Nonparametric Bayesian networks are typically faithful in the total variation metric

Philip Boeken Patrick Forré Joris M. Mooij

University of Amsterdam

October 22, 2024

Abstract

We show that for a given DAG G, among all observational distributions of Bayesian networks over G with arbitrary outcome spaces, the faithful distributions are 'typical': they constitute a dense, open set with respect to the total variation metric. As a consequence, the set of faithful distributions is non-empty, and the unfaithful distributions are nowhere dense. We extend this result to the space of Bayesian networks, where the properties hold for *Bayesian networks* instead of *distributions of Bayesian networks.* As special cases, we show that these results also hold for the faithful *parameters* of the subclasses of linear Gaussian– and discrete Bayesian networks, giving a topological analogue of the measure-zero results of Spirtes et al. (1993) and Meek (1995). Finally, we extend our topological results and the measure-zero results of Spirtes et al. and Meek to 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 (in)dependencies $X_A \perp\!\!\!\perp_{\mathbb{P}} X_B \mid X_C$ for multiple choices of $A, B, C \subseteq V$, and use this information to reconstruct G (up to certain equivalences). A core assumption of almost all 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

$$A \overset{d}{\underset{G}{\perp}} B \mid X \iff X_A \underset{\mathbb{P}}{\perp\!\!\!\perp} X_B \mid X_C. \tag{1}$$

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 3 below). A Bayesian network for which (1) holds is called *faithful.*

In absence of any knowledge of the true causal graph, 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 from $\Theta_{\mathcal{N}}$ or $\Theta_{\mathcal{D}}$ at random will give with probability one a faithful Bayesian network:

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 result is as follows:

> **For a given DAG G, among all distributions that are Markov with respect to G, 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'. The topological properties are with respect to the total variation metric on the joint distribution $\mathbb{P}(X_V)$ of all variables V of the Bayesian network. Our 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.

Considering nowhere dense sets as a notion of atypicality stands on the following theoretical footing. 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 as considered by Spirtes et al. (1993) and Meek (1995) is a σ-ideal, and so is the family of meager sets:

Definition 1. A set $I \subseteq M$ is *dense* in another set $F \subseteq M$ if every point in F 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.

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 (see Example 1), 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 are nowhere dense, which is an even a stronger notion of atypicality.

Example 1. The set of integers $\mathbb{Z}$ is nowhere dense in $\mathbb{R}$, and the rationals $\mathbb{Q}$ are meager in $\mathbb{R}$.

Example 2. A straightforward proof of the existence of nowhere differentiable continuous functions is to show that they are comeager in the space of continuous functions, hence dense, hence non-empty.[1]

We will use a similar reasoning to prove the existence of faithful distributions: they are the complement of a nowhere dense set, hence dense, hence non-empty.

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[2] are meager, which is a topological analogue of the Lebesgue measure-zero result from Bareinboim et al. (2022).

The contribution and 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 result: that faithful distributions are dense and open. In Section 4 we lift this result from the space of observational distributions to the space of Bayesian networks, i.e. the space of tuples of conditional distributions that define the Bayesian networks. In Section 4.1 we focus 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 (non)parametric topological results and the parametric measure-zero results of Spirtes et al. and Meek to Bayesian networks with latent variables.

[1] This result uses the Baire Category Theorem: comeager subsets of complete metric spaces are dense.

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

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. 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 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

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

where $D \in \mathcal{B}(\mathcal{X}_{A \cup B})$. Since $\mathcal{X}_V$ is standard Borel, there exists for any $A, B \subseteq V$ and joint distribution $\mathbb{P}(X_A, X_B)$ 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.[3]

Writing $\mathrm{pa}(v)$ for the set of parents of v in G, a *Bayesian network over* G is defined as 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. For sets of vertices $A, B, C \subseteq V$ we say that A and B are *d-separated* given C, written $A \perp_G^d B \,|\, C$, if for every path $\pi = a \ast\!\!-\!\!\ast \ldots \ast\!\!-\!\!\ast b$ 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.

