The Power of Intervention

Kevin B. Korb[*] and Erik Nyberg[†]

Abstract

We further develop the mathematical theory of causal interventions, extending earlier results of Korb et al. (2005) and Spirtes et al. (2000). Some of the skepticism surrounding causal discovery has concerned the fact that using only observational data can radically underdetermine the best explanatory causal model, with the true causal model appearing inferior to a simpler, faithful model (cf. Cartwright, 2001). Our results show that experimental data, together with some plausible assumptions, can reduce the space of viable explanatory causal models to one.

Keywords: Causal models, causal discovery, faithfulness, simplicity, intervention, Bayesian networks, underdetermination.

1 Introduction

The recent interest in Bayesian network technology amongst philosophers of science is largely due to the development of causal discovery algorithms, which, despite skeptics (e.g., Humphreys and Freedman, 1996; Cartwright, 2001), have made considerable progress in the last decade. Every year there are ever more sophisticated algorithms developed to handle ever more difficult discovery problems, such as improved treatment of noise in data, learning in the presence of latent variables and learning networks with different types of variables. Precisely because of such technological advances there is also a growing interest in coming to terms with a causal interpretation of the models that are being learned. Causal discovery is meant to discover *causal* models, whereas the standard semantics for Bayesian networks make no mention of causality. They are portrayed simply as representations of probability distributions that have nice computational properties, when the probabilistic structure happens to allow for a useful factorization. Making sense of a causal interpretation of Bayesian networks requires making sense of the relation between probability and causality, something which, for example, the research program of probabilistic causality certainly aims at, but has not yet finished, at least to the satisfaction of a wider audience. We claim membership in both communities of researchers, those of probabilistic causality theorists and of Bayesian network researchers, and here attempt to take a modest further step towards providing a satisfactory causal understanding of Bayesian networks.

To keep the discussion and the mathematics simple, we restrict our treatment to the linear path models of Sewell Wright (1934), and indeed to recursive path models (which are directed acyclic graphs, or dags), but the ideas are likely to generalize in some form to ordinary discrete Bayesian networks.

[*] School of Computer Science & Software Engineering; Monash University; Clayton, Victoria; Australia.

[†] Department of History & Philosophy of Science; University of Melbourne; Parkville, Victoria; Australia.

2 Simplicity and Faithfulness

The original motivation for this work lies in our Ockham-like bias in favor of simplicity. When attempting to learn the causal structure of a system it is natural to prefer a simpler causal model to a more complex one, if they both do equally well in accounting for the data, say in a maximum likelihood metric. Complexity here can be understood in the simplest way, as the number of arcs in a causal model (alternatively, you might measure complexity by the number of parameters needing estimation, but these two measures are closely related). The existence of this bias, which we confess to have, raises the question: why should we think the causal model is simpler than any alternative?

One intuitive argument is that Chickering transformations tell us so. Chickering (1995) introduced a transformation rule for reversing the direction of arcs in a Bayesian network while preserving its ability to represent the original class of probability distributions. If we begin with the simplest Bayesian network structure that can be parameterized to represent some target probability distribution (which, presumably, is generating some observed sample over the variables), then we can apply Chickering's rule to turn around any causal arrow in that model into a non-causal anti-arrow, and yet we can still represent the very same probability distribution! This was thought by some to indicate that there can't be anything special about a causal interpretation of Bayesian networks, since for every causal network there are many anti-causal networks doing the very same work. But a closer look at Chickering's rule suggests there is something fishy about this argument.

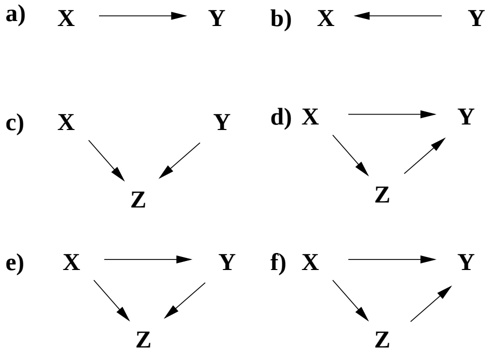

Figure 1: Chickering transformations.

