On the Intersection Property of Conditional Independence and its Application to Causal Discovery

Author(s):
Peters, Jonas

Publication Date:
2015-03

Permanent Link:

https://doi.org/10.3929/ethz-b-000383354 →

Originally published in:

Journal of Causal Inference 3(1), http://doi.org/10.1515/jci-2014-0015 →

Jonas Peters*

On the Intersection Property of Conditional Independence and its Application to Causal Discovery

Abstract: This work investigates the intersection property of conditional independence. It states that for random variables A, B, C and X we have that $X \perp\!\!\!\perp A \mid B, C$ and $X \perp\!\!\!\perp B \mid A, C$ implies $X \perp\!\!\!\perp (A, B) \mid C$. Here, "$\perp\!\!\!\perp$" stands for statistical independence. Under the assumption that the joint distribution has a density that is continuous in A, B and C, we provide necessary and sufficient conditions under which the intersection property holds. The result has direct applications to causal inference: it leads to strictly weaker conditions under which the graphical structure becomes identifiable from the joint distribution of an additive noise model.

Keywords: probability theory, causal discovery, graphical models

DOI 10.1515/jci-2014-0015

1 Introduction

This paper investigates the intersection property of conditional independence. For continuous random variables A, B, C and X this property states that $X \perp\!\!\!\perp A \mid B, C$ and $X \perp\!\!\!\perp B \mid A, C$ implies $X \perp\!\!\!\perp (A, B) \mid C$. Here, "$\perp\!\!\!\perp$" stands for statistical independence and "$\not\!\perp\!\!\!\perp$" for statistical dependence (see Section 1.2 for precise definitions). The intersection property does not necessarily hold if the joint distribution does not have a density (e.g. Dawid [1]). Dawid [2] provides measure-theoretic necessary and sufficient conditions for the intersection property. In this work we assume the existence of a density (A0), see below.

It is well known that the intersection property holds if the joint distribution has a strictly positive density (e.g. Pearl [3], 1.1.5). Proposition 1 shows that if the density is not strictly positive, a weaker condition than the intersection property still holds. Corollary 1 states necessary and sufficient conditions for the intersection property. The result about strictly positive densities is contained as a special case. Drton et al. ([4], exercise 6.6) and Fink [5] develop analogous results for the discrete case.

In the remainder of this introduction we discuss the paper's main contribution (Section 1.1) and introduce the required notation (Section 1.2).

1.1 Main contributions

In Section 3 we provide a sufficient and necessary condition on the density for the intersection property to hold (Corollary 1). This result is of interest in itself since the developed condition is weaker than strict positivity.

Studying the intersection property has direct applications to causal inference. Inferring causal relationships is a major challenge in science. In the last decades considerable effort has been made in order to learn causal statements from observational data. As a first step, causal discovery methods therefore estimate graphs from observational data and attach a causal meaning to these graphs (the terminology of causal inference is introduced in Section 4.1). Some causal discovery methods based on structural equation models

***Corresponding author: Jonas Peters,** ETH Zürich, Switzerland, E-mail: peters@stat.math.ethz.ch

(SEMs) require the intersection property for identification; they therefore rely on the strict positivity of the density. This is satisfied if the noise variables have full support, for example. Using the new characterization of the intersection property we can now replace the condition of strict positivity. In fact, we show in Section 4 that noise variables with a path-connected support are sufficient for identifiability of the graph (Proposition 3). This is already known for linear SEMs [6] but not for non-linear models. As an alternative, we provide a condition that excludes a specific kind of constant functions and leads to identifiability, too (Proposition 4).

In Section 2, we provide an example of an SEM that violates the intersection property. Its corresponding graph is not identifiable from the joint distribution. In correspondence to the theoretical results of this work, some noise densities in the example do not have a path-connected support and the functions are partially constant. We are not aware of any causal discovery method that is able to infer the correct graph or the correct Markov equivalence class; the example therefore shows current limits of causal inference techniques. It is non-generic in the case that it violates all sufficient assumptions mentioned in Section 4.

All proofs are provided in Appendix A.

1.2 Conditional independence and the intersection property

We now formally introduce the concept of conditional independence in the presence of densities and the intersection property. Let therefore A, B, C and X be (possibly multi-dimensional) random variables that take values in metric spaces $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ and $\mathcal{X}$, respectively. We first introduce assumptions regarding the existence of a density and some of its properties that appear in different parts of this paper.

(A0) The distribution is absolutely continuous with respect to a product measure of a metric space. We denote the density by $p(\cdot)$. This can be a probability mass function or a probability density function, for example.

(A1) The density $(a,b,c) \mapsto p(a,b,c)$ is continuous. If there is no variable C (or C is deterministic), then $(a,b) \mapsto p(a,b)$ is continuous.

(A2) For each c with $p(c) > 0$ the set $\mathrm{supp}_c(A,B) := \{(a,b) : p(a,b,c) > 0\}$ contains only one path-connected component (see Section 3).

(A2′) The density $p(\cdot)$ is strictly positive.

Condition (A2′) implies (A2). We assume (A0) throughout the whole work.

In this paper we work with the following definition of conditional independence.

Definition 1 (Conditional (In)dependence). *We call X independent of A conditional on B and write $X \perp\!\!\!\perp A \,|\, B$ if and only if*

$$p(x, a \,|\, b) = p(x \,|\, b) p(a \,|\, b) \tag{1}$$

for all x, a, b such that $p(b) > 0$. Otherwise, X and A are dependent conditional on B and we write $X \not\perp\!\!\!\perp A \,|\, B$.