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

$$A \underset{G}{\overset{d}{\perp}} B \,|\, C \implies X_A \underset{\mathbb{P}}{\perp\!\!\!\perp} X_B \,|\, X_C \tag{2}$$

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

In general, the set of all conditional independencies in $\mathbb{P}$ does not characterise the set of d-separations in G: we might have a d-connection $A \not\perp_G^d B \,|\, C$ 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

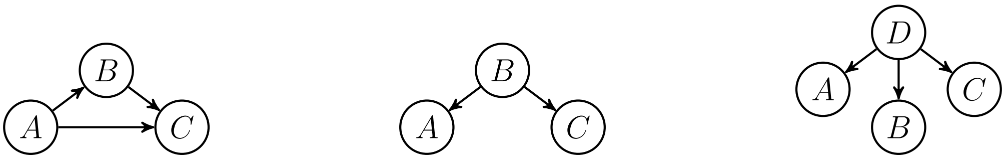
$$A \underset{G}{\overset{d}{\not\perp}} B \,|\, C \implies X_A \underset{\mathbb{P}}{\not\perp\!\!\!\perp} X_B \,|\, X_C.$$

Example 3. The following Bayesian networks are unfaithful. Corresponding graphs are shown in Figure 1.

a) Cancelling paths: let $\mathbb{P}(X_A)$ be any distribution and let $\mathbb{P}(X_B \,|\, X_A) = \mathcal{N}(\beta_{AB} X_A, \sigma_B^2)$, $\mathbb{P}(X_C \,|\, X_A, X_B) = \mathcal{N}(\beta_{AC} X_A + \beta_{BC} X_B, \sigma_C^2)$ for given parameters $\sigma_A^2, \sigma_B^2, \sigma_C^2 > 0$ and $\beta_{AC}, \beta_{AB}, \beta_{BC} \in \mathbb{R}$ with $\beta_{AC} = -\beta_{AB}\beta_{BC}$. Then $A \not\perp_{G^a}^d C$ and $X_A \perp\!\!\!\perp X_C$.[4]

[3] This is equivalent to independence of the 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 $p(x_C) > 0$.

[4] 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).



(a) DAG G^a

(b) DAG G^b

(c) DAG G^c

Figure 1: DAGs of the Bayesian networks given in Example 3.

b) Deterministic variables: let $\mathbb{P}(X_A \mid X_B)$ and $\mathbb{P}(X_C \mid X_B)$ 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 $\mathbb{P}(X_A \mid X_D)$ and $\mathbb{P}(X_C \mid X_D)$ be Markov kernels and $\mathbb{P}(X_D)$ any distribution and let $\mathbb{P}(X_B \mid X_D) = \delta_{X_D}$, so we deterministically set $X_B = X_D$. Then we have $A \not\perp^d_{G^c} C \mid B$ and $X_A \perp\!\!\!\perp X_C \mid X_B$.[5]

As an important step in the proof of the typicality of faithful distributions, we use that conditional independence is a preserved under taking limits. However, 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 $\mathbb{P}, \mathbb{P}_1, \mathbb{P}_2, \ldots \in \mathcal{P}(\mathcal{X}_V)$ 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 bounded, 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 $X_A \perp\!\!\!\perp_{\mathbb{P}_n} X_B \mid X_C$ for all $n \in \mathbb{N}$, we might have $X_A \not\perp\!\!\!\perp_{\mathbb{P}} X_B \mid X_C$; see e.g. Lauritzen (1996), pp. 38-39. Instead of weak convergence, we consider a different type of convergence:

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 $\mathbb{P}_n \overset{tv}{\to} \mathbb{P}$. It is equivalent to convergence $\mathbb{E}_{\mathbb{P}_n}[f] \to \mathbb{E}_{\mathbb{P}}[f]$ uniformly over all bounded measurable functions $f : \mathcal{X}_V \to [-1, 1]$, so it is (much) stronger than weak convergence. By Lauritzen (2024) we have the following result:

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

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:

$$M_G := \left\{ \mathbb{P} \in \mathcal{P}(\mathcal{X}_V) : A \overset{d}{\underset{G}{\perp}} B \mid C \implies X_A \underset{\mathbb{P}}{\perp\!\!\!\perp} X_B \mid X_C \text{ for all } A, B, C \subseteq V \right\} \tag{3}$$

$$F_G := \left\{ \mathbb{P} \in M_G : A \overset{d}{\underset{G}{\not\perp}} B \mid C \implies X_A \underset{\mathbb{P}}{\not\perp\!\!\!\perp} X_B \mid X_C \text{ for all } A, B, C \subseteq V \right\} \tag{4}$$

$$U_G := M_G \setminus F_G. \tag{5}$$

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

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

Theorem 5. *The set of faithful distributions F_G is a non-empty, dense and open set, and the unfaithful distributions U_G are nowhere dense.*

The proof refers to some technical lemmas that are given in Section 3.1.