Here is how Chickering's rule works. In the graphs of Figure 1 the Chickering rule will take each model on the left and return the model on the right. For example, Figure 1(a) becomes Figure 1(b). Indeed, the Chickering rule in this case will also transform (b) into (a). The implication is that, at least in maximum likelihood terms, there can be no empirical reason to prefer one model to the other: they can each be parameterized to represent the very same set of probability distributions; they are **statistically indistinguishable**.

The transformation of Figure 1(c) into (d) shows a **collider** (a child of two parents) in (c) being eliminated by the reversal of $Y \to Z$. Since the collider is **uncovered** (the parents are not themselves directly connected), it represents a probabilistic conditional dependency: the parents are marginally independent, but become dependent when something is known about the child variable. This kind of probabilistic structure cannot be represented if we simply reverse the arc $Y \to Z$; we must also *add* an arc between the original parents, producing (d), if the new model is to be able to represent any distribution which the original model could. Since we do add an arc to get (d), (d) will be capable of representing some additional distributions which (c) cannot; hence, the two models are statistically distinguishable. Indeed, as Verma and Pearl (1990) demonstrated, dag models are statistically indistinguishable if and only if they have exactly the same undirected arcs (meaning: the same arcs disregarding orientations) and exactly the same set of uncovered colliders. Whenever Chickering's rule introduces a new arc, we move from one equivalence class of models to another, more powerful class of models — more powerful in the sense of being capable of representing a broader class of probability distributions. This process can continue until we reach a fully connected model (every pair of nodes being directly connected), which has the complexity of a full joint probability distribution and so is capable of representing any such distribution over the variables.

The transformation of Figure 1(e) into (f) shows that a covered collider requires no additional arcs, and so, as with (a) to (b), both models are in a single equivalence class.

Aside from introducing some necessary terminology, the point of these examples is to emphasize that Chickering transformations either produce a model with the same complexity as the original or produce a model with increased complexity. They can never simplify. If, then, we are starting out with the true causal model and considering alternative Bayesian net representations of some probability distribution, Chickering transformations will lead us away from simplicity, computational efficiency and the truth. The ready availability of anti-causal models is an irrelevant distraction. We might codify this reaction in:

Principle 1 (Causal Simplicity) *The true causal model for a physical system is always the structurally simplest among those capable of representing the system's probability distribution.*

Unfortunately, Chickering transformations will not necessarily be irrelevant to causal discovery algorithms, for if these employ only observational data and maximum likelihood metrics, then not only will they be unable to distinguish statistically equivalent models, they will be unable to prefer simpler models to their Chickering transformations. Of course, we can demand a metric which implements Ockham's razor, giving preference to simpler models. If we could justify Principle 1, then we could justify such an approach to causal discovery. But matters are even worse than so far suggested for our simplicity bias: Principle 1 is known to be false.

Another intuitive requirement for causal discovery is that of **faithfulness**: causal models should be faithful to reality, meaning that corresponding to every causal path between variables that suggests a potential causal influence[1] there should in reality be some probabilistic dependence. So:

Principle 2 (Faithfulness) *True causal models are faithful to reality.*

Causal discovery algorithms assume Principle 2: if there is no measurable probabilistic relation in the data, they will not posit some causal relation in the discovered model. That seems very sensible — and yet it is precisely what is wrong with causal discovery, according to Nancy Cartwright (2001). The problem is not simply that the assumption can go wrong, which is a difficulty all of our inductive procedures are likely to have, but that it can go wrong frequently and systematically. Consider the strange case of the neutral contraceptive pills, due to Germund Hesslow (1976). This is a "Simpson's paradox" type case, where unfaithfulness is generated by multiple paths: two variables directly related, but also indirectly related through a third variable (as shown in Figure 2(a)). For the linear neutral

[1] Technically speaking, this is called a d-connecting path. We will provide the formal definition in the next section.

case, we suppose that the strengths along the two paths from *Pill* to *Thrombosis* exactly balance, so that there is no net correlation between *Pill* and *Thrombosis*. However, this model, by stipulation, is the true causal model — that is, our reality. Well, then we have a failure of faithfulness, since we have a direct causal arc from *Pill* to *Thrombosis* without any correlation wanting to be explained by it. In fact, causal discovery algorithms in this case will generally not return the original model of Figure 2(a), but rather the simpler model (assuming no temporal information is provided) of Figure 2(b). This simpler model has all and only the probabilistic dependencies of the original more complex model, given the scenario described.

