Minimal I-map dags: necessary and sufficient conditions

Helen Armstrong[†] and Kevin B. Korb[*]

[†]School of Mathematics
University of New South Wales
Sydney, NSW 2052
AUSTRALIA
helen@maths.unsw.edu.au

[*]School of Computer Science & Software Eng
Monash University
Clayton, VIC 3800
AUSTRALIA
korb@csse.monash.edu.au

Abstract Underlying Bayesian network modeling and causal discovery is Pearl's result that for every strictly positive distribution there exists a unique minimal undirected independency map (I-map), and, for every total ordering of variables, a unique minimal I-map dag [Pearl, 1988]. Whilst strict positivity is sufficient for uniqueness, we show that it is not necessary. We show that for the existence of a unique minimal I-map dag it is a necessary and sufficient condition that there be no two variables which are equivalent. This allows modeling of a wider class of processes, including those with fully effective causal interventions. Our result is as simple as, but more general than, that of Pearl.

1 INTRODUCTION

It is a seminal result underpinning the use of Bayesian networks that for every strictly positive distribution there exists a unique minimal undirected independency map (I-map), and, for every total ordering of variables, a unique minimal I-map directed acyclic graph (dag) [Pearl, 1988]. This underlies Pearl's network construction algorithm ([Pearl, 1988] §3.3) as well as Verma and Pearl's constraint-based conditional independence learning algorithm for building Bayesian networks from observational data [Verma and Pearl, 1990], adapted by Spirtes et al. in TETRAD II [Spirtes et al., 1993]. Whilst strict positivity is sufficient for uniqueness, we show that it is not necessary and give a necessary and sufficient condition for the existence of a unique minimal

I-map dag. For completeness, we also present a proof that every strictly positive distribution satisfies Pearl's "intersection property" (in Appendix C), which was an unproven assumption of his uniqueness result in [Pearl, 1988]. Our basic result is simply:

RESULT A probability distribution p has a unique minimal I-map dag representation for a given ordering if and only if for each subset of variables $\{X, Y, Z\}$ with X depending on $\{Y, Z\}$, there is no 1-1 onto function g such that $Z = g(Y)$ almost everywhere.
In other words, it is a necessary and sufficient condition that there be no two variables which are equivalent (unless they are not dependent upon any further variables). This result is just as simple as, but more general than, that of Pearl.

2 BACKGROUND

Since Pearl's ground-breaking *Probabilistic Reasoning in Intelligent Systems* [Pearl, 1988] some Bayesian net modelers will have been avoiding non-strictly positive models out of fear that Pearl's results for network construction will fail to apply. So one motivation for the current work is to demonstrate that those techniques apply for some non-positive distributions as well.

Our second motivation is to better develop the causal interpretation of Bayesian networks. Many have said or suggested that Bayesian networks are not inherently causal (e.g., [Borchelt and Kruse, 2002, pp. 263-4]), or that a causal interpretation cannot be sustained (e.g., [Williamson, 2002]). The former is certainly true at least in a trivial sense: namely, in the sense that for every causal structure in N variables that fixes a specific ordering, there are $(N-1)!$ non-causal orderings which will give rise to a non-causal Bayesian network. On the other hand, we propose that there is a deeper truth: that every Bayesian network which represents a probability distribution depends upon there being some Bayesian network which represents a causal process giving rise to that probability distribution.[1] The complexity of non-causal graphs is bounded below by the complexity of the true causal graph, which is a clear consequence from a consideration of Chickering's transformation rule [Chickering, 1995]. In other words, in a direct and compelling sense, the causal interpretation of Bayesian networks is the primary one upon which every other interpretation depends. This we call the **Thesis of Causal Primacy**. For the time being, we cannot turn the thesis into a theorem, although we can point at some empirical evidence for it [Korb and Boulton, 2003].

[1] And this holds also for Bayesian networks strictly representing an agent's subjective states of belief, since all such agents we know of are constructed out of causal, physical systems, such as brains.

Whether or not causal models are primary, Bayesian networks which also *are* causal models — that is, those where arcs represent direct causal relations between variables — are clearly important in the application of the technology. Whereas any Bayesian network which represents a probability distribution will inform you of the distribution conditional upon some set of observations, only a causal network can inform you of the implications of an *intervention* upon one or more nodes [Pearl, 2000]. Since manufacturing processes, environmental management, medical treatments, control processes, and many other applications are directly concerned with what happens to a system under intervention, simply finding *some* Bayesian network to represent a probability distribution is not enough.

When dealing with causal models, it is clear that there is a unique correct dag representing the causal system.[2] For example, given that X and Y are directly dependent, then given further knowledge, for example, that X precedes Y, we infer immediately that the correct model is $X \rightarrow Y$. And we are intuitively led to Pearl's seminal result [Pearl, 1988] that given an ordering, for every strictly positive probability distribution (or density) p, there exists a unique Bayesian network G which reports via d-separation the maximum number of independencies in p without reporting any false independencies. Thus, for the given ordering, such a G is uniquely the sparsest possible I-map representation. Pearl [Pearl, 1988] coined the term "minimal I-map" for such a structure.[3] Formally:

Definition 1 *A directed graph G on a set of nodes $\mathbf{V}$ is an* **independency map** *(or I-map) of a joint probability density/distribution p over a set of random variables $\mathbf{U}$ if there is a 1-1 correspondence between the elements of $\mathbf{U}$ and the nodes $\mathbf{V}$ of G such that for all disjoint subsets $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ of $\mathbf{V}$ we have $\mathbf{Z}$ d-separating $\mathbf{X}$ from $\mathbf{Y}$ implies $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z}$ (that is, X is conditionally independent of Y given Z). G is also a* **minimal** *I-map of p if deleting any edge of G produces a graph which is not an I-map.*

For succinctness, whenever we have $\mathbf{Z}$ d-separating $\mathbf{X}$ from $\mathbf{Y}$ in a graph G, we will say that $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z}$ is implied by G, or implied by G via d-separation.