Proof. First, we show for any given $A, B, C \subseteq V$ with $A \not\perp_G^d B \mid C$ that $M_G \setminus I_{A,B,C}$ is dense and open, where we write $I_{A,B,C} = \{\mathbb{P} \in M_G : X_A \perp\!\!\!\perp_{\mathbb{P}} X_B \mid X_C\}$. By Lemma 1, there exists a $\mathbb{P}_1$ that is Markov and has $X_A \not\!\perp\!\!\!\perp_{\mathbb{P}_1} X_B \mid X_C$, so $\mathbb{P}_1 \in M_G \setminus I_{A,B,C}$. Given any $\mathbb{P}_0 \in I_{A,B,C}$ (so $\mathbb{P}_0$ is unfaithful w.r.t. G) and $\mathbb{P}_1 \in M_G \setminus I_{A,B,C}$, there exists a net $(\mathbb{P}_\lambda)_{\lambda \in (0,\lambda^*)} \subseteq M_G \setminus I_{A,B,C}$ (Definition 4, Lemma 2) that interpolates between $\mathbb{P}_0$ and $\mathbb{P}_1$, and which converges in total variation to $\mathbb{P}_0$ as $\lambda \to 0$ (Lemma 3), hence $M_G \setminus I_{A,B,C}$ is dense. With this construction, we approximate any Markov distribution that has a particular faithfulness violation by Markov distributions that don't have this violation. By Theorem 4, $M_G \setminus I_{A,B,C}$ is open.

We can write $F_G = \bigcap_{A \not\perp_G^d B \mid C} (M_G \setminus I_{A,B,C})$ if there is a d-connection $A \not\perp_G^d B \mid C$, and $F_G = M_G$ otherwise. 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 nowhere dense since it is the complement of a dense open set. ∎

To conclude, unfaithful distributions are 'atypical': there is no open set of distributions that are Markov w.r.t. 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 Proof of Theorem 5

Lemma 1. *For any DAG G, standard Borel space $\mathcal{X}_V$ and subsets $A, B, C \subseteq V$ such that $A \not\perp_G^d B \mid C$, there exists a distribution $\mathbb{P} \in M_G$ with the conditional dependence $X_A \not\!\perp\!\!\!\perp_{\mathbb{P}} X_B \mid X_C$.*

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 $X_A \not\!\perp\!\!\!\perp_{\mathbb{P}} X_B \mid X_C$ also has $X_{A'} \not\!\perp\!\!\!\perp_{\mathbb{P}} X_{B'} \mid X_C$ 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 $p \in (0,1)$. Then $\mathbb{P}(X_V)$ is Markov and $X_A \not\!\perp\!\!\!\perp_{\mathbb{P}} X_B \mid X_C$. If $A \neq B$, then by Meek (1998) Lemma 3,[6] there exists a binary distribution $\tilde{\mathbb{P}}$ on $\{0,1\}^{|V|}$ that is Markov with respect to G and which has the conditional dependence $X_A \not\!\perp\!\!\!\perp_{\tilde{\mathbb{P}}} X_B \mid X_C$, so there are $\tilde{x}_A, \tilde{x}_B, \tilde{x}_C$ with $\tilde{\mathbb{P}}(\tilde{x}_C) > 0$ such that $\tilde{\mathbb{P}}(\tilde{x}_A, \tilde{x}_B \mid \tilde{x}_C) \neq \tilde{\mathbb{P}}(\tilde{x}_A \mid \tilde{x}_C)\tilde{\mathbb{P}}(\tilde{x}_B \mid \tilde{x}_C)$. Define the pushforward $\mathbb{P}(X_V) := \tilde{\mathbb{P}} \circ f_V^{-1}$, which has