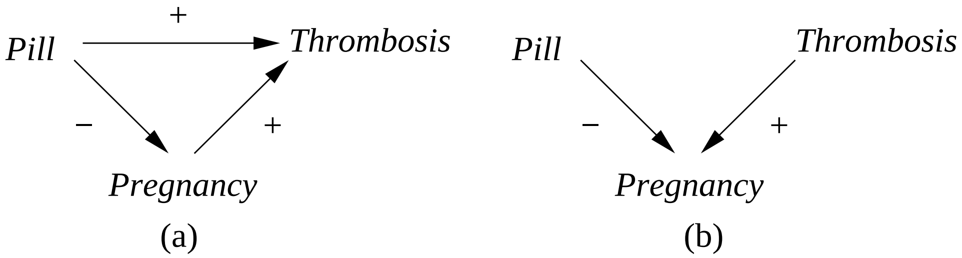

Figure 2: (a) Neutral Hesslow and (b) faithful Hesslow.

So, if we prefer the simpler model, we are preferring the wrong model; if we prefer the true model, we abandon faithfulness and also our causal discovery algorithms — leaving unexplained how we can hope to find the truth. Spirtes et al. (2000) (also known as SGS) respond to this kind of example by pointing out that it depends upon a precise parameterization of the original model: the two paths from *Pill* to *Thrombosis* must *exactly* balance, or else we shall be able to statistically distinguish the two models and, happily, prefer the original to the simpler imposter. The unhappy balancing act is a measure zero event; measure zero events ordinarily receive zero probability. So, presumably, we can dismiss neutral Hesslow cases as bad dreams. The response of Cartwright and others has been to point out classes of problems where measure theory doesn't dictate to probability theory. For example, in governance we often wish to neutralize some effect and may do so by introducing a new causal path from an existing cause; in such a case, it is public policy to calibrate parameters until they neutralize. Or, evolution can also introduce an element of "design". As Steel (2004) points out, every case of redundancy in DNA, where a backup allele maintains a characteristic when the primary has mutated, is a candidate for this. So it does not seem that we can rely on the probability remaining zero of confusing a faithful, simple model for a faithless, more complex reality.

The concept of causal intervention has been receiving more attention recently, as, for example, in Woodward (2003). Indeed, both Spirtes et al. (1993) and Spirtes et al. (2000) treat intervention in some detail, and provide a foundation for our work here, and yet they do not apply interventions to resolve the problems of simplicity and faithfulness. That is our intent here. In empirical science, when observational data fail to differentiate between competing hypotheses, a likely answer is to go beyond observation and experimentally intervene in nature: if we hold fixed known alternative causes of cancer and apply and withold a candidate carcinogen to experimental and control groups respectively, we can be in a position to resolve the issue of what the true causal model is, whether or not it is faithful to observable dependency structures. Intervention and experiment would seem to have the power to resolve our conflict between truth, on the one hand, and simplicity and faithfulness, on the other.