The intersection property of conditional independence is defined as follows (e.g. Pearl [3], 1.1.5).

Definition 2 (Intersection Property). *We say that the joint distribution of X, A, B, C satisfies the intersection property if*

$$X \perp\!\!\!\perp A \,|\, B, C \text{ and } X \perp\!\!\!\perp B \,|\, A, C \;\Rightarrow\; X \perp\!\!\!\perp (A,B) \,|\, C \,. \tag{2}$$

The intersection property (2) has been proven to hold for strictly positive densities (e.g. Pearl [3], 1.1.5). The other direction “$\Leftarrow$” is known as the “weak union” of conditional independence [3].

2 Counterexample

We now give an example of a distribution that does not satisfy the intersection property (2). Since the joint distribution has a continuous density, the example shows that the intersection property requires further restrictions on the density apart from its existence. We will later use the same idea to prove Proposition 2 that shows the necessity of our new condition.

Example 1. *Consider a so-called additive noise model (ANM; see Section 4.1) for random variables* X, A, B:

$$\begin{aligned} A &= N_A\,, \\ B &= A + N_B\,, \\ X &= f(B) + N_X\,, \end{aligned} \tag{3}$$

where N_A, N_B, N_X *are jointly independent, have continuous densities and satisfy* $\mathrm{supp}(N_A) := \{n\ :\ p_{N_A}(n) > 0\} = (-2;-1) \cup (1;2)$ *and* $\mathrm{supp}(N_B) = \mathrm{supp}(N_X) = (-0.3;0.3)$. *Let the function f be of the form*

$$f(b) = \begin{cases} +10 & \text{if } b > 0.5\,, \\ 0 & \text{if } b < -0.5\,, \\ g(b) & \quad \text{else}, \end{cases} \tag{4}$$

where the function g can be chosen to make f arbitrarily smooth. Some parts of this structural equation model (SEM) are summarized in Figure 1. The distribution satisfies $X \perp\!\!\!\perp A \,|\, B$ *and* $X \perp\!\!\!\perp B \,|\, A$ *but* $X \not\!\perp\!\!\!\perp A$ *and* $X \not\!\perp\!\!\!\perp B$. *The (intuitive) reason for this as follows: we see* $X \perp\!\!\!\perp A \,|\, B$ *from eq. (3). Further, if we know that A (or B) is positive, X has to take values close to ten and thus* $X \not\!\perp\!\!\!\perp A$ ($X \not\!\perp\!\!\!\perp B$)*; but when knowing that A is positive, the knowledge of B does not provide any additional information about X* ($X \perp\!\!\!\perp B \,|\, A$)*. This means that the intersection property is violated. A formal proof is provided in the more general setting of Proposition 2. Within each component, however, that is if we consider the areas* $A, B > 0$ *and* $A, B < 0$ *separately, we do have the independence statement* $X \perp\!\!\!\perp (A, B)$*; therefore the intersection property holds "locally". This observation will be formalized as the weak intersection property in Proposition 1.*

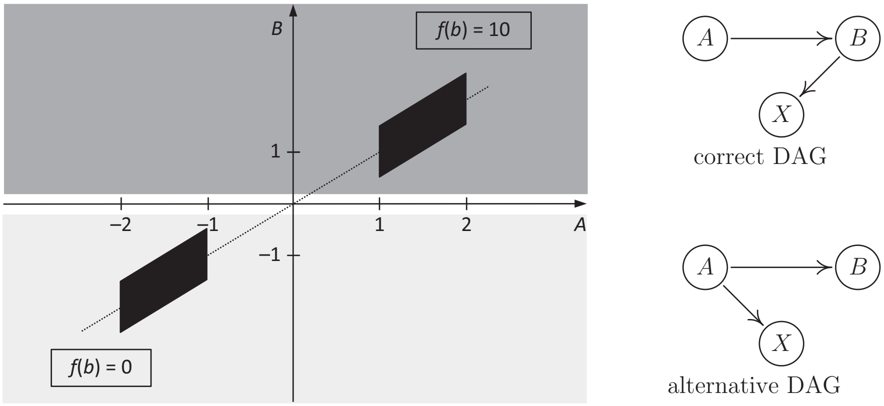

Figure 1: Example 1. The plot on the left-hand side shows the support of variables *A* and *B* in black. The function *f* takes values ten and zero in the areas filled with dark grey and light grey, respectively. The ANM (3) corresponds to the top graph on the right-hand side but the distribution can also be generated by an ANM with the bottom graph, this is explained in Remark 1.

It will turn out to be important that the two path-connected components of the support of *A* and *B* cannot be connected by an axis-parallel line. This motivates the notation introduced in Section 3. Remark 1 in Section 4 discusses the causal interpretation of Example 1.

3 Necessary and sufficient condition for the intersection property

This section characterizes the intersection property in terms of the joint density over the corresponding random variables. In particular, we state a weak intersection property (Proposition 1) that leads to a necessary and sufficient condition for the classical intersection property, see Corollary 1.

We will see that the intersection property fails in Example 1 because of the two "separated" components in Figure 1. In order to formulate our results we first require the notion of path-connectedness. A continuous mapping $\lambda : [0,1] \to \mathcal{X}$ into a metric space $\mathcal{X}$ is called a *path* between $\lambda(0)$ and $\lambda(1)$ in $\mathcal{X}$. A subset $\mathcal{S} \subseteq \mathcal{X}$ is called *path-connected* if every pair of points in $\mathcal{S}$ can be connected by a path in $\mathcal{S}$. We can always decompose $\mathcal{X}$ into its (disjoint) *path-connected components*.[1] The following definition provides a formalization of the intuition that the two components in Figure 1 are "separated".