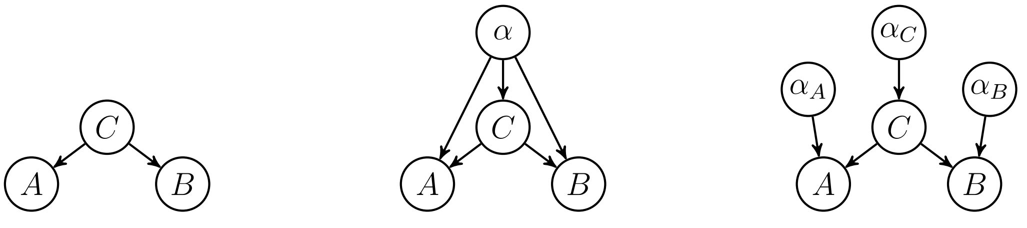
$$\begin{aligned}\mathbb{P}(X_A = f_A(\tilde{x}_A), X_B &= f_B(\tilde{x}_B) \mid X_C = f_C(\tilde{x}_C)) \\ &= \tilde{\mathbb{P}}(\tilde{x}_A, \tilde{x}_B \mid \tilde{x}_C) \neq \tilde{\mathbb{P}}(\tilde{x}_A \mid \tilde{x}_C)\tilde{\mathbb{P}}(\tilde{x}_B \mid \tilde{x}_C) \\ &= \mathbb{P}(X_A = f_A(\tilde{x}_A)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

so indeed $X_A \not\!\perp\!\!\!\perp_{\mathbb{P}} X_B \mid X_C$. By a similar reasoning the conditional independence $X_A \perp\!\!\!\perp_{\tilde{\mathbb{P}}} X_B \mid X_C$ implies $X_A \perp\!\!\!\perp_{\mathbb{P}} X_B \mid X_C$, and thus $\mathbb{P} \in M_G$. ∎

Next, we aim to construct an interpolation of $\mathbb{P}_0, \mathbb{P}_1 \in M_G$ 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:

[6] 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.

Example 4. Let $(\mathbb{P}_i(X_A \mid X_C), \mathbb{P}_i(X_B \mid X_C), \mathbb{P}_i(X_C))$ for $i \in \{0,1\}$ be Bayesian networks with DAG G as depicted in Figure 2a, which both have $X_A \perp\!\!\!\perp X_B \mid X_C$. 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 $(\mathbb{P}_\alpha(X_A \mid X_C), \mathbb{P}_\alpha(X_B \mid X_C), \mathbb{P}_\alpha(X_C), \mathbb{P}(\alpha))$ with $\alpha \sim \text{Bernoulli}(\lambda)$. Its graph is depicted in Figure 2b, from which we see that $\mathbb{P}_\lambda$ is not 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 corresponds to considering $(\mathbb{P}_{\alpha_A}(X_A \mid X_C), \mathbb{P}_{\alpha_B}(X_B \mid X_C), \mathbb{P}_{\alpha_C}(X_C), \mathbb{P}(\alpha_A), \mathbb{P}(\alpha_B), \mathbb{P}(\alpha_C))$ 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).



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

Figure 2

The preceding example motivates the following definition:

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

$$\begin{aligned} p_\lambda(x_V) &= \prod_{v \in V} \left((1-\lambda) p_0(x_v \mid x_{\mathrm{pa}(v)}) + \lambda p_1(x_v \mid x_{\mathrm{pa}(v)}) \right) \\ &= \sum_{\alpha \in \{0,1\}^d} (1-\lambda)^{d-|\alpha|} \lambda^{|\alpha|} p_{\alpha_d}(x_{v_d} \mid x_{\mathrm{pa}(v_d)}) ... p_{\alpha_1}(x_{v_1}) \end{aligned} \tag{6}$$

where $d = |V|$ and $(v_1, ..., v_d)$ is a topological ordering of G.

Lemma 2. *Given two distributions $\mathbb{P}_0, \mathbb{P}_1 \in M_G$ with independence $X_A \perp\!\!\!\perp_{\mathbb{P}_0} X_B \mid X_C$ and dependence $X_A \not\perp\!\!\!\perp_{\mathbb{P}_1} X_B \mid X_C$ and the interpolation $\mathbb{P}_\lambda$ from Definition 4, there exists a $\lambda^* \in (0,1)$ such that $X_A \not\perp\!\!\!\perp_{\mathbb{P}_\lambda} X_B \mid X_C$ 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 $E_A \in \mathcal{B}(\mathcal{X}_A), E_B \in \mathcal{B}(\mathcal{X}_B), E_C \in \mathcal{B}(\mathcal{X}_C)$ such that

