* for consistent constraint-based causal discovery. Weaker conditions have been proposed, like adjacency faithfulness (Spirtes et al., 1993), P-minimality (Pearl, 2009) and SGS-minimality (Spirtes et al., 1993). Zhang (2013) has shown that SGS-minimality is strictly weaker than P-minimality. P-minimality is sufficient for constraint-based causal discovery (Lin and Zhang, 2020), and SGS-minimality is necessary. Since these sets of Bayesian networks include all faithful Bayesian networks, they are the complement of a nowhere dense set if the faithful Bayesian networks are open and dense. In other words, the typicality of faithful Bayesian networks implies the typicality of even larger classes of Bayesian networks, like the P-minimal Bayesian networks.

Lin and Zhang (2020) show that there exist constraint-based causal discovery algorithms that are consistent on a *maximal* domain. In particular, they show that there exists a causal discovery algorithm $\hat{H}$ that is consistent for all P-minimal Bayesian networks, and for any other algorithm that is consistent for some Bayesian network $(G, \mathbb{P})$, $\hat{H}$ is also consistent for $(G, \mathbb{P})$, so the domain of consistency is maximal. Combined with Theorem 12 this gives the following result.

Corollary 5. *Given a set of vertices V, if for each DAG G with vertices V there exists a faithful Bayesian network in $\mathrm{BN}_G^{\mathrm{eqb}}$, then there exists a causal discovery algorithm that is consistent on a nowhere dense, maximal domain in* $\mathrm{BN}^{\mathrm{eqb}}$.

Faithful distributions can possess extremely weak dependencies that are hard to test for. For linear Gaussian networks, Zhang and Spirtes (2002) consider *strong faithfulness*, i.e. the condition that every d-connection in the graph has in the distribution a corresponding conditional dependence (partial correlation in their case) with some minimal strength. The set of parameters which exhibit a weak dependency is known to be of strict positive measure (Uhler et al., 2013), so strong faithfulness is not typical in the measure-theoretic sense. That every conditional dependence has a minimal strength implies the existence of a uniformly consistent test, and hence of uniformly consistent causal discovery algorithms. Namely, under tightness conditions, Boeken et al. (2026) show that uniformly consistent testability is implied by separation of the hypotheses in the *bounded Lipschitz metric* d_{BL} for the weak topology (i.e. the condition that $d_{BL}(H_0, H_1) > 0$) and that the nonparametric version of the minimal strength of the conditional dependence (i.e. that for some $\varepsilon > 0$ the alternative hypothesis H_1

consists of measures with $d_{BL}(\mathbb{P}(X \mid Z) \otimes \mathbb{P}(Y, Z), \mathbb{P}(X, Y, Z)) > \varepsilon$) satisfies this separation property under certain regularity conditions. We conjecture that under these regularity assumptions and this generalised notion of strong faithfulness, uniformly consistent constraint-based causal discovery can be achieved.

7.2 Concluding remarks

A number of follow-up questions remain. First, although there is no canonical measure on the space of Bayesian networks, one might construct specific ones. For example, conditional optional Pólya trees (Ma, 2017) provide a flexible class of random conditional measures. These can straightforwardly be extended to random Bayesian networks with a given graph. It would be of interest to know whether faithfulness has full measure.

Sadeghi (2017) characterises faithfulness in terms of the properties intersectionality, compositionality, singleton-transitivity, and ordered downward- and upward-stability. Our various typicality results for faithfulness directly imply corresponding typicality results for those constituent properties. Sadeghi equates very similar properties to faithfulness of distributions to more general classes of graphs (e.g. chain graphs and ancestral graphs). It would be interesting to see whether this approach leads to typicality results for faithfulness of distributions with respect to these more general classes of graphs.

Finally, our work focusses on acyclic causal models. However, constraint-based causal discovery algorithms also exist for certain classes of uniquely solvable cyclic models, like *simple SCMs* (Bongers et al., 2021). These causal discovery algorithms either rely on versions of faithfulness defined in terms of d-separations or σ-separation, see e.g. Richardson (1996), Strobl (2019) and Mooij and Claassen (2020). Our proof techniques for acyclic models do not immediately transfer to cyclic models. For example, it is not clear whether the interpolation of two simple SCMs is again a simple SCM. It therefore remains an open question whether simple SCMs are typically faithful.

8 Acknowledgements

This research was supported by Booking.com. We thank Konstantin Genin for helpful remarks.

References

Aliprantis, C. D. and Border, K. C. (2006). *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, Berlin ; New York, 3rd [rev. and enl.] ed edition.

Bareinboim, E., Correa, J. D., Ibeling, D., and Icard, T. (2022). On Pearl's Hierarchy and the Foundations of Causal Inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, volume 36, pages 507–556. Association for Computing Machinery, New York, NY, USA.

Barndorff-Nielsen, O. E. (2014). *Information and Exponential Families: In Statistical Theory*. John Wiley & Sons, Chichester [U.K.] New York.

Boeken, P., Skapinakis, E., Genin, K., and Mooij, J. M. (2026). Topological Criteria for Hypothesis Testing with Finite-Precision Measurements.

Bogachev, V. I. (2007). *Measure Theory Vol. II*. Springer, Berlin ; New York.

Bongers, S., Forré, P., Peters, J., and Mooij, J. M. (2021). Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915.