Definition 3. *(i) For each c with $p(c) > 0$ we define the (not necessarily closed) support of A and B as*

$$\mathrm{supp}_c(A,B) := \{(a,b) \;:\; p(a,b,c) > 0\}\,.$$

We further write for all sets $M \subseteq \mathcal{A} \times \mathcal{B}$

$$\mathrm{proj}_A(M) := \{a \in \mathcal{A} \;:\; \exists b \text{ with } (a,b) \in M\} \text{ and}$$

$$\mathrm{proj}_B(M) := \{b \in \mathcal{B} \;:\; \exists a \text{ with } (a,b) \in M\}\,.$$

(ii) We denote the path-connected components of $\mathrm{supp}_c(A,B)$ by Z_i^c, $i \in I_Z^c$, with some index set I_Z^c. Two path-connected components $Z_{i_1}^c$ and $Z_{i_2}^c$ are said to be coordinate-wise connected if

$$\mathrm{proj}_A(Z_{i_1}^c) \cap \mathrm{proj}_A(Z_{i_2}^c) \neq \emptyset \text{ or}$$

$$\mathrm{proj}_B(Z_{i_1}^c) \cap \mathrm{proj}_B(Z_{i_2}^c) \neq \emptyset\,.$$

(The intuition is that we can draw an axis-parallel line from $Z_{i_1}^c$ to $Z_{i_2}^c$.) We then say that Z_i^c and Z_j^c are equivalent if and only if there exists a sequence $Z_i^c = Z_{i_1}^c, \dots, Z_{i_m}^c = Z_j^c$ with all neighbours $Z_{i_k}^c$ and $Z_{i_{k+1}}^c (k = 1, \dots, m-1)$ being coordinate-wise connected. We represent the equivalence classes by the union of all its members. These unions we denote by U_i^c, $i \in I_U^c$.
We further introduce a deterministic function U^c of the variables A and B. We set

$$U^c := u^c(A,B) := \begin{cases} i & \text{if } (A,B) \in U_i^c \\ 0 & \text{if } p(A,B,c) = 0 \end{cases}.$$

We have that $U^c = i$ if and only if $A \in \mathrm{proj}_A(U_i^c)$ if and only if $B \in \mathrm{proj}_B(U_i^c)$. Furthermore, the projections $\mathrm{proj}_A(U_i^c)$ are disjoint for different i; the same holds for $\mathrm{proj}_B(U_i^c)$.

(iii) The case where there is no variable C can be treated as if C was deterministic: $p(c) = 1$ for some c.

In Example 1 there is no variable C. Figure 1 shows the support $\mathrm{supp}_c(A,B)$ in black. It contains two path-connected components Z_1^c and Z_2^c. Since they cannot be connected by axis-parallel lines, they are not equivalent; thus, one of them corresponds to the equivalence class U_1^c and the other to U_2^c. Figure 2 shows another example that contains three equivalence classes of path-connected components; again, there is no variable C; we formally introduce a deterministic variable C that always takes the value c.

[1] Formally, path-connected components are equivalence classes of points, where two points are equivalent if there exists a path in $\mathcal{X}$ connecting them. This equivalence should not be confused with the equivalence appearing in Definition 3(ii).

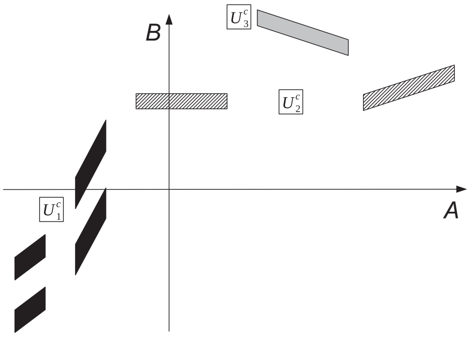

Figure 2: Each block represents one path-connected component Z_i^c of the support of $p(a,b)$. All blocks with the same filling are equivalent since they can be connected by axis-parallel lines (see Definition 3). There are three different fillings corresponding to the equivalence classes U_1^c, U_2^c and U_3^c.

Using Definition 3 we are now able to state the two main results, Propositions 1 and 2. As a direct consequence we obtain Corollary 1 which generalizes the condition of strictly positive densities.

Proposition 1 (Weak Intersection Property). *Assume (A0), (A1) and that $X \perp\!\!\!\perp A \mid B,C$ and $X \perp\!\!\!\perp B \mid A,C$. Consider now c with $p(c)>0$ and the variable U^c as defined in Definition 3(ii). We then have the weak intersection property*:

$$X \perp\!\!\!\perp (A,B) \mid C=c, U^c.$$

This means that

$$p(x \mid a,b,c) = p(x \mid c, u^c(a,b)) \tag{5}$$

for all x,a,b with $p(a,b,c)>0$. The values of A,B do not provide additional information if we already know $U^c = u^c(A,B)$.

We call this property the *weak* intersection property for the following reason: if $X \perp\!\!\!\perp (A,B) \mid C$, then by definition $u^c(a,b)=i$ if and only if $(a,b) \in U_i^c$ and therefore

$$p(x \mid a,b,c) = p(x \mid c) = p\Big(x \mid c, (a,b) \in U^c_{u^c(a,b)}\Big) = p(x \mid c, u^c(a,b)).$$

In this sense, eq. (5) is strictly weaker than $X \perp\!\!\!\perp (A,B) \mid C$.