3 UNIQUE MINIMALITY AND CAUSALITY

There is a clear connection between minimality and causal interpretation. When a Bayesian network is interpreted causally, each edge is intended to

[2] To be sure, on the assumption that there is *some* dag which represents the causal system. We will not be considering such questions as circular causality or the failure of the Markov condition in quantum mechanics.

[3] Some authors use the term "Markov network" for a minimal I-map, but we will use Pearl's terminology throughout.

denote the existence of a direct causal influence. The best model will display no extraneous edges, as these suggest non-existent direct causal influences. Although the suggestion can be nullified by parameters yielding no causal effect, there will remain a preference for minimality, as this reflects also a preference for simplicity. Adding a spurious arc does not limit that model's representational power, but it does add unnecessary complexity, violating the principle of Ockham.

Whilst it is possible for a minimal I-map to fail to display all independencies, if a given Bayesian network displays all and only the independencies in a distribution p, then from Chickering's transformational equivalence result [Chickering, 1995], any other Bayesian network which implies via d-separation the independencies in p must have at least as many edges. Hence, the true causal graph is minimal (though it may not be uniquely so, since other orderings may result in a graph with the same number of edges).

The uniqueness of a Bayesian network for a given ordering requires, for each node, the implication by the given independence constraints of a unique parent set. Such uniqueness requires what Pearl [Pearl, 1988] termed the "intersection property":

Definition 2 *Intersection property (IP):* $X \perp\!\!\!\perp Y|\{Z, W\}$ *and* $X \perp\!\!\!\perp Z|\{Y, W\}$ *implies*
$X \perp\!\!\!\perp \{Y, Z\}|W$.

Every strictly positive distribution (density) satisfies IP. However, there are many distributions (densities) which are not strictly positive which also satisfy this property. If we can find a weaker characterisation of those distributions which satisfy the intersection property, and so have a *unique* minimal I-map representation for a specific ordering, we can then model a wider class of distributions with Bayesian nets.

Distributions which are not strictly positive can be used to combine representations which otherwise would require separate graphs. For example, consider medical diagnosis for a particular condition which is dependent on (amongst many other things) the sex of the patient. Suppose pregnancy increases the incidence of this condition. Then clearly every instantiation of variables which includes *Sex=male*, *Pregnancy=yes* will have probability zero. Rather than using two separate graphs (one for each sex), the weaker condition which does not require strict positivity can be employed to infer a single network incorporating both the *Sex* and *Pregnancy* variables. Many modeling problems incorporate such impossible joint events.

The natural question raised by the above is whether uniqueness is necessary for non-causal Bayesian network modelling. If there exist two minimal I-maps for a given ordering, then will one of them necessarily be a "better" model of the system than the other? If probabilistic dependencies between variables are all that we are interested in, the answer is surely "no". But

very many of the applications for which Bayesian networks are inherently suitable involve their use in causal reasoning, for example, in such areas as medicine, sociology, economics and social policy. If we are to use the model for causal reasoning, and not merely for reasoning about the interdependencies between variables, then all but one minimal I-map must mislead about at least one causal intervention process. So, non-unique minimal I-maps are not good enough for such application purposes.

Furthermore, the causal interpretation is the only foundation we have for the automated learning of Bayesian networks. Whilst in the past the structure of the causal model was commonly elicited from human domain experts, the considerable progress over the last decade in the machine learning of Bayesian network structures from data is allowing the wider penetration of Bayesian net modelling into industry and science. All of this work assumes a causal interpretation.

Some might respond that neither the Verma-Pearl algorithm [Verma and Pearl, 1990] nor the metric learning algorithms (e.g., [Cooper and Herskovits, 1992]) makes reference to anything more than probabilistic dependencies or their absence, so the claimed assumption of an underlying causal interpretation is spurious. But probabilistic dependencies themselves depend upon something. The probabilistic dependencies we find in nature can themselves be *explained*: they are either a matter of semantic/logical/mathematical necessity (which we abbreviate as "logical" *passim*), or they arise because the dependent variables participate in a causal chain, or else the variables have some causal ancestor. The only further answer is that they are related by magick, an answer which is ruled out by Hans Reichenbach's *Principle of the Common Cause* [Reichenbach, 1956]. Reichenbach's rejection of magick is one we endorse. We shall resort to magick when the full resources of our science fail to come to grips with the causal mechanism underlying our probabilistic observations, and not before. On this methodological principle, underlying all of our sciences, rests the dependence of probability upon causality, and so then also our discovery algorithms based upon probability.[4]

4 IP AND STRONG VIOLATION

In order to prove the uniqueness of the minimal I-map dag for a strictly positive distribution, Pearl invokes the intersection property [Pearl, 1988]. We will show that there is no such dag only when IP is violated, and violated in a stronger way than that envisioned by Pearl.

IP is violated whenever $X \perp\!\!\!\perp Y \,|Z$ and $X \perp\!\!\!\perp Z|Y$ and $X \not\perp\!\!\!\perp \{Y, Z\}$. It is a standard result that $X \perp\!\!\!\perp \{Y, Z\}$ implies $X \perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp Z$ but not conversely

[4]We appreciate that some disagree, over and above difficulties with accommodating quantum mechanics. Jon Williamson is one [Williamson, 2002]. However, we must leave a response to those arguments to another occasion.