Boos, D. D. (1985). A Converse to Scheffe's Theorem. *The Annals of Statistics*, 13(1).

Dembo, A. and Peres, Y. (1994). A Topological Criterion for Hypothesis Testing. *The Annals of Statistics*, 22(1).

Ermakov, M. (2017). On Consistent Hypothesis Testing. *Journal of Mathematical Sciences*, 225(5):751–769.

Feigin, P. D. (1981). Conditional Exponential Families and a Representation Theorem for Asympotic Inference. *The Annals of Statistics*, 9(3):597–603.

Geiger, D., Verma, T., and Pearl, J. (1990). Identifying independence in Bayesian networks. *Networks*, 20(5):507–534.

Genin, K. and Kelly, K. T. (2017). The Topology of Statistical Verifiability. *Electronic Proceedings in Theoretical Computer Science*, 251:236–250.

Gunning, R. C. and Rossi, H. (1965). *Analytic Functions of Several Complex Variables*. Prentice-Hall.

Hunt, B. R., Sauer, T., and Yorke, J. A. (1992). Prevalence: A translation-invariant "almost every" on infinite-dimensional spaces. *Bulletin of the American Mathematical Society*, 27(2):217–238.

Ibeling, D. and Icard, T. (2021). A Topological Perspective on Causal Inference. In *Advances in Neural Information Processing Systems*, volume 34, pages 5608–5619. Curran Associates, Inc.

Kechris, A. (1995). *Classical descriptive set theory*. Springer.

Lang, R. (1986). A note on the measurability of convex sets. *Archiv der Mathematik*, 47(1):90–92.

Lauritzen, S. (1996). *Graphical models*. Clarendon Press.

Lauritzen, S. (2024). Total variation convergence preserves conditional independence. *Statistics & Probability Letters*, 214:110200.

Lin, H. and Zhang, J. (2020). On Learning Causal Structures from Non-Experimental Data without Any Faithfulness Assumption. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, pages 554–582. PMLR.

Ma, L. (2017). Recursive partitioning and multi-scale modeling on conditional densities. *Electronic Journal of Statistics*, 11(1).

Meek, C. (1995). Strong completeness and faithfulness in Bayesian networks. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, page 411–418, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Meek, C. (1998). *Graphical Models: Selecting Causal and Statistical Models*. PhD thesis, Carnegie Mellon University.

Mityagin, B. S. (2020). The Zero Set of a Real Analytic Function. *Mathematical Notes*, 107(3):529–530.

Mooij, J. and Claassen, T. (2020). Constraint-Based Causal Discovery using Partial Ancestral Graphs in the presence of Cycles. In *UAI2020*, pages 1159–1168. PMLR.

Munkres, J. R. (2014). *Topology*. Pearson, Harlow, 2. ed., pearson new internat. ed edition.

Neykov, M., Balakrishnan, S., and Wasserman, L. (2021). Minimax optimal conditional independence testing. *The Annals of Statistics*, 49(4):2151–2177.

Oxtoby, J. C. (1980). *Measure and Category*. Graduate Texts in Mathematics. Springer New York, New York, NY, Second edition.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. The Morgan Kaufmann Series in Representation and Reasoning. Morgan Kaufmann, San Francisco, Calif, rev. 2. ed., transferred to digital printing edition.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Richardson, T. (1996). A discovery algorithm for directed cyclic graphs. In *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence*, UAI'96, pages 454–461, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Richardson, T. (2003). Markov Properties for Acyclic Directed Mixed Graphs. *Scandinavian Journal of Statistics*, 30(1):145–157.

Sadeghi, K. (2017). Faithfulness of Probability Distributions and Graphs. *Journal of Machine Learning Research*, 18(148):1–29.

Scheffé, H. (1947). A Useful Convergence Theorem for Probability Distributions. *The Annals of Mathematical Statistics*, 18(3):434–438.

Shah, R. D. and Peters, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538.

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction, and Search*, volume 81 of *Lecture Notes in Statistics*. Springer, New York, NY.

Strobl, E. V. (2019). A constraint-based algorithm for causal discovery with cycles, latent variables and selection bias. *International Journal of Data Science and Analytics*, 8(1):33–56.

Uhler, C., Raskutti, G., Bühlmann, P., and Yu, B. (2013). Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, 41(2):436–463.

Verma, T. (1993). Graphical aspects of causal models. UCLA Cognitive Systems Laboratory, Technical Report (R-191).

Verma, T. and Pearl, J. (1990). Causal Networks: Semantics and Expressiveness. In Shachter, R. D., Levitt, T. S., Kanal, L. N., and Lemmer, J. F., editors, *Machine Intelligence and Pattern Recognition*, volume 9 of *Uncertainty in Artificial Intelligence*, pages 69–76. North-Holland.

Yang, E., Baker, Y., Ravikumar, P., Allen, G., and Liu, Z. (2014). Mixed Graphical Models via Exponential Families. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pages 1042–1050. PMLR.

Zhang, J. (2013). A Comparison of Three Occam's Razors for Markovian Causal Models. *The British Journal for the Philosophy of Science*, 64(2):423–448.

Zhang, J. and Spirtes, P. (2002). Strong faithfulness and uniform consistency in causal inference. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, UAI'03, pages 632–639, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Zhang, J. and Spirtes, P. (2008). Detection of Unfaithfulness and Robust Causal Inference. *Minds and Machines*, 18(2):239–271.