Furthermore, Proposition 1 includes the intersection property for positive densities as a special case. If the density is indeed strictly positive, then there is only a single path-connected component Z_1^c and a single equivalence class U_1^c. Therefore, U^c is constant and it follows from eq. (5) and Lemma 1 (see "Proof of Proposition 1" in Appendix A) that $X \perp\!\!\!\perp (A,B) \mid C$.

Proposition 2 (Failure of Intersection Property). *Assume (A0), (A1) and that there exist two different sets $U_1^{c^*} \neq U_2^{c^*}$ for some c^* with $p(c^*)>0$. Then there exists a random variable X such that the intersection property (2) does not hold for the joint distribution of X,A,B,C.*

As a direct corollary from these two propositions we obtain a characterization of the intersection property in the case of continuous densities.

Corollary 1 (Intersection Property). *Assume (A0) and (A1).*
The intersection property (2) *holds for all variables X if and only if all components Z_i^c are equivalent, i.e. there is only one set U_1^c.*

In particular, this is the case if (A2) holds (there is only one path-connected component) or (A2') holds (the density is strictly positive).

4 Application to causal discovery

We will first introduce some graph notation that we use for formulating the application to causal inference.

4.1 Notation and prerequisites

Standard graph definitions can be found in Lauritzen [7], Spirtes et al. [8] and many others. We follow the presentation of Section 1.1 in Peters et al. [9]. A **graph** $G = (\mathbf{V}, \mathcal{E})$ contains nodes $\mathbf{V} = \{1, \dots, p\}$ (often identified with random variables $X_1, \dots, X_p$) and edges $\mathcal{E} \subset \mathbf{V} \times \mathbf{V}$ between nodes. A graph $G_1 = (\mathbf{V}_1, \mathcal{E}_1)$ is called a **proper subgraph** of G if $\mathbf{V}_1 = \mathbf{V}$ and $\mathcal{E}_1 \subset \mathcal{E}$ with $\mathcal{E}_1 \neq \mathcal{E}$. A node i is called a **parent** of j if $(i,j) \in \mathcal{E}$ and a **child** if $(j,i) \in \mathcal{E}$. The set of parents of j is denoted by $\mathbf{PA}_j^G$, the set of its children by $\mathbf{CH}_j^G$. Two nodes i and j are **adjacent** if either $(i,j) \in \mathcal{E}$ or $(j,i) \in \mathcal{E}$. We say that there is an undirected edge between two adjacent nodes i and j if $(i,j) \in \mathcal{E}$ and $(j,i) \in \mathcal{E}$. An edge between two adjacent nodes is directed if it is not undirected. We then write $i \to j$ for $(i,j) \in \mathcal{E}$. Three nodes are called a **v-structure** if one node is a child of the two others that themselves are not adjacent. A **path** in G is a sequence of (at least two) distinct vertices $i_1, \dots, i_n$, such that there is an edge between i_k and i_{k+1} for all $k = 1, \dots, n-1$. If $i_k \to i_{k+1}$ for all k we speak of a **directed path** from i_1 to i_n and call i_n a **descendant** of i_1. We denote all descendants of i by $\mathbf{DE}_i^G$ and all non-descendants of i, excluding i, by $\mathbf{ND}_i^G$. In this work, i is neither a descendant nor a non-descendant of itself. G is called a **directed acyclic graph (DAG)**, if all edges are directed and there is no pair of nodes (j, k) such that there are directed paths from j to k and from k to j. In a DAG, a path between i_1 and i_n is **blocked by a set S** (with neither i_1 nor i_n in this set) whenever there is a node i_k, such that one of the following two possibilities hold: (1) $i_k \in \mathbf{S}$ and $i_{k-1} \to i_k \to i_{k+1}$ or $i_{k-1} \leftarrow i_k \leftarrow i_{k+1}$ or $i_{k-1} \leftarrow i_k \to i_{k+1}$, or (2) $i_{k-1} \to i_k \leftarrow i_{k+1}$ and neither i_k nor any of its descendants is in $\mathbf{S}$. We say that two disjoint subsets of vertices $\mathbf{A}$ and $\mathbf{B}$ are ***d*-separated** by a third (also disjoint) subset $\mathbf{S}$ if every path between nodes in $\mathbf{A}$ and $\mathbf{B}$ is blocked by $\mathbf{S}$. A joint distribution is said to be **Markov with respect to the DAG *G*** if $\mathbf{A}, \mathbf{B}$ d-sep. by $\mathbf{C} \Rightarrow \mathbf{A} \perp\!\!\!\perp \mathbf{B} | \mathbf{C}$ for all disjoint sets $\mathbf{A}, \mathbf{B}, \mathbf{C}$. It is said to be **faithful to the DAG *G*** if $\mathbf{A}, \mathbf{B}$ d-sep. by $\mathbf{C} \Leftarrow \mathbf{A} \perp\!\!\!\perp \mathbf{B} | \mathbf{C}$ for all disjoint sets $\mathbf{A}, \mathbf{B}, \mathbf{C}$. Finally, a distribution satisfies **causal minimality** with respect to G if it is Markov with respect to G, but not to any proper subgraph of G.

In order to infer graphs from distributions, one requires assumptions that relate the joint distribution with properties of the graph, which is often assumed to be a DAG. Constraint-based or independence-based methods [3, 8] and some score-based methods [10, 11] assume the Markov condition and faithfulness. These two assumptions make the Markov equivalence class of the correct graph identifiable from the joint distribution, i.e. the skeleton and the v-structures of the graph can be inferred from the joint distribution [12].