(there exist violating distributions such that $X \not\perp\!\!\!\perp \{Y, Z\}$ but both $X \perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp Z$ are true).

So, violation of IP would hold whenever $X \perp\!\!\!\perp Y|Z$ and $X \perp\!\!\!\perp Z|Y$ and any of the following held:

1. $X \not\perp\!\!\!\perp Y$ and $X \not\perp\!\!\!\perp Z$ (which implies $X \not\perp\!\!\!\perp \{Y, Z\}$)
2. $X \not\perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp Z$ (which implies $X \not\perp\!\!\!\perp \{Y, Z\}$)
3. $X \perp\!\!\!\perp Y$ and $X \not\perp\!\!\!\perp Z$ (which implies $X \not\perp\!\!\!\perp \{Y, Z\}$)
4. $X \perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp Z$ and $X \not\perp\!\!\!\perp \{Y, Z\}$

Theorem 1 *$X \perp\!\!\!\perp Y|Z$ and $X \perp\!\!\!\perp Z|Y$ and $X \not\perp\!\!\!\perp \{Y, Z\}$ together imply that both $X \not\perp\!\!\!\perp Y$ and $X \not\perp\!\!\!\perp Z$.*

Proof See Appendix A.

By the above theorem, the intersection property is violated whenever $X \perp\!\!\!\perp Y|Z$ and $X \perp\!\!\!\perp Z|Y$ and $X \not\perp\!\!\!\perp Y$ and $X \not\perp\!\!\!\perp Z$, and we will not need to consider the other cases.

In order to facilitate discussion, we introduce:

Definition 3 *Strong violation of the intersection property: $X \perp\!\!\!\perp Y|Z$ and $X \perp\!\!\!\perp Z|Y$ and $X \not\perp\!\!\!\perp Y$ and $X \not\perp\!\!\!\perp Z$.*

We will show that only those distributions satisfying Definition 3 are not representable by a unique minimal I-map.

5 IP IS EQUIVALENT TO UNIQUENESS OF THE MINIMAL I-MAP

Clearly, every distribution has a minimal I-map representation: simply delete from the complete graph an edge if such deletion does not result in the implication of a false independency until no more can be deleted. The representational problem is ensuring, for a given ordering, the uniqueness of such a representation.

That the intersection property is sufficient for uniqueness follows from Pearl [Pearl, 1988, p. 141]. To prove the uniqueness of the parent set for each variable, given an ordering, and hence the uniqueness of the minimal I-map which can be inferred from a set of independence constraints, Pearl invokes IP. Explicitly, he uses the fact that the Markov boundary of any given variable "is unique because the intersection property renders [the set of Markov blankets for each element] closed under intersection". Conversely, we illustrate how violation of IP results in non-uniqueness of the inferred representation.

Assume strong violation of the intersection property (which follows, by the preceding theorem, from violation and the positivity constraint on the marginals for Y and Z). Assume the ordering $< Z, Y, X >$. Then both $X \leftarrow Z \rightarrow Y$ and $X \leftarrow Y \leftarrow Z$ imply no false independencies and respect the given ordering (though both imply a false dependency and so neither is a D-map). Both are minimal as the deletion of any edge would imply a false independency. Hence, there is no unique minimal I-map representation. Note that if IP were satisfied (i.e., if $X \perp\!\!\!\perp \{Y, Z\}$), then the unique minimal representation in the case of $Y \perp\!\!\!\perp Z$ would be the empty graph, and if $Y \not\perp\!\!\!\perp Z$ then the unique minimal representation would be the graph containing a single edge $Y \leftarrow Z$.

To see how violation of IP can occur, and get some idea of the general characterisation of violating distributions, consider this example. Let X be your weight, Y be your height in centimetres and Z be your height in inches. Then $X \perp\!\!\!\perp Y | Z$ (once I know your height in inches, your height in centimetres is irrelevant to my belief about your weight) and $X \perp\!\!\!\perp Z | Y$, yet $X \not\perp\!\!\!\perp Y$ and $X \not\perp\!\!\!\perp Z$. Clearly, Y and Z contain the same information with respect to X, and there is a 1-1 onto relationship between Y and Z: these variables are logically constrained.

In the linear Gaussian case, it can be proven that this intuitive example does violate IP. Assume that X, Y, Z are Gaussian and defined as per the previous paragraph. Then the correlation ρ_{YZ} between Y and Z is clearly one (your height in inches is a scalar multiple of your height in centimetres), and Y and Z are equally correlated with X, i.e., $\rho_{XY} = \rho_{XZ}$. Hence, the partial correlation $\rho_{XY.Z} = \dfrac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{1 - \rho_{XZ}^2}\sqrt{1 - \rho_{YZ}^2}} = 0$, and similarly for $\rho_{XZ.Y}$. For Gaussian distributions zero correlations imply independencies, and so we have $X \perp\!\!\!\perp Y | Z$ and $X \perp\!\!\!\perp Z | Y$, yet $X \not\perp\!\!\!\perp Y$ and $X \not\perp\!\!\!\perp Z$.

Furthermore, in the linear Gaussian case, it is easy to prove that when IP is violated we have Y and Z perfectly correlated; i.e., they are logically constrained. If $X \perp\!\!\!\perp Y | Z$, then we know that $\rho_{XY.Z} = 0$. But by definition $\rho_{XY.Z} = 0$ implies that $\rho_{XY} = \rho_{XZ}\rho_{YZ}$. Similarly, the condition $X \perp\!\!\!\perp Z | Y$ implies that $\rho_{XZ} = \rho_{XY}\rho_{YZ}$, and these together imply that $\rho_{YZ} = \pm 1$ (whenever ρ_{XY} and ρ_{XZ} are non-zero).

