Are Bayesian networks typically faithful?

Philip Boeken Patrick Forré Joris M. Mooij
University of Amsterdam

January 22, 2025

arXiv:2410.16004v2 [math.ST] 20 Jan 2025

Abstract

Faithfulness is a ubiquitous 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. However, this does not imply that faithfulness is typical in restricted classes of Bayesian networks, as 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 mild regularity conditions, the faithful parameters constitute a dense, open set and the unfaithful parameters have Lebesgue measure zero, extending the existing results for linear Gaussian and discrete Bayesian networks. Finally, we show that the aforementioned results also hold for Bayesian networks with latent variables.

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 $X_A \perp\!\!\!\perp_{\mathbb{P}} X_B \mid X_C$ for multiple 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

$$A \perp^d_G B \mid X \iff X_A \perp\!\!\!\perp_{\mathbb{P}} X_B \mid X_C.$$

Bayesian networks for which this condition holds are called *faithful.* The implication from left to right holds for all Bayesian networks, and is called the *Markov property.* 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 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 $d \in \mathbb{N}$) and show that drawing the parameters 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 set of parameters whose distribution is unfaithful to* G *is measure-zero.*

Theorem 2 (Meek, 1995)**.** *With respect to Lebesgue measure over* $\Theta_{\mathcal{D}}$*, the set of parameters 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 main nonparametric result 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 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.

In practice, one often imposes parametric assumptions on the data to facilitate statistical inference. To this end, we consider the class of Bayesian networks parametrised by conditional exponential families. Under mild regularity conditions, we obtain the following generalisation of Theorems 1 and 2:

> **Considering a conditional exponential family parametrisation over a given DAG, the faithful parameters constitute a dense, open set, and the set of unfaithful parameters has Lebesgue measure zero.**

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 $\emptyset$, which are closed under taking subsets and countable unions. The family of Lebesgue measure 0 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 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, *open* sets are nowhere dense. Comeager sets (complements of meager sets) are commonly referred to as *typical* (Kechris, 1995). We show 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 nowhere denseness of unfaithful parameters of discrete Bayesian network to relieve consistency requirements of causal discovery methods for discrete variables.

The outline of this paper is as follows. In Section 2 we provide some technical prerequisites about Bayesian networks and the total variation metric. In Section 3 we state and prove our main nonparametric result: that faithful distributions are dense and open. In Section 4 we lift this result from the space of distributions to the space of Bayesian networks. We focus in particular on finite dimensional parametrisations of Bayesian networks, and we specifically prove the topological analogue of the measure-zero results of Spirtes et al. and Meek for linear Gaussian and discrete Bayesian networks. In Section 5 we extend our results to Bayesian networks with latent variables.

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* $\mathbb{P}(X_B \,|\, X_A)$ is a measurable map $\mathcal{X}_A \to \mathcal{P}(\mathcal{X}_B)$, where $\mathcal{P}(\mathcal{X}_B)$ is equipped with the smallest

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

σ-algebra that makes for all $D \in \mathcal{B}(\mathcal{X}_B)$ the evaluation map $\mathrm{ev}_D : \mathcal{P}(\mathcal{X}_B) \to [0,1], \mathbb{P} \mapsto \mathbb{P}(X_B \in D)$ 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 $\mathbb{P}(X_A, X_B) = \mathbb{P}(X_B \,|\, X_A) \otimes \mathbb{P}(X_A)$. 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 $\mathbb{P}(X_A, X_B \,|\, X_C) = \mathbb{P}(X_A \,|\, X_C) \otimes \mathbb{P}(X_B \,|\, X_C)$ holds $\mathbb{P}(X_C)$ almost surely. This is equivalent to independence of the generated σ-algebras $\sigma(X_A) \perp\!\!\!\perp_{\mathbb{P}} \sigma(X_B) \,|\, \sigma(X_C)$ or, if $\mathbb{P}(X_A, X_B, X_C)$ has a density $p(x_A, x_B, x_C)$, to $p(x_A, x_B \,|\, x_C) = p(x_A \,|\, x_C) p(x_B \,|\, x_C)$ for all x_A, x_B, x_C with $p(x_C) > 0$.