Alternatively [6, 9, 13], we can assume an additive noise models (ANMs). In these models, the joint distribution over $X_1, \dots, X_p$ is generated by an SEM

$$X_i = f_i(X_{\mathbf{PA}_i}) + N_i, \quad i = 1, \dots, p, \tag{6}$$

with continuous, non-constant functions f_i, additive and jointly independent noise variables N_i with mean zero and sets $\mathbf{PA}_i$ that are the parents of i in a DAG G. To simplify notation, we have identified variable X_i with its index (or node) i. These models can be shown to satisfy the Markov condition (Pearl [3], theorem 1.4.1); the functions f_i being non-constant correspond to causal minimality (Peters et al. [9], proposition 17), which is strictly weaker than faithfulness. We now define what we mean by identifiability of the DAG in continuous ANMs. Consider a certain class of SEMs and suppose that the distribution $P = P(X_1, \dots, X_p)$ is generated from such an SEM. We say that G is identifiable from P if P cannot be generated by an SEM from the same class but with a different graph $H \neq G$.

Loosely speaking, Peters et al. ([9], theorem 28) prove that

(∗) The identifiability of model classes extends from DAGs with two nodes to DAGs with an arbitrary number of variables.

4.2 Intersection property and causal discovery

We first revisit Example 1 and interpret it from a causal point of view.

Remark 1 (Example 1 continued). *Example 1 has the following important implication for causal inference. The distribution can be generated by two different DAGs, namely $A \rightarrow B \rightarrow X$ and $X \leftarrow A \rightarrow B$, see Figure 1. The SEM (3) corresponds to the former DAG. A slightly modified version of eq. (3) where $X = \tilde{f}(A) + N_X$ replaces the last equation in eq. (3) corresponds to the latter DAG. The distribution satisfies causal minimality with respect to both DAGs. Since it violates faithfulness and the intersection property, we are not aware of any causal inference method that is able to recover the correct graph structure based on observational data only. Recall that Peters et al. [9] assume strictly positive densities in order to assure the intersection property. More precisely, Example 1 shows that lemma 38 in Peters et al. [9], see Appendix B, does not hold anymore when the positivity is violated.*

In order to prove $(*)$, Peters et al. [9] require a strictly positive density. This is because the key results used in the proof is proposition 29 which is proved using lemma 38, which itself relies on the intersection property (proposition 29 and lemma 38 are provided in Appendix B). But since Corollary 1 provides weaker assumption for the intersection property, we are now able to obtain new identifiability results.

Proposition 3. *Assume that a joint distribution over $X_1, \ldots, X_p$ is generated by an ANM (6). Assume further that the noise variables have continuous densities and that the support of each noise variable N_i, $i = 1, \ldots, p$ is path-connected. Then, statement $(*)$ holds.*

Example 1 violates the assumption of Proposition 3 since the support of A is not path-connected. It satisfies another important property, too: the function f is constant on some intervals. The following proposition shows that this is necessary to violate identifiability.

Proposition 4. *Assume that a joint distribution over $X_1, \ldots, X_p$ is generated by an ANM (6) with graph G. Let us denote the non-descendants of X_i by $\mathbf{ND}_i^G$. Assume that the structural equations are non-constant in the following way: for all X_i, for all its parents $X_j \in \mathbf{PA}_i$ and for all $X_{\mathbf{C}} \subseteq \mathbf{ND}_i^G \setminus \{X_j\}$, there are (x_j, x_j', x_k, x_c) such that $f_i(x_j, x_k) \neq f_i(x_j', x_k)$ and $p(x_j, x_k, x_c) > 0$ and $p(x_j', x_k, x_c) > 0$. Here, x_k represents the value of all parents of X_i except X_j. Then for any $\mathbf{PA}_i \setminus \{j\} \subseteq \mathbf{S} \subseteq \mathbf{ND}_i^G \setminus \{j\}$, it holds that $X_i \not\perp\!\!\!\perp X_j \mid \mathbf{S}$. Therefore, statement $(*)$ follows.*

Proposition 4 provides an alternative way to prove identifiability. The results are summarized in Table 1.

Table 1: This table shows conditions for continuous additive noise models (ANMs) that lead to identi_ability of the directed acyclic graph from the joint distributions. Using the characterization of the intersection property we could weaken the condition of a strictly positive density.

Additional assumption on continuous ANMs	Identifiability of graph, see (*)
Noise variables with full support	✓ Peters et al. [9]
Noise variables with path-connected support	✓ Proposition 3
Non-constant functions, see Proposition 4	✓ Proposition 4
None of the above satisfied	✗ Example 1

5 Conclusions

It is possible to prove the intersection property of conditional independence for variables whose distributions do not have a strictly positive density. A necessary and sufficient condition for the intersection property is that all path-connected components of the support of the density are equivalent, that is, they can be connected by axis-parallel lines. In particular, this condition is satisfied for densities whose support is path-connected. In the general case, the intersection property still holds after conditioning on an equivalence class of path-connected components; we call this the weak intersection property. We believe that the assumption of a density that is continuous in A, B and C can be weakened even further.

This insight has a direct application in causal inference (which is rather of theoretical nature than having implications for practical methods). In the context of continuous ANMs, we relax important conditions for identifiability of the graph from the joint distribution. Furthermore, there is some interest in uniform consistency in causal inference. For linear Gaussian SEMs, for example, the PC algorithm [8] exploits conditional independences, that is, vanishing partial correlations. Zhang and Spirtes [14] prove uniform consistency under the assumption that non-vanishing partial correlations cannot be arbitrarily close to zero (this condition is referred to as “strong faithfulness”). Our work suggests that in order to prove uniform consistency for continuous ANMs, one may need to be “bounded away” from Example 1.