Intuitively, Y and Z are not capable of being represented as separate nodes as their "identical" dependency with X ensures there can be no justification for preferring the adjacency of $X \leftarrow Y$ over $X \leftarrow Z$.

For an illustration of strong violation of IP, and hence no unique graph, for a discrete case consider the discrete distribution $p(X, Y, Z)$ given in Appendix E. $X \not\perp\!\!\!\perp Y$ and $X \not\perp\!\!\!\perp Z$ and, whenever it is defined, $p(x|yz)$ is equal to both $p(x|y)$ and $p(x|z)$. Once again, we find that Y and Z are logically constrained and each of the four conditional distributions $p(Y|Z = 1)$, $p(Y|Z = 0)$, $p(Z|Y = 1)$, $p(Z|Y = 0)$ is degenerate (i.e., is a zero/one distribution).

6 STRICT POSITIVITY IS NOT NECESSARY FOR UNIQUENESS

Consider the following discrete distribution. X, Y and Z are each dichotomous random variables, such that $X \perp\!\!\!\perp Y$. For example, X and Y are the outcomes of two coin tosses, with $p(X = 1) = \frac{1}{3}$ and $p(Y = 1) = \frac{1}{4}$ (with one representing a head and zero a tail). Z is a variable dependent on the outcome of X and Y, with probabilities as given in Appendix F. Note that the event $\{X = 0, Z = 1, Y = 1\}$ has probability zero, so the joint distribution is not strictly positive.

The event $\{Z = 0, Y = 1\}$ has non-zero probability, so the conditional probability
$p(X = 1|Z = 0, Y = 1)$ is well defined and is equal to $\frac{1}{7}$. Furthermore, $p(X = 1|Y = 1) = \frac{1}{3}$, and so $X \not\perp\!\!\!\perp Z|Y$. Similarly $X \not\perp\!\!\!\perp Y|Z$ [$p(X = 1|Z = 0) = \frac{7}{16} \neq p(X = 1|Z = 0, Y = 1)$] and $Y \not\perp\!\!\!\perp Z|X$, so IP is not violated. The unique graph for the ordering $< X, Y, Z >$ is $X \rightarrow Z \leftarrow Y$.

It is significant that none of the four conditional distributions $p(Y|Z = 1)$, $p(Y|Z = 0)$, $p(Z|Y = 1)$, $p(Z|Y = 0)$ is degenerate; i.e., none of these is a zero/one distribution.

A particular kind of application of this result, of some significance, is the modeling of causal interventions. If we add to the subnetwork $X \leftarrow Y$ a control variable for X we get $D_X \rightarrow X \leftarrow Y$, having the same structure as above. Now in the general case we would like to allow the control variable to *not* strictly determine the value of the variable being manipulated, since real-world controls are not invariably effective. But in particular cases we would like to model perfect controls, perhaps just as a simplifying feature. Our result here endorses that option: the violation of strict positivity, on the assumption that D_X renders the effect of Y null but not vice versa, is no impediment to finding a unique minimal I-map. Our result also endorses the application of causal discovery algorithms to learning causal structures from measurements which include explicit representation of effective causal interventions.

7 A NECESSARY AND SUFFICIENT CONDITION

Pearl uses the overly strong condition of strict positivity to guarantee that a distribution satisfies IP. The question arises, what is a weaker characterisation than strict positivity which remains sufficient for uniqueness of the minimal I-map representation? That is, one which is both necessary and sufficient.

Consider again height and weight. As we saw, letting X be your weight, Y be your height in centimetres (cm) and Z be your height in inches, if your

height in cm is known, then your height in inches is irrelevant to inference about your weight, and conversely; i.e., $X \not\perp\!\!\!\perp Y$ and $X \not\perp\!\!\!\perp Z$ but $X \perp\!\!\!\perp Y|Z$ and $X \perp\!\!\!\perp Z|Y$.

If $X \not\perp\!\!\!\perp \{Y, Z\}$ but $X \perp\!\!\!\perp Y|Z$, then Y contains no extra relevant information to X than what is already contained in Z. If $X \not\perp\!\!\!\perp \{Y, Z\}$ but $X \perp\!\!\!\perp Z|Y$, then Z contains no extra relevant information to X than what is already contained in Y. In short, there must be a 1-1 relation between the information in Y that is relevant to X, and the information in Z that is relevant to X. Following our intuition, we see that such a situation occurs if, and only if, there is a logical constraint relating the sample spaces Ω_Y and Ω_Z; i.e., that each of the conditional distributions or densities $\{p(Y|z) : z \in \Omega_Z\}$ is degenerate, and similarly for $\{p(Z|y) : y \in \Omega_Y\}$.

More formally, assume $X \not\perp\!\!\!\perp \{Y, Z\}$ but $X \perp\!\!\!\perp Y|Z$ and $X \perp\!\!\!\perp Z|Y$. For each point $y \in \Omega_Y$ such that $p(x, y) \neq 0$, define $g : \Omega_Y \to \Omega_Z$ by the following

$$g(y_o) = \{z \in \Omega_Z \text{ s.t. } \forall x \in \Omega_X \; p(x|y_o) = p(x|y_o, z)\} \tag{1}$$

Theorem 2 *g is a well defined 1-1 function. That is, the set of all such $z \in \Omega_Z$ has one and only one member for all y_o such that $p(y_o) \neq 0$. Moreover $p(Z = g(Y)) = 1$.*

Proof. See Appendix D.

