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

Jonas Peters

peters@stat.math.ethz.ch

ETH Zurich

March 5, 2014

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 continuous density, 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.

1 Introduction

1.1 Application 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. Causal discovery methods make assumptions that relate the joint distribution with properties of the causal graph. Constraint-based or independence-based methods [Pearl, 2009, Spirtes et al., 2000] and some score-based methods [Chickering, 2002, Heckerman et al., 1999] assume the Markov condition and faithfulness. A distribution is said to be Markov with respect to a directed acyclic graph (DAG) G if each d-separation in the graph implies the corresponding (conditional) independence; the distribution is faithful with respect to G if the reverse statement holds. These

1

two assumptions render 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 [Verma and Pearl, 1991]. Methods like LiNGAM [Shimizu et al., 2006] or additive noise models [Hoyer et al., 2009, Peters et al., 2013] assume the Markov condition, too, but do not require faithfulness; instead, these methods assume that the structural equations come from a restricted model class (e.g. linear with non-Gaussian noise or non-linear with additive Gaussian noise). In order to prove that the directed acyclic graph (DAG) is identifiable from the joint distribution Peters et al. [2013] require a strictly positive density. Their proof makes use of the intersection property of conditional independence (Definition 2) which is known to hold for positive densities [e.g. Pearl, 2009, 1.1.5].

1.2 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.

As mentioned above, some causal discovery methods based on structural equation models require the intersection property for identification; they therefore rely on the strict positivity of the density. This can be achieved by fully supported noise variables, 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 structural equation models [Shimizu et al., 2006] but not for non-linear models. As an alternative, we provide a condition that excludes constant functions and leads to identifiability, too (Proposition 4).

In Section 2, we provide an example of a structural equation model that violates the intersection property (but satisfies causal minimality). 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 are 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 DAG 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.

2

1.3 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.

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

The intersection property of conditional independence is defined as follows [e.g. Pearl, 2009, 1.1.5].

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

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

The intersection property (2) has been proven to hold for strictly positive densities [e.g. Pearl, 2009, 1.1.5]. It is also known that the intersection property does not necessarily hold if the joint distribution does not have a density [e.g. Dawid, 1979b]. Dawid [1980] provides measure-theoretic necessary and sufficient conditions for the intersection property. In this work we assume the existence of a density (A0) and provide more detailed conditions under which the intersection property holds.

2 Counter Example

We now give an example of a distribution that does not satisfy the intersection property (2). Since the joint distribution is absolutely continuous with respect to the Lebesgue measure, 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 structural equation model for random variables* X, A, B:

$$A = N_A \,,$$
$$B = A + N_B \,,$$
$$X = f(B) + N_X \,,$$

with $N_A \sim \mathcal{U}([-2; -1] \cup [1; 2])$, $N_B, N_X \sim \mathcal{U}([-0.3; 0.3])$ *being jointly independent. Let the function* f *be of the form*

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

where the function g *can be chosen to make* f *arbitrarily smooth. Some parts of this structural equation model are summarized in Figure 1. We clearly have* $X \perp\!\!\!\perp A \mid B$ *and* $X \perp\!\!\!\perp B \mid A$ *but* $X \not\!\perp\!\!\!\perp A$ *and* $X \not\!\perp\!\!\!\perp B$. *A formal proof of this statement is provided in the more general setting of Proposition 2. It will turn out to be important that the two connected components of the support of* A *and* B *cannot be connected by an axis-parallel line. In the notation introduced in Definition 3 below, this means* Z_1 *and* Z_2 *are not equivalent. 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* $X \perp\!\!\!\perp A, B$ *that is predicted by the intersection property. This observation will be formalized as the weak intersection property in Proposition 1.*

Example 1 has the following important implication for causal inference. The distribution satisfies causal minimality with respect to two different graphs, namely $A \rightarrow B \rightarrow X$ and $X \leftarrow A \rightarrow B$ (see Figure 1). 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. [2013] assume
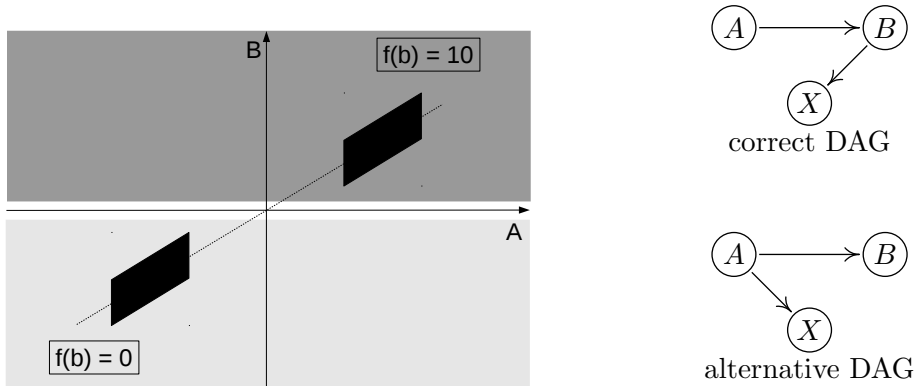
Figure 1: Example 1. The plot on the left hand side shows the support of variables A and B in black. In the areas filled with dark gray and light gray the function f takes values ten and minus ten, respectively. The structural equation model corresponds to the top graph but the distribution can also be generated by a structural equation model with the bottom graph.

strictly positive densities in order to assure the intersection property. More precisely, the example shows that Lemma 37 in [Peters et al., 2013] does not hold anymore when the positivity is violated.