Acknowledgements: The author thanks the anonymous reviewers for their insightful and constructive comments. He further thanks Thomas Kahle and Mathias Drton for pointing out the link to algebraic statistics for discrete variables. The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union’s Seventh Framework Programme (FP7/2007-2013) under REA grant agreement no 326496.

Appendix A

Proofs

Proof of Proposition 1

We require the following well-known lemma (e.g. Dawid [15]).

Lemma 1. *We have* $X \perp\!\!\!\perp A \,|\, B$ *if and only if*

$$p(x\,|\,a,b) = p(x\,|\,b)$$

for all x, a, b *such that* $p(a,b) > 0$.

Proof. (of Proposition 1) To simplify notation we write $u^c := u^c(a,b)$. We have by Lemma 1

$$p(x\,|\,b,c) = p(x\,|\,a,b,c) = p(x\,|\,a,c) \tag{7}$$

for all x, a, b, c with $p(a,b,c) > 0$. As the main argument we show that

$$p(x\,|\,b,c) = p(x\,|\,\tilde{b},c) \tag{8}$$

for all $x, b, \tilde{b}, c$ with $b, \tilde{b} \in \mathrm{proj}_B(U_i^c)$ for the same i.

Step 1, we prove eq. (8) for $b, \tilde{b} \in Z_i^c$. We first show that there is a path $\lambda : t \mapsto (a(t), b(t))$, such that $p(a(t), b(t), c) > 0$ for all $0 \leq t \leq 1$, and $b(0) = b$ and $b(1) = \tilde{b}$. Since the interval $[0,1]$ is compact and λ is continuous, the path $\{(a(t), b(t)) : 0 \leq t \leq 1\}$ is compact, too (for notational simplicity we identify the path λ with its image). Define for each point $(a(t), b(t))$ on the path an open ball with radius small enough such that all (a, b) in the ball satisfy $p(a, b, c) > 0$ (this is possible because $(a, b, c) \mapsto p(a, b, c)$ is assumed to be continuous). Because these balls are path-connected, they also lie in Z_i^c. They form an open cover of the path $\{(a(t), b(t)) : 0 \leq t \leq 1\}$, and we can thus choose a finite subset of balls, of size n say, that still provides an open cover of the path. Without loss of generality let $(a(0), b(0))$ be the centre of ball 1 and $(a(1), b(1))$ be the centre of ball n. It suffices to show that eq. (8) holds for the centres of two neighbouring balls, say (a_1, b_1) and (a_2, b_2). Choose a point (a^*, b^*) from the non-empty intersection of those two balls. Since $d((a_1, b_1), (a^*, b_1)) < d((a_1, b_1), (a^*, b^*))$ and $d((a_2, b_2), (a_2, b^*)) < d((a_2, b_2), (a^*, b^*))$ for the Euclidean metric d, we have that $p(a_1, b_1, c)$, $p(a^*, b_1, c)$, $p(a^*, b^*, c)$, $p(a_2, b^*, c)$ and $p(a_2, b_2, c)$ are all greater than zero. Therefore, using eq. (7) several times,

$$\begin{aligned} p(x\,|\,b_1, c) &= p(x\,|\,a_1, c) = p(x\,|\,a^*, c) \\ &= p(x\,|\,b^*, c) = p(x\,|\,a_2, c) = p(x\,|\,b_2, c) \end{aligned}$$

This shows eq. (8) for $b, \tilde{b} \in Z_i^c$.

Step 2, we prove eq. (8) for $b \in Z_i^c$ and $\tilde{b} \in Z_{i+1}^c$, where Z_i^c and Z_{i+1}^c are coordinate-wise connected (and thus equivalent). If $b^* \in \mathrm{proj}_B(Z_i^c) \cap \mathrm{proj}_B(Z_{i+1}^c)$, we know that

$$p(x\,|\,b, c) = p(x\,|\,b^*, c) = p(x\,|\,\tilde{b}, c)$$

from the argument given in step 1. If $a^* \in \mathrm{proj}_A(Z_i^c) \cap \mathrm{proj}_A(Z_{i+1}^c)$, then there is a b_i, b_{i+1} such that $(a^*, b_i) \in Z_i^c$ and $(a^*, b_{i+1}) \in Z_{i+1}^c$. By eq. (7) and the argument from step 1 we have

$$p(x\,|\,b, c) = p(x\,|\,b_i, c) = p(x\,|\,b_{i+1}, c) = p(x\,|\,\tilde{b}, c).$$

We can now combine these two steps in order to prove the original claim from eq. (8). If $b, \tilde{b} \in \mathrm{proj}_B(U_i^c)$ then $b \in \mathrm{proj}_B(Z_1^c)$ and $\tilde{b} \in \mathrm{proj}_B(Z_n^c)$, say. Further, there is a sequence $Z_1^c, \ldots, Z_n^c$ with Z_k^c and Z_{k+1}^c being coordinate-wise connected for $k = 1, \ldots, n-1$. Combining steps 1 and 2 proves eq. (8).