Writing $\mathrm{pa}(v)$ for the set of parents of v in G, a *Bayesian network over* G is 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 $\pi = a \ast\!\!-\!\!\ast \dots \ast\!\!-\!\!\ast b$, a *collider* is a vertex v with $\dots \to v \leftarrow \dots$ in π. 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 $\pi = a \ast\!\!-\!\!\ast \dots \ast\!\!-\!\!\ast b$ between every $a \in A$ and $b \in B$, there is a collider on π that is not an ancestor of C, or if 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 $A \not\perp^d_G B \,|\, C$.

Theorem 3 (Verma and Pearl, 1990)**.** *For any Bayesian network over DAG G with observational distribution $\mathbb{P}$ the* global Markov property *holds:*

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

for all $A, B, C \subseteq V$.

In general, the set of conditional independencies in $\mathbb{P}$ 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 2. A Bayesian network is called *faithful* if for all $A, B, C \subseteq V$ we have

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

Example 1. The following Bayesian networks are unfaithful. Corresponding graphs 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 $X_A \perp\!\!\!\perp X_C$.[2]

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 $X_A \perp\!\!\!\perp X_C$.

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 $A \not\perp^d_{G^c} C \,|\, B$ and $X_A \perp\!\!\!\perp X_C \,|\, X_B$.[3]

An important step in our proof of the typicality of faithful distributions, is that conditional independence is preserved when 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

[2] A realistic example of this phenomenon is when opening a window (A) and subsequently turning up the heating (B) has no net effect on room temperature (C).

[3] For Bayesian networks with known deterministic variables or relations, Geiger et al. (1990) introduced the *D-separation* criterion.

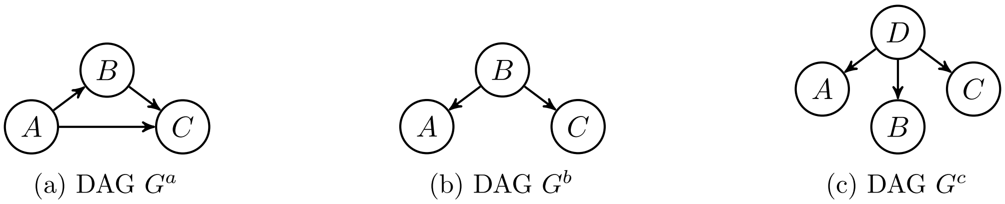


Figure 1: DAGs of the Bayesian networks that are given in Example 1.

probability measures $\mathbb{P}, \mathbb{P}_1, \mathbb{P}_2, \ldots \in \mathcal{P}(\mathcal{X}_V)$ 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 $f : \mathcal{X}_V \to [-1,1]$. 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 $X_A \not\!\perp\!\!\!\perp_{\mathbb{P}} X_B \,|\, X_C$; see e.g. Lauritzen (1996), pp. 38-39. Instead of weak convergence, we consider a different topology:

Definition 3. The *total variation metric* d_{TV} on $\mathcal{P}(\mathcal{X}_V)$ is defined as

$$d_{TV}(\mathbb{P}, \mathbb{Q}) := \sup_{A \in \mathcal{B}(\mathcal{X}_V)} |\mathbb{P}(A) - \mathbb{Q}(A)|.$$

Convergence in this metric is denoted by $\mathbb{P}_n \xrightarrow{tv} \mathbb{P}$. It is equivalent to convergence $\mathbb{E}_{\mathbb{P}_n}[f] \to \mathbb{E}_{\mathbb{P}}[f]$ uniformly over all measurable functions $f : \mathcal{X}_V \to [-1,1]$, so it is (much) stronger than weak convergence. By Lauritzen (2024) we have that conditional independence is closed in total variation:

Theorem 4 (Lauritzen, 2024)**.** *Given probability measures $\mathbb{P}, \mathbb{P}_1, \mathbb{P}_2, \ldots \in \mathcal{P}(\mathcal{X}_V)$ 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 Typicality of faithful distributions of Bayesian networks