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

and $p_1(x_C) > 0$ for all $x_C \in E_C$.[7] Define

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

so we have $q(0, x_C) = 0 \neq q(1, x_C)$ for all $x_C \in E_C$. From (6) 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 $X_A \not\perp\!\!\!\perp_{\mathbb{P}_\lambda} X_B \mid X_C$ for all $\lambda \in (0, \lambda^*)$. Define $E_C^n := \{x_C \in E_C : \lambda^*(x_C) > 1/n\}$, then $E_C^1 \subseteq E_C^2 \subseteq \dots \subseteq E_C$ with $\lim_n \mathbb{Q}(E_C^n) = \mathbb{Q}(E_C) > 0$, so there exists a N such that $\mathbb{Q}(E_C^n) > 0$ 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 $\mathbb{Q} \ll \mathbb{P}_\lambda$ for all $\lambda \in (0,1)$ we also have $\mathbb{P}_\lambda(E_C^*) > 0$, implying that $X_A \not\perp\!\!\!\perp_{\mathbb{P}_\lambda} X_B \mid X_C$ 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 (6) we get the expression

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

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 5 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 $\mathcal{X}_V$, define *the space of Bayesian networks* as

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

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

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

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

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 $m_0, m_1 \in \mathrm{BN}_G$, the pseudometric[8] d° on BN_G is defined as

$$d^\circ(m_0, m_1) := d_{TV}(D(m_0), D(m_1)).$$

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 $d^\circ(m_0, m_1) = 0$ for any two m_0, m_1 that have the same observational distribution.

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, which is at the core of the main result of this section:

Theorem 6. *The set of faithful Bayesian networks $D^{-1}(F_G)$ is a non-empty, dense and open set, and the unfaithful Bayesian networks $D^{-1}(U_G)$ are nowhere dense.*

Proof. By Theorem 5, F_G is a dense open set. For every $\mathbb{P} \in M_G$ we can pick a version of the tuple of disintegrations $(\mathbb{P}(X_v \mid X_{\mathrm{pa}(v)}))_{v \in V}$, which is an element of BN_G, hence D is surjective. Since D is surjective, for every $\mathbb{P} \in M_G$ there is a $m \in \mathrm{BN}_G$ such that $\mathbb{P} = D(m)$, and so we have $D^{-1}(B(\mathbb{P}, r)) = D^{-1}(B(D(m), r)) = B(m, r)$, so D is continuous, and therefore $D^{-1}(F_G)$ is open as well.

Similarly, we have $D(B(m, r)) = B(D(m), r)$, so D is open. Since F_G is dense, for any open $O \subseteq \mathrm{BN}_G$ the open set $D(O)$ intersects F_G, and so $D^{-1}(F_G)$ intersects with O, implying that $D^{-1}(F_G)$ is dense and non-empty.

Finally, $D^{-1}(U_G)$ is nowhere dense since it is the complement of a dense open set. ■

4.1 Parametrised subclasses 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.

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

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

and the corresponding *distribution map* is defined as

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

Remark 1. Similar to the previous section the question is whether $T^{-1}(F_G)$ is typical in Θ – a sufficient condition is that T is open and continuous. Given the fact that $D : \mathrm{BN}_G \to M_G$ is continuous and open, T is continuous (open) if and only if φ is continuous (open).

In the following sections we treat the linear Gaussian and discrete parametrisations from Spirtes et al. (1993) and Meek (1995) separately.

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

4.1.1 Linear Gaussian

For linear Gaussian Bayesian networks with mean zero, Spirtes et al. (1993) parametrise for each $v \in V$ the conditional distribution $\mathbb{P}(X_v \,|\, x_{\mathrm{pa}(v)})$ by a linear coefficient β_v and a variance σ_v^2. This gives the parameter space

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

and the map $\varphi_{\mathcal{N}}(\theta)$ is described by the correspondence $\mathbb{P}_\theta(X_v \,|\, x_{\mathrm{pa}(v)}) = \mathcal{N}(\beta_v^T x_{\mathrm{pa}(v)}, \sigma_v^2)$ for each $v \in V$.