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. For these results, the notion of path-connectedness becomes important. 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 from $\mathcal{S}$ can be connected by a path in $\mathcal{S}$. We require the following definition.

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

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

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

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

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

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

We then say that Z_i^c and Z_j^c are equivalent if and only if there is a sequence $Z_i^c = Z_{i_1}^c, \ldots, Z_{i_m}^c = Z_j^c$ with two neighbours $Z_{i_k}^c$ and $Z_{i_{k+1}}^c$ being coordinate-wise connected. We represent these equivalence classes by the union of all its members. These unions we denote by $(U_i^c)_i$.

We further introduce a deterministic function U^c of the variables A and B. We set

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

We have that $U^c = i$ if and only if $A \in \mathrm{proj}_A(U_i^c)$ if and only if $B \in \mathrm{proj}_B(U_i^c)$.

Note that the projections $\mathrm{proj}_A(U_i^c)$ are disjoint (for different i); similarly 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. Since they cannot be connected by axis-parallel lines, they are not equivalent; thus, one of them corresponds to 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.

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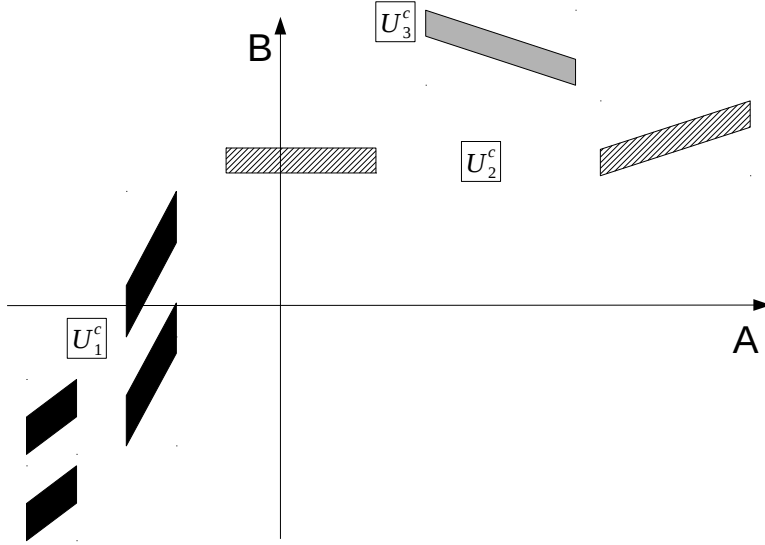
Figure 2: Each block represents one path-connected component Z_i^c of the support of $p(a, b)$. All blocks with the same filling are equivalent since they can be connected by axis-parallel lines. There are three different fillings corresponding to the equivalence classes U_1^c, U_2^c and U_3^c.

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

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

for all x, a, b with $p(a, b, c) > 0$.

Proposition 2 (Failure of Intersection Property)**.** *Assume (A0), (A1) and that there are two different sets $U_1^{c^*} \neq U_2^{c^*}$ for some c^* with $p(c^*) > 0$. Then there is 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). Then*

> *The intersection property* (2) *holds for all variables X.*
>
> $\iff$ *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 now define what we mean by identifiability of the graph in continuous additive noise models. Assume that a joint distribution over $X_1, \ldots, X_p$ is generated by a structural equation model (SEM)

$$X_i = f_i(X_{\mathbf{PA}_i}) + N_i, \tag{3}$$

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 directed acyclic graph $\mathcal{G}$. To simplify notation, we identify variables X_i with its index (or node) i. We consider the following statement

> $(*)$ $\mathcal{G}$ *is identifiable from the joint distribution, i.e. it cannot be generated by an SEM with different graph $\mathcal{H} \neq \mathcal{G}$.*

Peters et al. [2013, Theorem 27] prove this identifiability by extending the identifiability from graphs with two nodes to graphs with an arbitrary number of variables. Because they require the intersection property, it is shown only for strictly positive densities. But since Corollary 1 provides weaker assumption for the intersection property, we can use it to obtain new identifiability results.