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

$$\begin{aligned} M_G &:= \Big\{ \mathbb{P} \in \mathcal{P}(\mathcal{X}_V) : A \perp^d_G B \,|\, C \implies X_A \perp\!\!\!\perp_{\mathbb{P}} X_B \,|\, X_C \text{ for all } A, B, C \subseteq V \Big\} \\ F_G &:= \Big\{ \mathbb{P} \in M_G : A \not\perp^d_G B \,|\, C \implies X_A \not\!\perp\!\!\!\perp_{\mathbb{P}} X_B \,|\, X_C \text{ for all } A, B, C \subseteq V \Big\} \\ U_G &:= M_G \setminus F_G. \end{aligned}$$

We will derive properties of F_G and U_G as subsets of the (complete) metric space (M_G, d_{TV}). 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

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

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 main 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.

Theorem 5. *For every $\mathbb{P} \in M_G$ and every $A, B, C \subseteq V$, there is a sequence $\mathbb{P}_n \in M_G$ such that $X_A \not\!\perp\!\!\!\perp_{\mathbb{P}_n} X_B \,|\, X_C$ for all $n \in \mathbb{N}$ and $\mathbb{P}_n \xrightarrow{tv} \mathbb{P}$.*

Proof. By Lemma 1, there exists a $\mathbb{P}_1$ that is Markov and has $X_A \not\!\perp\!\!\!\perp_{\mathbb{P}_1} X_B \,|\, X_C$. Let $\mathbb{P} \in M_G$ be given, then there exists an interpolation $(\mathbb{P}_\lambda)_{\lambda \in (0,1)}$ between $\mathbb{P}$ and $\mathbb{P}_1$ in M_G (Definition 4) such that $X_A \not\!\perp\!\!\!\perp_{\mathbb{P}_\lambda} X_B \,|\, X_C$ for all positive λ below some $\lambda^* \in (0,1)$ (Lemma 2), which converges in total variation to $\mathbb{P}$ as $\lambda \to 0$ (Lemma 3). 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. Given a finite index set V, let G be a fully connected DAG, in which case $M_G = \mathcal{P}(\mathcal{X}_V)$ and the result follows from Theorem 5. ■

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

Theorem 6. *The set of unfaithful distributions F_G is a non-empty, dense and open set, and the unfaithful distributions U_G are nowhere dense.*

Proof. By Theorems 4 and 5 we have for any given $A, B, C \subseteq V$ with $A \not\perp^d_G B \,|\, C$ that $M_G \setminus I_{A,B,C}$ is dense and open. 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 binary 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 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 Conditional dependence is dense in total variation

Lemma 1. *For any DAG G, standard Borel space $\mathcal{X}_V$ and subsets $A, B, C \subseteq V$ such that $A \not\perp^d_G B \,|\, C$, 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,[4] 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 $\tilde{x}_A, \tilde{x}_B, \tilde{x}_C$ 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

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

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 $\mathbb{P}_\lambda(X_A, X_B, X_C) = (1 - \lambda)\mathbb{P}_0(X_A, X_B, X_C) + \lambda\mathbb{P}_1(X_A, X_B, X_C)$ 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 $\alpha \sim \text{Bernoulli}(\lambda)$.

[4] 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.

Its graph is depicted in Figure 2b, from which we see that $\mathbb{P}_\lambda$ need not be Markov with respect to G, as we might have $X_A \not\perp\!\!\!\perp_{\mathbb{P}_\lambda} X_B \mid X_C$. Instead, taking a mixture of the conditional distributions of the Bayesian networks gives $(\mathbb{P}_{\alpha_A}(X_A \mid X_C), \mathbb{P}_{\alpha_B}(X_B \mid X_C), \mathbb{P}_{\alpha_C}(X_C), \mathbb{P}(\alpha_A), \mathbb{P}(\alpha_B), \mathbb{P}(\alpha_C))$ with $\alpha_A, \alpha_B, \alpha_C \sim \text{Bernoulli}(\lambda)$ i.i.d., whose $(A \cup B \cup C)$-marginal $\mathbb{P}_\lambda(X_A, X_B, X_C)$ is Markov with respect to G (see Figure 2c).