Note that uniqueness of $z \in \Omega_Z$ clearly fails when $X \perp\!\!\!\perp \{Y, Z\}$. We can similarly define $g^{-1} : \Omega_Z \to \Omega_Y$ on all points $z_o \in \Omega_Z$ such that $p(z_o) \neq 0$.

Such a g, establishing an invertible correspondence between Ω_Y and Ω_Z, can be defined whenever the "dependency relevance" of Y and Z to a variable X is equal (i.e., $X \perp\!\!\!\perp Y|Z$ and $X \perp\!\!\!\perp Z|Y$).

Conversely, it is a classical result that if a 1-1 function g exists such that $Z = g(Y)$, then the sigma algebras generated by Y and Z are the same, hence the "dependency relevance" of Y and Z to a variable X is equal and IP is violated. Note that the existence of g ensures that the distribution has degenerate conditionals, which accords with the examples given above.

To see that a functional dependency between the spaces is not sufficient, and that both the 1-1 and onto properties are critical to ensure a direct correspondence between the information in Y and the information in Z (or, more formally, between their sigma algebras), consider this example. Let X be the effect on health of drinking red wine, Y the absolute value of percentage change in cholesterol levels and Z be directional percentage change in cholesterol levels (so Z can be positive or negative, but $Y \geq 0$). Then there is no onto function from Y to Z which is also 1-1 and, conversely, any 1-1 function from Z to Y cannot be onto. So even though Y is a deterministic function of Z, and a subspace of Z is a deterministic function of Y, in relation to X, knowing Z is more useful than knowing Y. In particular, $X \not\perp\!\!\!\perp Z|Y$, but $X \perp\!\!\!\perp Y|Z$. Hence, the intersection property is not violated, and the unique minimal I-map representation is $X \leftarrow Z \rightarrow Y$.

8 CONCLUSION

Pearl's seminal result underpinning the whole of Bayesian net modelling was that every strictly positive distribution can be represented by a unique minimal I-map (with respect to a given ordering). We have demonstrated that that result was overly restrictive. A weaker condition that is both necessary and sufficient for a given distribution to have a unique minimal I-map dag representation is that there be no logical constraints between sets of variables; i.e., that there be no degenerate conditional distributions over any subset of variables. More formally, the restriction is that there exist no 1-1 function $g : \Omega_Y \rightarrow \Omega_Z$ such that the full joint distribution $p_{\mathbf{V}}$ has a marginal p_{XYZ} over $X \in \mathbf{V}$ and $Y, Z \subset \mathbf{V}$ such that $p_{XYZ}(Z = g(Y)) = 1$.

This result opens up the standard Bayesian network modelling techniques to distributions that are not strictly positive, where some of the values of some variables are strictly determined by those of others, including networks incorporating fully effective causal interventions. Our result also may help us understand the relation between causality and Bayesian networks, since unique minimality appears to be of the essence of the causal interpretation.

Acknowledgement

We are grateful to Charles Twardy and Ann Nicholson for spotting errors in an earlier draft.

APPENDICES

NOTATION

Let $p_{\mathbf{V}}$ be a distribution over a set of variables $\mathbf{V} = \{X_1, X_2, \cdots, X_n\}$ with Cartesian product sample space $\Omega_{\mathbf{V}} = \Omega_{X_1} \times \Omega_{X_2} \times \cdots \times \Omega_{X_n}$. For brevity, denote any marginal distribution derived from $p_{\mathbf{V}}$ over a subset of variables $\{X, Y, Z\}$ by p and use the shorthand $p(x, y, z)$ to denote the measure under p of the singleton event $\{X = x, Y = y, Z = z\}$ in the product sample space $\Omega_X \times \Omega_Y \times \Omega_Z$. Analogously, denote all conditional distributions and probabilities, and all continuous joint, marginal and conditional densities whenever they are defined.

APPENDIX A: Proof that $X \perp\!\!\!\perp Y | Z$ and $X \perp\!\!\!\perp Z | Y$ and $X \not\perp\!\!\!\perp \{Y, Z\}$ implies both $X \not\perp\!\!\!\perp Y$ and $X \not\perp\!\!\!\perp Z$

The following proof holds for both p a discrete distribution and p a continuous density.

Assume that $X \perp\!\!\!\perp Y|Z$ and $X \perp\!\!\!\perp Z|Y$ and $X \not\perp\!\!\!\perp \{Y, Z\}$. We show that the assumption of either $X \perp\!\!\!\perp Y$ or $X \perp\!\!\!\perp Z$ leads to a contradiction.

First assume additionally that $X \perp\!\!\!\perp Y$ and that there exists at least one $y \in \Omega_Y$ and $z \in \Omega_Z$ such that $p(y, z) \neq 0$. Then

$$\begin{aligned} p(x|y,z) &= p(x|y) \text{ since } X \perp\!\!\!\perp Z|Y \\ &= p(x) \text{ by assumption that } X \perp\!\!\!\perp Y \end{aligned} \tag{1}$$

Thus $X \perp\!\!\!\perp \{Y, Z\}$ whenever $p(y, z) \neq 0$, which contradicts the dependency assumption.

Swapping the roles of z and y in the above, we similarly find that the assumption $X \perp\!\!\!\perp Z$ also leads to a contradiction.

Hence if $X \perp\!\!\!\perp Y|Z$ and $X \perp\!\!\!\perp Z|Y$ and $X \not\perp\!\!\!\perp \{Y, Z\}$, then both $X \not\perp\!\!\!\perp Y$ and $X \not\perp\!\!\!\perp Z$.