For multivariate Gaussians $\mathbb{P}_\theta(X_V)$, conditional independence $X_A \perp\!\!\!\perp_{\mathbb{P}_\theta} X_B \,|\, X_C$ with subsets $A, B, C \subseteq V$ is equivalent to zero partial covariance $q_{AB.C}(\theta) = 0$ (Baba et al., 2004). Spirtes et al. show that the partial covariance $q_{AB.C}(\theta)$ is a polynomial in θ, and that there are $\theta_0, \theta_1 \in \Theta_{\mathcal{N}}$ such that $q_{AB.C}(\theta_0) = 0 \neq q_{AB.C}(\theta_1)$, so $q_{AB.C}$ is non-trivial. The unfaithful parameters can be expressed as follows:

$$T_{\mathcal{N}}^{-1}(U_G) = \bigcup_{A \not\perp_G^d B \,|\, C} \{\theta \in \Theta_{\mathcal{N}} : q_{AB.C}(\theta) = 0\}. \tag{7}$$

Since the roots of a multivariate real polynomial have Lebesgue measure zero and finite unions of measure-zero sets have measure zero, we get $\lambda[T_{\mathcal{N}}^{-1}(U_G)] = 0$. This result immediately extends to any distribution over $\Theta_{\mathcal{N}}$ that has a density with respect to Lebesgue measure. As a corollary the set $T_{\mathcal{N}}^{-1}(F_G)$ has strictly positive measure, implying the existence of a faithful linear Gaussian Bayesian network.

To prove the topological analogue, we cannot employ Remark 1 because $T_{\mathcal{N}}(\Theta_{\mathcal{N}})$ is not open: one can for example approximate standard Gaussian densities with standard Student's t densities, which by Scheffé (1947) implies that $T_{\mathcal{N}}(\Theta_{\mathcal{N}})$ can be approximated in total variation by its complement. Instead, we give a direct proof relying on properties of $q_{AB.C}$:

Theorem 7. *The set of faithful parameters $T_{\mathcal{N}}^{-1}(F_G)$ is a non-empty, dense and open set, and the unfaithful parameters $T_{\mathcal{N}}^{-1}(U_G)$ are nowhere dense.*

Proof. By (7) we can write $T_{\mathcal{N}}^{-1}(F_G) = \cap_{A \not\perp_G^d B \,|\, C} \{\theta \in \Theta_{\mathcal{N}} : q_{AB.C} \neq 0\}$ if there is a d-connection $A \not\perp_G^d B \,|\, C$, and $T_{\mathcal{N}}^{-1}(F_G) = \Theta_{\mathcal{N}}$ otherwise. Since $\mathbb{R} \setminus \{0\}$ is open and $q_{AB.C}$ is continuous, the set $T_{\mathcal{N}}^{-1}(F_G)$ is open. The complement of the set of roots of any non-trivial real polynomial is dense. Finite intersections of dense, open sets are dense, so $T_{\mathcal{N}}^{-1}(F_G)$ is dense, and therefore non-empty. Finally, $T_{\mathcal{N}}^{-1}(U_G)$ is nowhere dense since it is the complement of a dense open set. ■

For any two parameters $\theta_0, \theta_1 \in \Theta_{\mathcal{N}}$ with $\theta_0 \neq \theta_1$ we have $T_{\mathcal{N}}(\theta_0) \neq T_{\mathcal{N}}(\theta_1)$, so on the set of linear Gaussian Bayesian networks $\varphi_{\mathcal{N}}(\Theta_{\mathcal{N}})$, the pseudometric d° is a proper metric.

4.1.2 Discrete

For discrete distributions with finite state space, Meek (1995) considers the parametrisation $\varphi_{\mathcal{D}}(\theta)$ described for each $v \in V$ by the correspondence $\mathbb{P}_\theta(x_v \,|\, x_{\mathrm{pa}(v)}) = \theta_{v, x_v, x_{\mathrm{pa}(v)}}$ for some parameter $\theta_{v, x_v, x_{\mathrm{pa}(v)}}$. This gives the parameter space

$$\Theta_{\mathcal{D}} := \prod_{v \in V} \left\{ \theta_{v, x_v, x_{\mathrm{pa}(v)}} \in [0,1] : x_v \in \mathcal{X}_v, x_{\mathrm{pa}(v)} \in \mathcal{X}_{\mathrm{pa}(v)}, \sum_{x_v \in \mathcal{X}_v} \theta_{v, x_v, x_{\mathrm{pa}(v)}} = 1 \right\}.$$