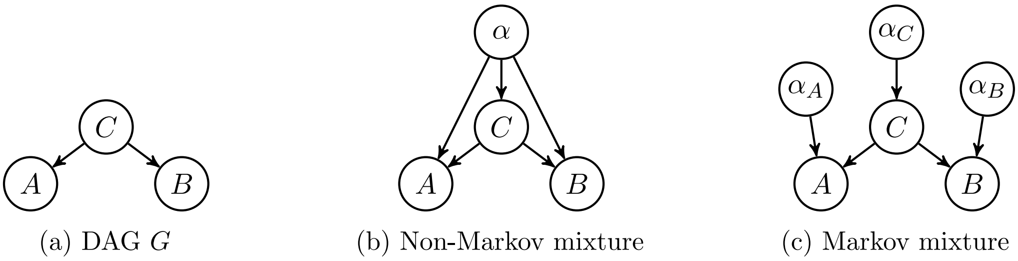


Figure 2: Graphs relating to different mixtures of Bayesian networks with graph G.

This 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

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

It is 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

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

where $d = |V|$ and $(v_1, \ldots, v_d)$ is a topological ordering of G. Our goal is to show that if we have conditional dependence $X_A \not\perp\!\!\!\perp_{\mathbb{P}_1} X_B \mid X_C$ 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.

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.

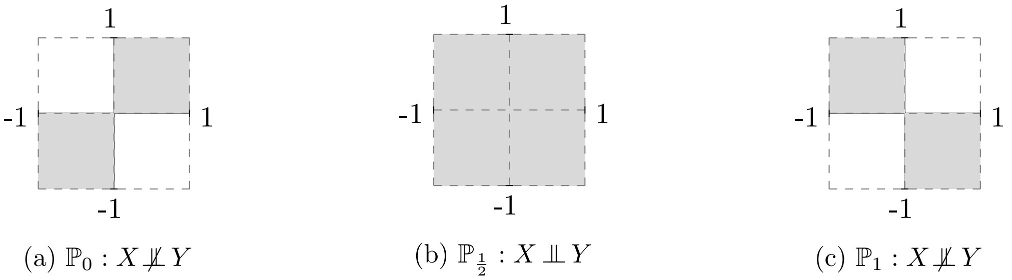

1
-1
1
-1
(a) $\mathbb{P}_0 : X \not\perp\!\!\!\perp Y$
1
-1
1
-1
(b) $\mathbb{P}_{\frac{1}{2}} : X \perp\!\!\!\perp Y$
1
-1
1
-1
(c) $\mathbb{P}_1 : X \not\perp\!\!\!\perp Y$

Figure 3: Mixtures of dependent variables might become independent.

Lemma 2. *Given two distributions* $\mathbb{P}_0, \mathbb{P}_1 \in M_G$ *with independence* $X_A \perp\!\!\!\perp_{\mathbb{P}_0} X_B \,|\, X_C$ *and dependence* $X_A \not\perp\!\!\!\perp_{\mathbb{P}_1} X_B \,|\, X_C$ *and the interpolation* $\mathbb{P}_\lambda$ *from Definition 4, there exists 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 $\mathbb{Q}$. There exist $E_A \in \mathcal{B}(\mathcal{X}_A), E_B \in \mathcal{B}(\mathcal{X}_B), E_C \in \mathcal{B}(\mathcal{X}_C)$ with $\mathbb{P}_1(E_C) > 0$ and $p_1(x_C) > 0$ for all $x_C \in E_C$, and[5]