In order to explore the power of intervention, we will consider a particular kind of intervention, with some ideal features. Much of the literature idealizes every possible feature (e.g., the do-calculus of Pearl, 2000, and the manipulation theorem of Spirtes et al., 2000): interventions are themselves uncaused (they are root nodes), and so multiple interventions are uncorrelated; interventions impact upon exactly one variable in the original model; interventions are deterministic, definitely resulting in the intervened upon variable adopting a unique state. Such extreme idealization is not a promising starting point for a *general* theory of causal modeling, and the actual interventions available to us often fall well short of the ideal (cf. Korb et al., 2004). For our purposes here, however, we shall buy into all of them except the last: when an intervention definitely fixes the state of the target variable, we shall mark this by calling it a **perfect intervention**; by default, our interventions will not be perfect.[2] Perfect interventions can be represented in Bayesian networks simply by setting the target variable to the desired state and cutting all of its inbound arcs. Our less perfect interventions cannot be so represented: existing parents retain influence over their children. In order to represent these interventions we augment our dags with intervention variables. Indeed, our augmented models will be *fully* augmented, meaning *every* original variable X gets a new parent I_X, doubling the number of variables.[3] In the case of the two Hesslow models, faithless (true) and faithful (false), this results in Figure 3.

What we suggest, then, is that the argument over whether faithfulness is an acceptable assumption for causal discovery (Principle 2) is simply misdirected, ignoring the power of interventions. Faithfulness is not the issue; the real issue is **admissibility under augmentation**: A causal model M is admissible under augmentation if and only if its fully augmented model M' is capable of representing the system's fully augmented probability distribution.

Since Causal Simplicity (Principle 1) has had to be abandoned under fire from neutral Hesslow, it seemed to us good to adopt some variation to champion, namely

Principle 3 (Augmented Causal Simplicity) *The true causal model is always the structurally simplest among those capable of representing the system's fully augmented probability distribution.*

Whereas in Korb et al. (2005) we left open the status of this principle (in a slightly different form), here we will give it a proper assessment by the end.

Spirtes et al. (2000) proved a theorem about augmentation which already suggests the value of interventions for distinguishing between causal models:

SGS Theorem 4.6 *No two distinct causal models that are statistically indistinguishable remain so under intervention.*[4]

[2] We will assume that our interventions have *some* positive probability of affecting the target variable, of course; but then via our plausibility assumptions below we shall assume the same for every parent variable.

[3] Note that although considering the fully augmented model implies a consideration of a joint prior probability over the doubled variable space, we do not require joint interventions, but only independent interventions (as an inspection of the proof in the Appendix will reveal); nor do we restrict the particular prior distribution over any intervention variable, except to say that they must not be extreme distributions.

[4] The terminology is changed somewhat to fit our current usage. For example, what we call statistical indistinguishability Spirtes et al. (2000) call strong statistical indistinguishability (which, except for a minimality condition, is the same thing); also, they do not write of augmenting models, but discuss "rigid indistinguishability", which amounts to the same thing. SGS also write of probability distributions being