Consider now x, b, c such that $p(b, c) > 0$ (which implies $p(c) > 0$) and consider $u^c = i$, say. Observe further that $p(a, c) > 0$ for $a \in \mathrm{proj}_A(U_i^c)$. We thus have

$$\begin{aligned} p(x, u^c\,|\,c) &= \int_a p(x, a, u^c\,|\,c)\; da = \int_{a \in \mathrm{proj}_A(U_i^c)} p(x, a\,|\,c)\; da \\ &= \int_{a \in \mathrm{proj}_A(U_i^c)} \frac{p(x, a, c) p(a, c)}{p(c) p(a, c)}\; da \\ &= \int_{a \in \mathrm{proj}_A(U_i^c)} p(x\,|\,a, c) p(a\,|\,c)\; da \\ &= \int_{a \in \mathrm{proj}_A(U_i^c), p(a,b,c) > 0} p(x\,|\,a, c) p(a\,|\,c)\; da \\ &\qquad + \int_{a \in \mathrm{proj}_A(U_i^c), p(a,b,c) = 0} p(x\,|\,a, c) p(a\,|\,c)\; da \\ &\overset{(7)}{=} p(x\,|\,b, c) \int_{a \in \mathrm{proj}_A(U_i^c), p(a,b,c) > 0} p(a\,|\,c)\; da + \int_{\mathcal{A}_b} p(x\,|\,a, c) p(a\,|\,c)\; da \\ &=: (\#) \end{aligned}$$

with $\mathcal{A}_b = \{a \in \mathrm{proj}_A(U_i^c) : p(a, b, c) = 0\}$. It is the case, however, that for all $a \in \mathcal{A}_b$ there is a $\tilde{b}(a) \in \mathrm{proj}_B(U_i^c)$ with $p(a, \tilde{b}(a), c) > 0$. But since also $b \in \mathrm{proj}_B(U_i^c)$ we have $p(x\,|\,\tilde{b}, c) = p(x\,|\,b, c)$ by eq. (8). Ergo,

$$
\begin{aligned}
(\#) &= p(x\,|\,b,c)\int_{a\in \mathrm{proj}_A(U_i^c),p(a,b,c)>0} p(a\,|\,c)\;da + \int_{\mathcal{A}_b} p(x\,|\,a,\tilde{b}(a),c)p(a\,|\,c)\;da\\
&= p(x\,|\,b,c)\int_{a\in \mathrm{proj}_A(U_i^c),p(a,b,c)>0} p(a\,|\,c)\;da + p(x\,|\,b,c)\int_{\mathcal{A}_b} p(a\,|\,c)\;da\\
&= p(x\,|\,b,c)\int_{a\in \mathrm{proj}_A(U_i^c)} p(a\,|\,c)\;da\\
&= p(x\,|\,b,c)\;p(u^c\,|\,c)
\end{aligned}
$$

This implies

$$p(x\,|\,c,u^c) = p(x\,|\,b,c)\;.$$

Together with eq. (7) this leads to

$$p(x\,|\,a,b,c,u^c) = p(x\,|\,a,b,c) = p(x\,|\,c,u^c)\,.$$

□

Proof of Proposition 2

Proof. Define X according to

$$X = g(C, U^C) + N_X,$$

where $N_X \sim \mathcal{U}([-0.1, 0.1])$ is uniformly distributed with N_X independent of (A, B, C). Define g according to

$$g(c,u^c) = \begin{cases} 10 & \text{if } C = c^* \text{ and } u^{c^*} = 1\\ 0 & \text{otherwise}\end{cases}$$

Fix a value c with $p(c) > 0$. We then have for all a, b with $p(a,b,c) > 0$ that

$$p(x\,|\,a,b,c) = p(x\,|\,c,u^c) = p(x\,|\,a,c) = p(x\,|\,b,c)$$

because U^c can be written as a function of A or of B. We therefore have that $X \perp\!\!\!\perp A\,|\,B, C$ and $X \perp\!\!\!\perp B\,|\,A, C$. Depending on whether b is in $\mathrm{proj}_B(U_1^{c^*})$ or not we have $p(x = 0\,|\,b, c^*) = 0$ or $p(x = 10\,|\,b, c^*) = 0$, respectively. Thus,

$$p(x = 10\,|\,b, c^*) \cdot p(x = 0\,|\,b, c^*) = 0\,, \text{whereas}$$

$$p(x = 10\,|\,c^*) \cdot p(x = 0\,|\,c^*) \neq 0\,.$$

This shows that $X \not\perp\!\!\!\perp B\,|\,C = c^*$. Note that $(x, a, b, c) \mapsto p(x, a, b, c)$ is not necessarily continuous, see (A1). □

Proof of Proposition 3

Proof. Since the true structure corresponds to a DAG, we can find a causal ordering, i.e. a permutation $\pi : \{1, \ldots, p\} \to \{1, \ldots, p\}$ such that

$$\mathbf{PA}_{\pi(i)} \subseteq \{\pi(1), \ldots, \pi(i-1)\}\,.$$

In this ordering, $\pi(1)$ is a source node and $\pi(p)$ is a sink node. We can then rewrite the structural equation model in eq. (6) as

$$X_{\pi(i)} = \tilde{f}_{\pi(i)}\big(X_{\pi(1)}, \ldots, X_{\pi(i-1)}\big) + N_{\pi(i)}\,,$$

where the functions $\tilde{f}_i$ are the same as f_i except they are constant in the additional input arguments.