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

Define

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

for which we have $q(0, x_C) = 0 \neq q(1, x_C)$ for all $x_C \in E_C$. From (2) we see that $q(\lambda, x_C)$ is a non-trivial polynomial in λ for every $x_C \in \mathcal{X}_C$, and so $q(\lambda, x_C) \neq 0$ for all $\lambda \in (0, \lambda^*(x_C))$ with $\lambda^*(x_C)$ the smallest strictly positive root of the polynomial. 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 $\lambda \in (0, \lambda^*)$. Define $E_C^n := \{x_C \in E_C : \lambda^*(x_C) > 1/n\}$, then $E_C^1 \subseteq E_C^2 \subseteq \ldots \subseteq E_C$ with $\lim_n \mathbb{P}_1(E_C^n) = \mathbb{P}_1(E_C) > 0$, so there exists a 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 $\mathbb{P}_1 \ll \mathbb{P}_\lambda$ 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 3. *Given two distributions* $\mathbb{P}_0, \mathbb{P}_1 \in M_G$ *and the interpolation* $\mathbb{P}_\lambda$ *from Definition 4, we have* $\mathbb{P}_\lambda \overset{tv}{\to} \mathbb{P}_0$ *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

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

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

4 Typicality of faithful Bayesian networks

In this section we extend Theorem 6 from the space of observational distributions of Bayesian networks to the space of Bayesian networks:

Definition 5. Given a DAG G with finite index set V, standard Borel $\mathcal{X}_V$, *the space of Bayesian networks* is defined as

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

[5] Note that conditional independence does not imply $\mathbb{P}_1(X_A \in E_A, X_B \in E_B \,|\, X_C \in E_C) \neq \mathbb{P}_1(X_A \in E_A \,|\, X_C \in E_C)\mathbb{P}_1(X_B \in E_B \,|\, X_C \in E_C)$. See also Neykov et al. (2021), p.3.

Whether a Bayesian network is faithful depends on its observational distribution $\mathbb{P} \in M_G$. To formalise the relation between the Bayesian network and the observational distribution we introduce the following mapping:

Definition 6. The *distribution* map is defined as

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

4.1 A pseudometric topology on the space of Bayesian Networks

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.

Definition 7. For $m, m' \in \mathrm{BN}_G$, the pseudometric[6] d° on BN_G is defined as

$$d^\circ(m, m') := d_{TV}(D(m), D(m')).$$

We equip BN_G with the topology generated by the open balls $B(m, r) := \{m' \in \mathrm{BN}_G : d^\circ(m, m') < r\}$ for all $m \in \mathrm{BN}_G$ and $r > 0$. Note that this space is not T_0, meaning that points are not necessarily topologically distinguishable. In particular, we have $d^\circ(m, m') = 0$ for any two m, m' that have the same observational distribution.

The preimage of a dense open set through a function that is open and continuous is dense and open. Hence, a sufficient condition for the faithful Bayesian networks $D^{-1}(F_G)$ to be typical is that the map $D : (\mathrm{BN}, d^\circ) \to (M, d_{TV})$ is open and continuous. This is immediate from the definition of the pseudometric d°, so we get the following result:

Theorem 7. *The set of unfaithful Bayesian networks $D^{-1}(U_G)$ is nowhere dense.*

4.2 Exponential family parametrisations of Bayesian networks

The preceding section begs 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 exponential family parametrisations of Bayesian networks.

Definition 8. A *parametrisation* of a Bayesian network is a set $\Theta \subseteq \mathbb{R}^d$ with $d \in \mathbb{N}$ and a map

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

The corresponding map from the parameter 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 families (Feigin, 1981).

Definition 9. Let $\mathcal{X} \subseteq \mathbb{R}^m$ for some $m \in \mathbb{N}$ and $\mathcal{Y} \subseteq \mathbb{R}$ be codomains of random variables X and Y, then a *conditional exponential family* parametrised by $\theta \in \Theta \subseteq \mathbb{R}^n$ is a set of conditional densities with respect to a σ-finite measure μ on $\mathcal{Y}$ of the form

$$p_\theta(y \mid x) = e^{\eta(\theta)^\top t(y,x) - A(\eta(\theta),x)}$$

for a given sufficient statistic $t : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}^k$ and natural parameter $\eta : \Theta \to \mathbb{R}^k$ such that $A(\eta(\theta), x) < \infty$ for all $\theta \in \Theta$ and $x \in \mathcal{X}$, where $A(\eta, x) := \log \int e^{\eta^\top t(y,x)} \mathrm{d}\mu(y)$.