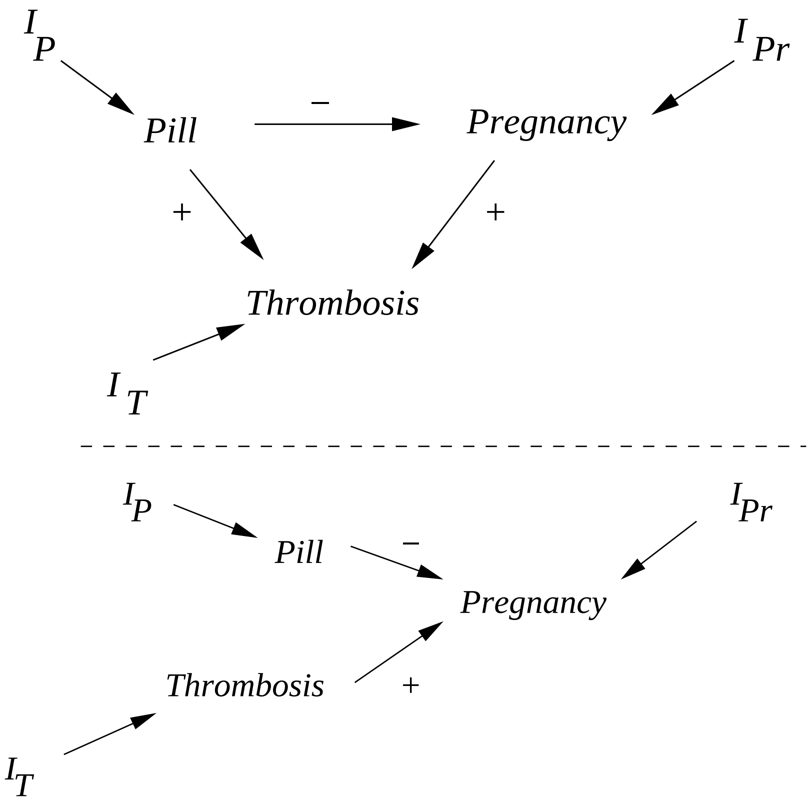

Figure 3: Augmented Hesslow models: the faithless, but true, causal model (top); the faithful, but false, model (bottom).

This has a trivial corollary:

Corollary 1 *All distinct causal models are statistically distinguishable under intervention.*

These results are fairly theoretical: they tell us that when we augment models, even statistically indistinguishable models, we can then find *some* probability distribution that will distinguish them. This does not provide much practical guidance. In particular, these results tell us nothing about the value of intervention when we are dealing with a specific physical system and its single probability distribution, as we typically are in science. Here we begin to apply the concepts of intervention and augmentation to answering questions about particular physical systems rather than the set of all possible physical systems.

The very first questions have already been answered in two theorems of ours (see Korb et al., 2005, Appendix B). Let M_1 represent the original neutral Hesslow model and M_2 the simpler, faithful (and false) model. And let P_M indicate the probability distribution represented by M; to be precise, we need a vector θ of parameters first, so we may write this $P_{M(\theta)}$. Under the circumstances of the neutral Hesslow case $P_{M_1(\theta_1)} = P_{M_2(\theta_2)}$. We will use primes to indicate fully augmented models, so: M_1'. Then:

Theorem 1 (Distinguishability under Perfect Intervention)
If $P_{M_1(\theta_1)} = P_{M_2(\theta_2)}$, then under perfect interventions $P_{M_1'(\theta_1)} \neq P_{M_2'(\theta_2)}$.

The proof is in Korb et al. (2005) and is related to that of SGS Theorem 4.6. The main difference is that we begin with a particular probability distribution and a particular

faithful to models, etc., whereas, since the probabilities are furnished by reality and the models aim to represent them, we find their usage backwards.

inability to distinguish two models using that distribution.[5] Using Wright's path modeling rules we are able to find the needed difference between the augmented models, when there is none in their unaugmented originals. Although this result is particular to cases isomorphic to the neutral Hesslow case, the result is of wider interest than that suggests, since the neutral Hesslow structure is the only way in which a true causal model can be unfaithful to the probability distribution which it generates, through balancing multiple causal paths of influence.[6] We also generalized the result to imperfect interventions:

Theorem 2 (Distinguishability under Imperfect Intervention)
If $P_{M_1(\theta_1)} = P_{M_2(\theta_2)}$, then under imperfect interventions $P_{M_1'(\theta_1)} \neq P_{M_2'(\theta_2)}$.

So, the questions before us are: Can the Augmented Causal Simplicity Principle (Principle 3) be justified? How far can augmentation help us narrow the field of candidate models? We see in these first two theorems that faithful, but dishonest, Hesslow models can be eliminated. But how much remains after intervention to confuse the process of induction? In order to answer this we need a slightly more general form of Theorems 1 and 2.

Theorem 3 (Distinguishability under Intervention)
For any $M_1(\theta_1), M_2(\theta_2)$ (which model the same space and are positive) such that $M_1 \neq M_2$, if $P_{M_1(\theta_1)} = P_{M_2(\theta_2)}$, then under interventions $P_{M_1'(\theta_1)} \neq P_{M_2'(\theta_2)}$.

The proof is in the Appendix to this paper.

To jump to our conclusion, the answer to the last question above is that, if we exclude the implausible, nothing remains but the true model. In the ideal world of full augmentation (even without strictly perfect interventions) we can entirely eliminate the problem of the underdetermination of theory by data (to be sure, assuming also perfect, or infinite, data). To get this result we shall need to generate a modest mathematics of causal intervention, which we do in the traditional order, definitions first.