The density of the random vector $(X_1, \dots, X_p)$ has path-connected support by the following argument: consider a one-dimensional random variable N with mean zero and a (possibly multivariate) random vector X both with path-connected support and a continuous function f. Then, the support of the random vector $(X, f(X) + N)$ is path-connected, too. Indeed, consider two points (x_0, y_0) and (x_1, y_1) from the support of $(X, f(X) + N)$. The path can then be constructed by concatenating three sub-paths: (1) the path between (x_0, y_0) and $(x_0, f(x_0))$ (N's support is path-connected), (2) the path between $(x_0, f(x_0))$ and $(x_1, f(x_1))$ on the graph of f (which is path-connected due to the continuity of f) and (3) the path between $(x_1, f(x_1))$ and (x_1, y_1), analogously to (1).

Therefore, the intersection property (2) holds for any disjoint sets of variables $X, A, B, C \in \{X_1, \dots, X_p\}$ by Proposition 1. Thus, the statements of lemma 38 and thus proposition 29 from Peters et al. [9] remain correct, which proves $(*)$ for noise variables with continuous densities and path-connected support. □

Proof of Proposition 4

Proof. The proof is immediate. Since $p(x_i | x_j, x_k, x_c) \neq p(x_i | x_j', x_k, x_c)$ (the means are not the same) the statement follows from Lemma 1.

In this case, lemma 38 might not hold but more importantly proposition 29 does (both from Peters et al. [9]. This proves $(*)$. □

Appendix B

Technical results for identifiability in additive noise models

We provide the two key results required for proving property $(*)$ in Section 4.1. The intersection property is used to prove the “only if” part of lemma 38, which itself is used to prove proposition 29.

Lemma 38 [**9**] Consider the random vector $\mathbf{X}$ and assume that the joint distribution has a (strictly) positive density. Then the joint distribution over $\mathbf{X}$ satisfies causal minimality with respect to a DAG G if and only if $\forall B \in \mathbf{X}$ $\forall A \in \mathrm{PA}_B^G$ and $\forall \mathbf{S} \subset \mathbf{X}$ with $\mathbf{PA}_B^G \backslash \{A\} \subseteq \mathbf{S} \subseteq \mathbf{ND}_B^G \backslash \{A\}$ we have that

$$B \not\perp\!\!\!\perp A \,|\, \mathbf{S}.$$

Proposition 29 [9] Let G and G' be two different DAGs over variables $\mathbf{X}$. Assume that the joint distribution over $\mathbf{X}$ has a strictly positive density and satisfies the Markov condition and causal minimality with respect to G and G'. Then there are variables $L, Y \in \mathbf{X}$ such that for the sets $\mathbf{Q} := \mathbf{PA}_L^G \backslash \{Y\}$, $\mathbf{R} := \mathbf{PA}_Y^{G'} \backslash \{L\}$ and $\mathbf{S} := \mathbf{Q} \cup \mathbf{R}$ we have

- $Y \to L$ in G and $L \to Y$ in G'
- $\mathbf{S} \subseteq \mathbf{ND}_L^G \backslash \{Y\}$ and $\mathbf{S} \subseteq \mathbf{ND}_Y^{G'} \backslash \{L\}$

References

1. Dawid AP. Some misleading arguments involving conditional independence. J R Stat Soc Ser B 1979;41:249–52.
2. Dawid AP. Conditional independence for statistical operations. Ann Stat 1980;8:598–617.
3. Pearl J. Causality: models, reasoning, and inference, 2nd ed. New York, NY: Cambridge University Press, 2009.

4. Drton M, Sturmfels B, Sullivant S. Lectures on algebraic statistics. Volume 39 of *Oberwolfach Seminars*. Basel: Birkhäuser Verlag, 2009.
5. Fink A. The binomial ideal of the intersection axiom for conditional probabilities. J Algebraic Combinatorics 2011;33:455–63.
6. Shimizu S, Hoyer PO, Hyvärinen A, Kerminen AJ. A linear non-Gaussian acyclic model for causal discovery. J Mach Learn Res 2006;7:2003–30.
7. Lauritzen S. Graphical models. New York, NY: Oxford University Press, 1996.
8. Spirtes P, Glymour C, Scheines R. Causation, prediction, and search, 2nd ed. Cambridge, MA: MIT Press, 2000.
9. Peters J, Mooij JM, Janzing D, Schölkopf B. Causal discovery with continuous additive noise models. J Mach Learn Res 2014;15:2009–53.
10. Chickering DM. Optimal structure identification with greedy search. J Mach Learn Res 2002;3:507–54.
11. Heckerman D, Meek C, Cooper G. A Bayesian approach to causal discovery. In Glymour C, Cooper G, editors. Computation, causation, and discovery. Cambridge, MA: MIT Press, 1999:141–65.
12. Verma T, Pearl J. Equivalence and synthesis of causal models. In: P.B. Bonissone and M. Henrion and L.N. Kanal and J.F. Lemmer editors. Proceedings of the 6th annual conference on uncertainty in artificial intelligence (UAI), San Francisco, CA: Morgan Kaufmann, 1991:255–70.
13. Hoyer PO, Janzing D, Mooij JM, Peters J, Schölkopf B. Nonlinear causal discovery with additive noise models. In: D. Koller and D. Schuurmans and Y. Bengio and L. Bottou editors. Advances in neural information processing systems 21 (NIPS), Red Hook, NY: Curran Associates, Inc., 2009:689–696.
14. Zhang J, Spirtes P. Strong faithfulness and uniform consistency in causal inference. In: C. Meek and U. Kjærulff editors. Proceedings of the 19th annual conference on uncertainty in artificial intelligence (UAI), San Francisco, CA: Morgan Kaufmann, 2003:632–9.
15. Dawid AP. Conditional independence in statistical theory. J R Stat Soc Ser B 1979;41:1–31.