APPENDIX B: Proof that $X \perp\!\!\!\perp Y|Z$ and $X \perp\!\!\!\perp Z|Y$ and $X \not\perp\!\!\!\perp \{Y, Z\} \Leftrightarrow \forall (x, y, z) \in \Omega_X \times \Omega_Y \times \Omega_Z$ such that $p(y, z) \neq 0$, $p(x|y, z) = p(x|y) = p(x|z)$

In the following proof, the notation p stands for a joint density or discrete distribution.

Assume $X \perp\!\!\!\perp Y|Z$ and $X \perp\!\!\!\perp Z|Y$ and $X \not\perp\!\!\!\perp \{Y, Z\}$. Consider all points in $\Omega_X \times \Omega_Y \times \Omega_Z$ where the conditional distribution $p(x|y, z)$ is defined; i.e., $p(y, z) \neq 0$. Then $\forall x \in \Omega_X$

$$\begin{aligned} p(x|y,z) &= \frac{p(x,y,z)}{p(y,z)} \\ &= \frac{p(x,y|z)p(z)}{p(y|z)p(z)} \\ &= \frac{p(x|z)p(y|z)}{p(y|z)} \text{ since } X \perp\!\!\!\perp Y|Z \\ &= p(x|z). \end{aligned} \tag{2}$$

Similarly, by conditioning on y rather than z, we have that $p(x|y, z) = p(x|y)$.

Hence

$$p(x|y,z) = p(x|z) \tag{3}$$
$$= p(x|y) \tag{4}$$

whenever $p(y,z) \neq 0$.

Conversely, assume $p(x|y,z) = p(x|z)$ whenever $p(y,z) \neq 0$. Then

$$\begin{aligned} p(x,y|z) &= \frac{p(x,y,z)}{p(z)} \\ &= \frac{p(x|y,z)p(y,z)}{p(z)} \\ &= p(x|z)p(y|z) \end{aligned} \tag{5}$$

Hence $X \perp\!\!\!\perp Y|Z$. By interchanging x and y, we can similarly show that $X \perp\!\!\!\perp Z|Y$, and the result follows.

APPENDIX C: Proof that strict positivity implies IP

Pearl's seminal result relies on the fact that if p is strictly positive, it will satisfy IP. Pearl does not prove this result and for completeness we do so here.

Assume p is a strictly positive discrete distribution such that $X \perp\!\!\!\perp Y|Z$ and $X \perp\!\!\!\perp Z|Y$. We will need the following lemma, proven in Appendix B.

Lemma 1 *For a continuous or discrete joint measure p over $\{X,Y,Z\}$, $X \perp\!\!\!\perp Y|Z$ and $X \perp\!\!\!\perp Z|Y$ and $X \not\perp\!\!\!\perp \{Y,Z\} \Leftrightarrow \forall (x,y,z) \in \Omega_X \times \Omega_Y \times \Omega_Z$ such that $p(y,z) \neq 0$, $p(x|y,z) = p(x|y) = p(x|z)$*

By positivity, the conditionals $p(x|y,z)$, $p(x|y)$ and $p(x|z)$ are defined everywhere. Hence for all $x \in \Omega_X$ and for all $y \in \Omega_Y$ and for all $z \in \Omega_Z$ we have, by Lemma 1,

$$p(x|y,z) = p(x|y) = p(x|z). \tag{6}$$

Fix $z = z_1$. Then by equation (6), for all $x \in \Omega_X$ we have that $p(x|z_1) = p(x|y)$ for all $y \in \Omega_Y$; i.e., $p(x|y)$ is constant with respect to y and so $X \perp\!\!\!\perp Y$.

By interchanging y and z in the above argument, we similarly find that $X \perp\!\!\!\perp Z$ and the intersection property is satisfied.

Assume f is a strictly positive continuous joint density over $\{X,Y,Z\}$ with associated measure F under which $X \perp\!\!\!\perp Y|Z$ and $X \perp\!\!\!\perp Z|Y$. By positivity, the conditionals $f(x|y,z)$, $f(x|y)$ and $f(x|z)$ are defined everywhere. Hence for all $x \in \Omega_X$ and for all $y \in \Omega_Y$ and for all $z \in \Omega_Z$ we have (by Lemma 1)

$$f(x|y) = f(x|z). \tag{7}$$

By definition, equation (7) implies that

$$\frac{f(x,y)}{f(y)} = \frac{f(x,z)}{f(z)}. \tag{8}$$

Fix $y \in \Omega_y$. By integrating both sides of equation (8) with respect to dz, the marginal measure derived from F, we have (in each case, limited to measurable sets B):

$$\int_{z\in B} \frac{f(x,y)}{f(y)}dz = \int_{z\in B} \frac{f(x,z)}{f(z)}dz \tag{9}$$

$$\Leftrightarrow \int_{z\in B} f(x,y)f(z)dz = \int_{z\in B} f(x,z)f(y)dz \tag{10}$$

$$\Leftrightarrow f(x,y)\int_{z\in B} f(z)dz = f(y)\int_{z\in B} f(x,z)dz \tag{11}$$

$$\Leftrightarrow f(x,y)F(B) = f(y)\int_{z\in B} f(x,z)dz \tag{12}$$

But the left hand side of the above is constant for fixed y, and the only way $\int_{z\in B} f(x,z)dz$ can be a constant for all measurable B is if $f(x,z)$ is constant with respect to z. Hence $X \perp\!\!\!\perp Z$ (or $F(B) = 0$ which contradicts our assumption of strict positivity).

By interchanging y and z in the above argument, we similarly find that $X \perp\!\!\!\perp Z$ and the intersection property is satisfied.