Proposition 3. *Assume that a joint distribution over $X_1, \ldots, X_p$ is generated by a structural equation model* (3)*. If all densities of $N_1, \ldots, N_p$ are path-connected, then the density of $X_1, \ldots, X_p$ is path-connected, too. Thus, the intersection property* (2) *holds for any disjoint sets of variables $X, A, B, C \in \{X_1, \ldots, X_p\}$ (see Corollary 1). Therefore, statement $(*)$ holds if the noise variables have continuous densities and path-connected support.*

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.

additional assumption on continuous ANMs	identifiability of graph, see $(*)$
noise variables with full support	✓ [Peters et al., 2013]
noise variables with path-connected support	✓ Proposition 3
non-constant functions, see Proposition 4	✓ Proposition 4
none of the above satisfied	✗ Example 1

Table 1: This table shows conditions for continuous additive noise models (ANMs) that lead to identifiability 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.

Proposition 4. *Assume that a joint distribution over $X_1, \ldots, X_p$ is generated by a structural equation model (3) with graph $\mathcal{G}$. Let us denote the non-descendants of X_i by $\mathbf{ND}_i^{\mathcal{G}}$. 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 $X_{\mathbf{C}} \subseteq \mathbf{ND}_i^{\mathcal{G}} \backslash \{X_j\}$, 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 $\mathbf{PA}_i \setminus \{j\} \subseteq \mathbf{S} \subseteq \mathbf{ND}_i^{\mathcal{G}} \setminus \{j\}$, 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.

5 Conclusion

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 conditioning on any equivalence class of path-connected components, we call this the weak intersection property.

This insight has a direct application in causal inference. For continuous additive noise models we can prove identifiability of the graph from the joint

distribution using strictly weaker assumptions than before.

6 Proofs

6.1 Proof of Proposition 1

We require the following well-known lemma [e.g. Dawid, 1979a].

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 $p(a, b) > 0$ and $p(b) > 0$.

Proof. (of Proposition 1) We have by Lemma 1

$$p(x \,|\, b, c) = p(x \,|\, a, b, c) = p(x \,|\, a, c) \tag{4}$$

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

$$p(x \,|\, b, c) = p(x \,|\, \tilde{b}, c) \tag{5}$$

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 equation (5) for $b, \tilde{b} \in Z_i^c$, that is there is a path $(a(t), b(t))$, such that $p(a(t), b(t), c) > 0$ for all $0 \le t \le 1$, and $b(0) = b$ and $b(1) = \tilde{b}$. Since the interval $[0, 1]$ is compact and p is continuous, the path $\{(a(t), b(t)) : 0 \le t \le 1\}$ is compact, too. 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 $p(a, b, c) > 0$. Since this is an open cover of the space, choose a finite subset, of size n say, of all those balls that still provide an open cover of the path. Without loss of generality generality let $(a(0), b(0))$ be the center of ball 1 and $(a(1), b(1))$ be the center of ball n. It suffices to show that equation (5) holds for the centres of two neighbouring balls, say (a_1, b_1) and (a_2, b_2). Choose one 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 zero. Therefore, using equation (4) 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 equation (5) for $b, \tilde{b} \in Z_i^c$.

Step 2, we prove equation (5) 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 above. 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 equation (4) 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 equation (5). 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 $Z_1^c, \dots, Z_n^c$ coordinate-connecting these components. Combining steps 1 and 2 proves equation (5).

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

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

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 $b \in \mathrm{proj}_B(\tilde{U}_i^c)$ we have $p(x \,|\, \tilde{b}, c) = p(x \,|\, b, c)$ by equation (5). Ergo,

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

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

$$= p(x \,|\, b, c) \int_{a \in \mathrm{proj}_A(U_i^c)} p(a \,|\, c) \; da$$

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

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

Together with equation (4) this leads to

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

$\square$

6.2   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, A, B, C) being jointly independent. 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,

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

This shows that $X \not\perp\!\!\!\perp B \,|\, C = c^*$.

$\square$

6.3 Proof of Proposition 3

Proof. Since the true structure corresponds to a directed acyclic graph, 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 (3) as

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

where the functions $\tilde{f}_i$ are the same as f_i except they are constant in the additional input arguments. The statement of the proposition then follows 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 statements of Lemma 37 and thus Proposition 28 from Peters et al. [2013] remain correct, which proves $(*)$ for noise variables with continuous densities and path-connected support. $\qquad\square$

6.4 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 37 might not hold but more importantly Proposition 28 does [both from Peters et al., 2013]. This proves $(*)$. $\qquad\square$

References

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

A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B*, 41:1–31, 1979a.

A. P. Dawid. Some misleading arguments involving conditional independence. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41:249–252, 1979b.

A. P. Dawid. Conditional independence for statistical operations. *Annals of Statistics*, 8:598–617, 1980.

D. Heckerman, C. Meek, and G. Cooper. A Bayesian approach to causal discovery. In C. Glymour and G. Cooper, editors, *Computation, Causation, and Discovery*, pages 141–165, Cambridge, MA, 1999. MIT Press.

P.O. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21 (NIPS)*, pages 689–696. MIT Press, 2009.

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

J. Peters, J. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models, 2013. arXiv:1309.6779.

S. Shimizu, P.O. Hoyer, A. Hyvärinen, and A.J. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

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

T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the 6th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 1991.