Meek shows for every $\theta \in \Theta_{\mathcal{D}}$ and subsets $A, B, C \subseteq V$ that $X_A \perp\!\!\!\perp_{\mathbb{P}_\theta} X_B \,|\, X_C$ is equivalent to $q_{A,B,C}(\theta) = 0$ for some non-trivial polynomial $q_{A,B,C} : \Theta_{\mathcal{D}} \to \mathbb{R}$. Similar to the Gaussian case we express the unfaithful parameters as

$$T_{\mathcal{D}}^{-1}(U_G) = \bigcup_{A \not\perp_G^d B \,|\, C} \{\theta \in \Theta_{\mathcal{D}} : q_{A,B,C}(\theta) = 0\}$$

which implies $\lambda[T_{\mathcal{D}}^{-1}(U_G)] = 0$. This result also immediately extends to any distribution over $\Theta_{\mathcal{D}}$ that has a density with respect to Lebesgue measure. As in the Gaussian case, the set $T_{\mathcal{D}}^{-1}(F_G)$ has strictly positive measure, so there exists a faithful discrete Bayesian network.

For this model class we get the following topological analogue:

Theorem 8. *The set of faithful parameters $T_{\mathcal{D}}^{-1}(F_G)$ is a non-empty, dense and open set, and the unfaithful parameters $T_{\mathcal{D}}^{-1}(U_G)$ are nowhere dense.*

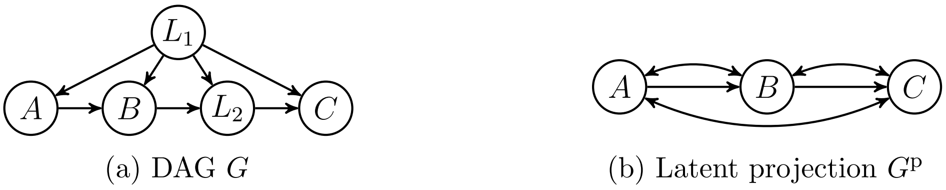
Proof. The proof is analogous to the proof of Theorem 7 and is therefore omitted. ■

For any two parameters $\theta_0 \neq \theta_1 \in \mathrm{Int}(\Theta_{\mathcal{D}})$ we have $T_{\mathcal{D}}(\theta_0) \neq T_{\mathcal{D}}(\theta_1)$, so on the set $\varphi(\mathrm{Int}(\Theta_{\mathcal{D}}))$ of discrete Bayesian networks with strictly positive distributions, the pseudometric d° is a proper metric. On the extreme points of $\Theta_{\mathcal{D}}$, this property might be violated.

5 Typicality of faithful Bayesian networks with latent variables

In practice, the assumption that all variables in the Bayesian network must be observed is often too restrictive. When certain variables remain unobserved, a suitable modelling class is that of Bayesian networks with observed variables V and latent variables W. Of particular interest is the resulting *semi-Markovian* model over the observed variables.

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 $a \to w_1 \to \dots \to w_n \to b$ in G with $w_i \in W$ for all $i = 1, \dots, n$ (if any), and bi-directed edges $a \leftrightarrow b$ if there is a fork $a \leftarrow w_1 \leftarrow \dots \leftarrow w_k \to \dots \to w_n \to b$ in G with $w_i \in W$ for all $i = 1, \dots, n$ (Verma, 1993). An example of a DAG G and its latent projection G^{p} is given in Figure 3.



(a) DAG G

(b) Latent projection G^{p}

Figure 3: DAG G and latent projection G^{p} of G onto $\{A, B, C\}$.

The definition of d-separation for ADMGs (also known as *m-separation* (Richardson, 2003)) employs an extended notion of a collider: given ADMG G^{p} with path $\pi = a \ast\!\!\to\!\!\ast \dots \ast\!\!\to\!\!\ast\, b$, 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 $\pi = a \ast\!\!\to\!\!\ast \dots \ast\!\!\to\!\!\ast\, b$ 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 $A \perp_G^d B \,|\, C$ if and only if $A \perp_{G^{\mathrm{p}}}^d B \,|\, C$ (Verma, 1993); as a corollary the Markov property (2) also holds for the latent projection G^{p} of Bayesian networks with latent variables.