It is important to note that without the assumption of strict positivity, equation (6) holds only on the sub sample space where $p(y)p(z) \neq 0$ (analogously where $f(y)f(z) \neq 0$) which is not sufficient to prove the independence of X and $\{Y, Z\}$.

APPENDIX D: Proof that IP holds for all p which do not have variables Y, Z such that $Z = g(Y)$

Necessity: If there is no such g, then IP holds

Proof is by contrapositive; i.e., that if the intersection property does not hold, then there exists a g. Formally, it is required to prove that if $X \perp\!\!\!\perp Y|Z$ and $X \perp\!\!\!\perp Z|Y$ and $X \not\perp\!\!\!\perp \{Y, Z\}$, then there exists almost everywhere in Ω_Y a 1-1 function g that is invertible almost everywhere in Ω_Z and such that $Z = g(Y)$; i.e., there exists a 1-1 function g such that $p(Z = g(Y)) = 1$, and $p(Y = g^{-1}(Z)) = 1$.

We prove the result formally for the discrete case, and note that it holds analogously for the continuous case with the distribution p replaced by a density, and sums by integrals throughout.

Since the result is almost everywhere, we make use of the following two assumptions:

$$\forall y \in \Omega_Y \ \exists \text{ at least one } x \in \Omega_X \text{ s.t. } p(x,y) \neq 0 \tag{13}$$

$$\forall z \in \Omega_Z \ \exists \text{ at least one } x \in \Omega_X \text{ s.t. } p(x,z) \neq 0 \tag{14}$$

Define $g : \Omega_Y \to \Omega_Z$ such that

$$g(y_o) = \{z \in \Omega_Z : \forall x \in \Omega_X p(x|y_o) = p(x|y_o, z)\} \tag{15}$$

Claim: Such a z exists for almost all y and is unique, hence g is a well defined and 1-1 function.

Existence Assume no such z exists. By definition of the space and Appendix B, we know that $p(x|y,z) = p(x|y)$ for all $(x,y,z) \in \Omega_X \times \Omega_Y \times \Omega_Z$ such that $p(y,z) \neq 0$. So if $p(y_o, z) \neq 0$ for a single z, then $p(x|y_o, z) = p(x|y_o)$. Hence, if no such z exists, we must have $p(y_o, z) = 0 \ \forall z \in \Omega_Z$. If $p(y_o, z) = 0 \ \forall z \in \Omega_Z$ then $p(x, y_o, z) = 0 \ \forall x \in \Omega_X$ and $\forall z \in \Omega_Z$. Hence

$$p(x, y_o) = \sum_{z \in \Omega_Z} p(x, y_o, z) = 0$$

which contradicts assumption (13). Thus there exists such a z.

Uniqueness Assume z is not unique, and let $\{z_1, z_2\} \subseteq \Omega_Z$ be such that

$$\forall x \in \Omega_X, \ p(x|y_o) = p(x|y_o, z_1) = p(x|y_o, z_2).$$

Fix $x \in \Omega_X$ and define $\alpha_o^x := p(x|y_o)$. Then, by definition of the space,

$$\alpha_o^x = p(x|y_o, z_i) = p(x|z_i), i = 1, 2. \tag{16}$$

Now let y_j be an arbitrary element in Ω_Y. For $i = 1, 2$ either $p(y_j, z_i) \neq 0$ or $p(y_j, z_i) = 0$. If $p(y_j, z_i) \neq 0$, then by definition of the space, $p(x|y_j, z_i) = p(x|y_j)$ and so

$$p(x|y_j) = p(x|y_j, z_i) = p(x|z_i) = \alpha_o^x. \tag{17}$$

For all points $z_m \in \Omega_Z$ where $p(y_j, z_m) \neq 0$, by definition of the space we must have

$$p(x|y_j, z_m) = p(x|y_j) = \alpha_o^x \text{ (by (17))}. \tag{18}$$

That is,

$$p(x|y_j, z_m) = \alpha_o^x \ \forall y_j \text{ and } z_m \text{ s.t. } p(y_j, z_m) \neq 0. \tag{19}$$

Noting that y_j and z_m are arbitrary save that $p(y_j, z_m) \neq 0$, we have

$$\begin{aligned} p(x) &= \sum_{y \in \Omega_Y, \ z \in \Omega_Z} p(x|y,z)p(y,z) \\ &= \sum_{j,m} p(x|y_j, z_m)p(y_j, z_m) \\ &= \sum_{j,m} \alpha_o^x p(y_j, z_m) \\ &= \alpha_o^x. \end{aligned} \tag{20}$$

And for all $z_m \in \Omega_Z$

$$\begin{aligned} p(x|z_m) &= \sum_{j,m} p(x|y_j, z_m) p(y_j|z_m) \\ &= \sum_{j,m} \alpha_o^x p(y_j, z_m) \\ &= \alpha_o^x \end{aligned} \tag{21}$$

Since the x chosen was arbitrary, we have $p(x|z_m) = p(x)$ for all $x \in \Omega_X$ and so $X \perp\!\!\!\perp Z$, which, by the earlier proof in Appendix A, contradicts the fact that $X \not\perp\!\!\!\perp \{Y, Z\}$.

If $p(y, x) = 0$ for all $x \in \Omega_X$, then

$$\begin{aligned} p(y) &= \sum_x p(x, y) \\ &= \sum_x 0 \\ &= 0. \end{aligned} \tag{22}$$

Hence for all $y \in \Omega_Y$ such that $p(y)$ is non zero, we have $g : \Omega_Y \to \Omega_Z$ well defined and 1-1 and so $p(z = g(y)) = 1$ almost everywhere in Ω_Y; i.e., $p(Y = g(Z)) = 1$.