[6] A pseudometric can have $d(m, m') = 0$ for $m \neq m'$; it is a metric if $d(m, m') > 0$ for all $m \neq m'$.

Given DAG G, we can consider a parametrisation of a Bayesian network over G where every conditional distribution is parametrised by a conditional exponential family, so where for every $v \in V$ we have a sufficient statistic $t_v : \mathcal{X}_{v \cup \mathrm{pa}(v)} \to \mathbb{R}^{k_v}$ for some k_v, a parameter space $\Theta_v \subseteq \mathbb{R}^{n_v}$ for some n_v, a natural parameter $\eta_v : \Theta_v \to N_v$, and consider the conditional distribution $p_{\theta_v}(x_v \,|\, x_{\mathrm{pa}(v)}) = e^{\eta_v(\theta_v)^\top t_v(x_v, x_{\mathrm{pa}(v)}) - A(\eta_v(\theta_v), x_{\mathrm{pa}(v)})}$ with respect to some underlying measure μ_v. This gives rise to a joint density $p_\theta(x_V) = \prod_{v \in V} p_{\theta_v}(x_v \,|\, x_{\mathrm{pa}(v)})$ and the parameter space $\Theta := \prod_{v \in V} \Theta_v$. Note that 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 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 $\eta_v(\theta_v)^\top = (\beta_v^\top/\sigma_v^2, -1/(2\sigma_v^2))$, and dominating measure $\mu_v = \lambda/\sqrt{2\pi}$, where λ denotes 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 $|\mathcal{X}_v|$-dimensional simplex. This gives the parameter space

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

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, and the dominating measure μ_v is the counting measure.

In the main result of this section, we require the natural parameters to be sufficiently regular, such that the marginal densities $p_\theta(x_A)$ are analytic functions in the parameter θ.

Theorem 8. *Let G be a DAG and let $\varphi : \Theta \to \mathrm{BN}_G$ be a parametrisation of the Bayesian network with Θ open and connected, such that the marginal density $p_\theta(x_A)$ is analytic in θ for each $A \subseteq V$. If there is at least one faithful parameter in Θ, then the set of unfaithful parameters is nowhere dense and has Lebesgue measure zero.*

Proof. Let $A, B, C \subseteq V$ such that $A \not\perp_G^d B \,|\, C$, 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 be such that $p_{\theta_1}(x_A, x_B \,|\, x_C) \neq p_{\theta_1}(x_A \,|\, x_C) p_{\theta_1}(x_B \,|\, x_C)$ and define

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

for which we have $q(\theta_1) \neq 0$. For any $\theta \in \Theta$ with conditional independence $X_A \perp\!\!\!\perp_{\mathbb{P}_\theta} X_B \,|\, X_C$ we have $q(\theta) = 0$, so $\{\theta \in \Theta : X_A \perp\!\!\!\perp_{\mathbb{P}_\theta} X_B \,|\, X_C\} \subseteq \{\theta \in \Theta : q(\theta) = 0\}$. It follows from the identity theorem (see e.g. Krantz and Parks (2002)) that the zero set of a nonconstant real analytic function on an open and connected domain is nowhere dense and has Lebesgue measure zero. Since $q(\theta)$ is analytic, we have that $T^{-1}(U_G) = \bigcup_{A \not\perp_G^d B \,|\, C} \{\theta \in \Theta : X_A \perp\!\!\!\perp_{\mathbb{P}_\theta} X_B \,|\, X_C\}$ is nowhere dense and has Lebesgue measure zero, which is the desired result. ■

Remark 1. The proof of Theorem 8 only requires that for every d-connection $A \not\perp_G^d B \,|\, C$ 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 with respect to all d-connections. 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.

Remark 2. Theorem 8 also holds when Θ is the closure of an open convex set with a faithful parameter in its interior, since the boundary of an open set is nowhere dense, and the boundary of a convex set has Lebesgue measure zero.