The question that we consider is whether Bayesian networks with latent variables are typically faithful to their latent projection. More formally, let G be a DAG over $V \cup W$, and let G^{p} be the latent projection of G onto V. Recall the distribution map $D : \mathrm{BN}_G \to M_G$ from Definition 6 and define the *observational distribution map*

$$D^{\mathrm{p}} : \mathrm{BN}_G \to M_{G^{\mathrm{p}}}, \quad D^{\mathrm{p}} := \pi_V \circ D$$

where $\pi_V : M_G \to M_{G^{\mathrm{p}}}$ is the projection $\mathbb{P}(X_{V \cup W}) \mapsto \mathbb{P}(X_V)$. Write $M_{G^{\mathrm{p}}}, U_{G^{\mathrm{p}}}, F_{G^{\mathrm{p}}}$ for the distributions over $\mathcal{X}_V$ that are Markov, unfaithful and faithful with respect to the ADMG G^{p} respectively, defined similarly as in (3–4).[9] We call a Bayesian network $m \in \mathrm{BN}_G$ *unfaithful with respect to the ADMG* G^{p} if $D^{\mathrm{p}}(m) \in U_{G^{\mathrm{p}}}$. The core observation for extending results of Section 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 Bayesian network in BN_G that is unfaithful with respect to G^{p} is also unfaithful with respect to G.*

Proof. If the Bayesian network $m \in \mathrm{BN}_G$ is unfaithful, i.e. $D^{\mathrm{p}}(m) \in U_{G^{\mathrm{p}}}$, then there are $A, B, C \subseteq V$ such that $A \not\perp^d_{G^{\mathrm{p}}} B \mid C$ and $X_A \perp\!\!\!\perp_{D^{\mathrm{p}}(m)} X_B \mid X_C$. Since $\pi_{A \cup B \cup C} \circ \pi_V = \pi_{A \cup B \cup C}$ we have $\pi_{A \cup B \cup C}(D^{\mathrm{p}}(m)) = \pi_{A \cup B \cup C}(D(m))$, so we have $X_A \perp\!\!\!\perp_{D(m)} X_B \mid X_C$ as well. As the latent projection preserves d-separations we have $A \not\perp^d_G B \mid C$, so the Bayesian network m is unfaithful with respect to U_G, i.e. $D(m) \in U_G$. ■

Now, we can extend Theorem 6 to ADMGs as follows:

Theorem 9. *Given ADMG G^{p} with vertices V, for any DAG G with vertices $V \cup W$ such that G^{p} is the latent projection of G onto V, the subset $(D^{\mathrm{p}})^{-1}(U_{G^{\mathrm{p}}}) \subseteq \mathrm{BN}_G$ of the Bayesian networks that are unfaithful with respect to G^{p} are nowhere dense.*

Proof. We have from Theorem 6 that $D^{-1}(U_G)$ is nowhere dense in BN_G. By Lemma 4 we have $(D^{\mathrm{p}})^{-1}(U_{G^{\mathrm{p}}}) \subseteq D^{-1}(U_G)$ and hence $(D^{\mathrm{p}})^{-1}(U_{G^{\mathrm{p}}})$ is nowhere dense as well. ■

For any parametrisation $\varphi : \Theta \to \mathrm{BN}_G$ define $T^{\mathrm{p}} := D^{\mathrm{p}} \circ \varphi$; it is immediate that the parameters $(T^{\mathrm{p}})^{-1}(U_{G^{\mathrm{p}}})$ unfaithful to the latent projection G^{p} are a subset of the parameters $T^{-1}(U_G)$ unfaithful to G. Of particular interest is the following result:

Theorem 10. *The set of parameters of linear Gaussian or discrete Bayesian networks with latent variables that are unfaithful to latent projection G^{p}, is nowhere dense and measure-zero.*

Proof. By Theorems 1, 2, 7 and 8 we have for both $\Theta_{\mathcal{N}}$ and $\Theta_{\mathcal{D}}$ that $T^{-1}(U_G)$ is nowhere dense and measure-zero. By Lemma 4 we have $(D^{\mathrm{p}})^{-1}(U_{G^{\mathrm{p}}}) \subseteq D^{-1}(U_G)$ and by pre-composing D and D^{p} with φ we get $(T^{\mathrm{p}})^{-1}(U_{G^{\mathrm{p}}}) \subseteq T^{-1}(U_G)$, so the parameters unfaithful with respect to G^{p} are nowhere dense and measure-zero. ■

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 factors. 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,

[9] Note that for fixed G and $\mathcal{X}_W$ the map D^{p} is not necessarily surjective, contrary to the map D.

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 interventional distributions but the same observational distribution.

Typicality of faithful distributions in any sense might still be 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 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.

References

Baba, K., Shibata, R., and Sibuya, M. (2004). Partial Correlation and Conditional Correlation as Measures of Conditional Independence. *Australian & New Zealand Journal of Statistics*, 46(4):657–664.

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.

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.

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

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

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.

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.