By interchanging y and z in the argument, we have the inverse function $g^{-1} : \Omega_Z \to \Omega_Y$ similarly well defined and 1-1 for almost all z. Thus there exists almost everywhere in $\Omega_X \times \Omega_Y \times \Omega_Z$ an invertible function $g : \Omega_Y \to \Omega_Z$, and $p(Z = g(Y)) = 1$.

Sufficiency: If there is such a g, then IP does not hold

The converse is simple and the same for the discrete and continuous cases. It is well known that if g is a 1-1 function and $Z = g(Y)$, then the sigma algebras $\sigma(Y)$ and $\sigma(Z)$ generated by Y and Z respectively are the same. Trivially, it will then hold that any variable X that depends on Y will also depend on Z, and that $p(x|z) = p(x|y)$ for all $\{x, y, z\}$ such that $p(y, z) \neq 0$. Hence the intersection property does not hold.

APPENDIX E: Discrete illustration of strong violation of IP

X	Y	Z	$p(x,y,z)$	$p(x\|y,z)$	$p(x\|y)$	$p(x\|z)$
1	1	1	$\frac{1}{3}$	$\frac{\frac{1}{3}}{\frac{1}{3}+\frac{1}{5}}$	$\frac{\frac{1}{3}}{\frac{1}{3}+\frac{1}{5}}$	$\frac{\frac{1}{3}}{\frac{1}{3}+\frac{1}{5}}$
1	1	0	0	not defined		
1	0	1	0	not defined		
1	0	0	$\frac{1}{4}$	$\frac{\frac{1}{4}}{\frac{1}{4}+\frac{13}{60}}$	$\frac{\frac{1}{4}}{\frac{1}{4}+\frac{13}{60}}$	$\frac{\frac{1}{4}}{\frac{1}{4}+\frac{13}{60}}$
0	1	1	$\frac{1}{5}$	$\frac{\frac{1}{5}}{\frac{1}{5}+\frac{1}{3}}$	$\frac{\frac{1}{5}}{\frac{1}{5}+\frac{1}{3}}$	$\frac{\frac{1}{5}}{\frac{1}{5}+\frac{1}{3}}$
0	1	0	0	not defined		
0	0	1	0	not defined		
0	0	0	$\frac{13}{60}$	$\frac{\frac{13}{60}}{\frac{13}{60}+\frac{1}{4}}$	$\frac{\frac{13}{60}}{\frac{13}{60}+\frac{1}{4}}$	$\frac{\frac{13}{60}}{\frac{13}{60}+\frac{1}{4}}$

X	Z	$p(x,z)$
1	1	$\frac{1}{3}$
1	0	$\frac{1}{4}$
0	1	$\frac{1}{5}$
0	0	$\frac{13}{60}$

X	Y	$p(x,y)$
1	1	$\frac{1}{3}$
1	0	$\frac{1}{4}$
0	1	$\frac{1}{5}$
0	0	$\frac{13}{60}$

Y	Z	$p(y,z)$
1	1	$\frac{1}{5}+\frac{1}{3}$
1	0	0
0	1	0
0	0	$\frac{13}{60}+\frac{1}{4}$

APPENDIX F: Example of a non strictly positive distribution that has a unique minimal I-map

X	Z	Y	$p(x,z,y)$
1	1	1	$\frac{1}{18}$
1	0	1	$\frac{1}{36}$
0	1	1	0
0	0	1	$\frac{2}{12}$
1	1	0	$\frac{1}{12}$
1	0	0	$\frac{2}{12}$
0	1	0	$\frac{5}{12}$
0	0	0	$\frac{1}{12}$

X	Y	$p(x,y)$
1	1	$\frac{1}{12}$
0	1	$\frac{2}{12}$
1	0	$\frac{3}{12}$
0	0	$\frac{6}{12}$

X	Z	$p(x,z)$
1	1	$\frac{5}{36}$
0	1	$\frac{5}{12}$
1	0	$\frac{7}{36}$
0	0	$\frac{3}{12}$

Z	Y	$p(z,y)$
1	1	$\frac{1}{18}$
0	1	$\frac{7}{36}$
1	0	$\frac{6}{12}$
0	0	$\frac{3}{12}$

References

C. Borchelt and R. Kruse. (2002). *Graphical Models: Methods for Data Analysis and Mining*. Wiley.

D. M. Chickering. (1995). A tranformational characterization of equivalent Bayesian network structures. In P. Besnard and S. Hanks (Eds.) – *Proc of the 11th Conf on Uncertainty in AI*, pages 87–98, San Francisco.

G. F. Cooper and E. Herskovits. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347.

K. Korb and D. Boulton. (2003). An empirical study of minimality and causality. Tech report 2003/133. Computer Science, Monash University.

J. Pearl. (1988). *Probabilistic reasoning in intelligent systems.* Morgan Kaufmann.

J. Pearl. (2000). *Causality: models, reasoning and inference.* Cambridge.

Reichenbach, H. (1956). *The direction of time.* University of California.

P. Spirtes, C. Glymour, and R. Scheines. (1993). *Causation, Prediction and Search.* Springer.

T. S. Verma and J. Pearl. (1990). Equivalence and synthesis of causal models. In *Proc of the 6th Conference on Uncertainty in AI*, pages 220–227. Morgan Kaufmann.

J. Williamson. (2002). Foundations for Bayesian networks. In Corfield and Williamson (Eds.) *Foundations of Bayesianism*, pages 75–115. Kluwer.