We specifically focus on exponential families because their marginal densities satisfy the required regularity condition, provided the natural parameters are sufficiently regular.

Theorem 9. *For an exponential family parametrisation of a Bayesian network with $\eta_v(\theta_v)$ analytic, every marginal density $p_\theta(x_A)$ is analytic in θ.*

Proof. Let $\tilde{p}_\theta(x_v \,|\, x_{\mathrm{pa}(v)}) = e^{\eta(\theta_v)^\top t(x_v, x_{\mathrm{pa}(v)})}$ be the unnormalised density. Considering a topological ordering $(v_1, ..., v_d)$ of G we can write $\eta(\theta)^\top = (\eta_{v_1}(\theta_{v_1})^\top, ..., \eta_{v_d}(\theta_{v_d})^\top)$ and $t(x_V)^\top = (t_{v_1}(x_{v_1}, x_{\mathrm{pa}(v_1)})^\top, ..., t_{v_d}(x_{v_d})^\top)$ so that the unnormalised joint density can be expressed as $\tilde{p}_\theta(x_V) = e^{\eta(\theta)^\top t(x_V)}$. From Brown (1986), Lemma 2.8 it follows that $\eta \mapsto \int_{\mathcal{X}_I} e^{\eta^\top t(x_V)} \mathrm{d}\mu(x_I)$ is analytic for every $I \subseteq V$. As a composition of analytic functions, so is $\theta \mapsto \int_{\mathcal{X}_I} e^{\eta(\theta)^\top t(x_V)} \mathrm{d}\mu(x_I)$. Then the following functions are analytic in θ: the unnormalised density $\tilde{p}_\theta(x_A) = \int_{\mathcal{X}_{V \setminus A}} e^{\eta(\theta)^\top t(x_V)} \mathrm{d}\mu(x_{V \setminus A})$, the normalisation constant $Z(\theta) := \int_{\mathcal{X}_V} e^{\eta(\theta)^\top t(x_V)} \mathrm{d}\mu(x_V)$, and finally the density $p_\theta(x_A) = \tilde{p}_\theta(x_A)/Z(\theta)$. ■

Given that Spirtes et al. (1993) and Meek (1995) have shown for every DAG G the existence of faithful parameters in $\Theta_{\mathcal{N}}$ and $\Theta_{\mathcal{D}}$, we obtain Theorems 1 and 2 and their topological analogues as corollaries of Theorems 8 and 9:

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

Corollary 3. *The set of unfaithful parameters $\{\theta \in \Theta_{\mathcal{D}} : T(\theta) \in U_G\}$ is nowhere dense and Lebesgue measure zero.*[7]

5 Bayesian networks with latent variables

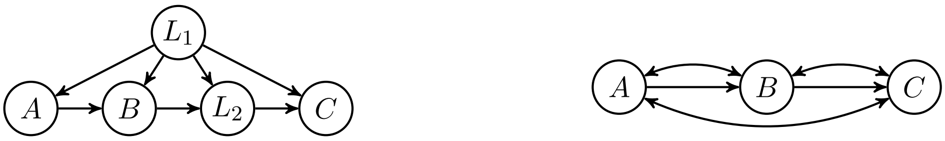
The assumption that all variables in the Bayesian network must be observed is often too restrictive in practice. When certain variables remain unobserved, a suitable modelling class is that of Bayesian networks with observed variables V and latent variables W.

Given a DAG G over $V \cup W$, the *latent projection* of G onto V is the *Acyclic Directed Mixed Graph* (ADMG) G^{p} 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^{p} is given in Figure 4.

The definition of d-separation for ADMGs (also known as *m-separation* (Richardson, 2003)) employs an extended notion of a collider: given ADMG G^{p} with path $\pi = a \ast\!\!-\!\!\ast \dots \ast\!\!-\!\!\ast b$, 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 $A \perp^d_G B \,|\, C$, if for every path $\pi = a \ast\!\!-\!\!\ast \dots \ast\!\!-\!\!\ast b$ between every $a \in A$ and $b \in B$, there is a collider in π that is not an ancestor of C, or if there is a non-collider in π in C. The independence models of G and G^{p} with respect to V are equal: for any $A, B, C \subseteq V$ we have $A \perp^d_G B \,|\, C$ if and only if $A \perp^d_{G^{\mathrm{p}}} B \,|\, C$ (Verma, 1993). As a corollary the Markov property (1) also holds for the latent projection G^{p} 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 $U_{G^{\mathrm{p}}}, F_{G^{\mathrm{p}}}$ for the distributions over $\mathcal{X}_{V \cup W}$ that are

[7] That this set is nowhere dense has also been shown by Lin and Zhang (2020).



(a) DAG G

(b) Latent projection G^{p}

Figure 4: DAG G and latent projection G^{p} onto $\{A, B, C\}$.

unfaithful and faithful with respect to the ADMG G^{p} respectively. The core observation for extending results of Sections 3 and 4 from DAGs to ADMGs is the following:

Lemma 4. *Given DAG G with vertices $V \cup W$ and its latent projection G^{p} onto V, any distribution over $V \cup W$ that is unfaithful with respect to G^{p} 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:

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

For $U_{G^{\mathrm{p}}}$ 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 $U_{G^{\mathrm{p}}} \subseteq U_G$. ■

Now, we can extend preceding results to ADMGs as follows:

Theorem 10. *Theorems 6, 7, 8 and Corollaries 2 and 3 also hold for Bayesian networks with latent variables, when faithfulness is only required to hold with respect to the latent projection.*

Proof. Subsets of nowhere dense (Lebesgue measure zero) sets are nowhere dense (Lebesgue measure zero). The extension of Theorem 6 immediately follows from Lemma 4. Considering the distribution map $D : \mathrm{BN}_G \to M_G$ we get the extension of Theorem 7 from the inclusion $D^{-1}(U_{G^{\mathrm{p}}}) \subseteq D^{-1}(U_G)$, and given a parametrisation $\varphi : \Theta \to \mathrm{BN}_G$ and map $T = D \circ \varphi$, the extension of Theorem 8 (and the related corollaries) follows from $T^{-1}(U_{G^{\mathrm{p}}}) \subseteq T^{-1}(U_G)$. ■

6 Discussion

One should be careful with interpreting the typicality results from this work and from Spirtes et al. (1993) and Meek (1995), as the employed notion of 'typicality' depends on somewhat arbitrary choices. The choice of σ-ideal makes an essential difference: the σ-ideals of null sets and meager sets 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 null set (Oxtoby, 1980, Theorem 1.6): a set that is small in one sense may be large in the other sense. When considering the σ-ideal of measure-zero sets, the results depend on the choice of σ-algebra and the probability measure. For the σ-ideal of meager sets, the results depend on the choice of the topology. The pseudometric topology that we consider on the space BN_G might be too weak for purposes of causal modelling, as it does not distinguish between two causal models that have different causal mechanisms but the same observational distribution.

Faithfulness might be an assumption that is too weak for the purposes of causal discovery, as faithful distributions can have extremely weak dependencies that are undetectable from finite samples. The (perhaps more practically relevant) notion of *strong faithfulness* of linear Gaussian Bayesian networks (Zhang and Spirtes, 2002) is not measure-zero, as is shown by Uhler et al. (2013). It is unclear whether or not it is typical in a topological sense.

From a philosophical perspective, it is absolutely unclear whether 'in nature, unfaithful Bayesian networks are nowhere dense', just as there is no reason to believe that 'nature picks parametric Bayesian networks via a distribution that has a density'. At least we can view it as a positive result that the opposite of our result, i.e. that unfaithful distributions are typical, does *not* hold.

7 Acknowledgements

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

References

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.

Brown, L. D. (1986). Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory. *Lecture Notes-Monograph Series*, 9:i–279.

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.

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.

Krantz, S. G. and Parks, H. R. (2002). *A Primer of Real Analytic Functions.* Birkhäuser Advanced Texts, Basler Lehrbücher. Birkhäuser, Boston, MA, second edition edition.

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.

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.

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.

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

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

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

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. 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.