3 Definitions

Here we define the concepts we need for our proofs, repeating some prior definitions for reference. As we indicated above, we restrict consideration to recursive (dag) path models. Furthermore, all models discussed below are models over the same space; that is, M, N, R, etc. are models over the same sets of variables, and again their augmented models are models over the same augmented space. Variables may be eliminated from a model, by disconnecting them. Finally, we only consider fully measured models, systems without latent variables ("causally sufficient" models, in the language of SGS).[7] This restriction is again aimed at a useful simplification, to get a mathematical theory of intervention started.

We begin with some real system R and consider models relative to that reality. Specifically, we consider how well models can reflect the probability distribution which reality determines, namely P_R, and how well interventions reveal features of reality which remain hidden to simple observations.

[5] In SGS terminology, this is an application of the idea of rigid distinguishability to models which are weakly indistinguishable, that is, having some probability distribution which they can both represent.

[6] In saying this we ignore non-linear causal relationships, which can result in unfaithful isolated causal paths. In general, such things cannot be ignored; however, they do not make any special difficulties for causal discovery (Korb et al., 2005).

[7] To be sure, since we are limiting ourselves to Wright's path models, this already implies causal sufficiency, as well as that the error terms are independently and positively distributed.

Definition 1 (Implausibility) *Implausibility comes in three types:*

Type I: *An arc with a zero path coefficient.*

Type II: *An intervention variable which is the child of some other variable.*

Type III: *An intervention variable with more than one child.*

A Type I implausibility asserts that there is a direct probabilistic dependency which makes no difference to the probability distribution induced by the model. Such models are simply absurd. Note that this is not a neutral Hesslow type case: those are cases where multiple causal paths exactly counterbalance each other, so that there is no net dependency. The absurd case is where there is a direct dependency without a direct dependency. Models, and their probability distributions, which are plausible in this sense will also be called positive.

Type II and Type III implausibility, on the other hand, are not intrinsically implausible. There are some experimental contexts where an attempted intervention may not be independent of the system under investigation (Type II) or not be precisely targeted at one variable (Type III). But there are many other contexts where such possibilities are indeed implausible in the ordinary sense. Randomized designs, for example, are used specifically to break dependencies between experimental interventions and other causally relevant factors, and double-blind controls can help to limit interventions to the single experimental factor under study. In any case, for our purposes we are defining 'implausibility' as a technical term for the duration of this paper.

Definition 2 (Reality) *Reality is a dag path model R parameterized with non-zero path coefficients.*[8]

Definition 3 $P_{M(\theta)}$ *is the probability distribution induced by dag M parameterized by the vector of coefficients θ.*

Thus, the probability distribution induced by reality, P_R, is an aspect of reality. Where a model is mentioned without parameters we may mean either the parameterized or the structural model (without parameters) depending upon context.

Let $\Phi(X,Y)$ be the set of all undirected paths (i.e., paths that ignore arc directions) between X and Y in M.

Definition 4 (d-separation) *Z d-separates X and Y — $I_M(X,Y|Z)$ — iff for each $\phi \in \Phi(X,Y)$ for some variables W_1 and W_2 in M either*

1. $W_1 \rightarrow Z \rightarrow W_2 \in \phi$

2. or $W_1 \leftarrow Z \rightarrow W_2 \in \phi$

and for no W_3

3. $W_1 \rightarrow W_3 \leftarrow W_2 \in \phi$ *if either $Z = W_3$ or Z is a descendant of W_3.*

[8] For those unprepared to acknowledge the dagginess of reality, we note explicitly that we are here introducing a technical language with a scope limited to the confines of this paper (as with implausibility).

In English: observing nodes in a chain cuts the probabilistic influence of the chain; likewise observing common causes cuts the relation between siblings (these are cases of Reichenbachian "screening off"); while observing common descendants of parents establishes a dependency between them.[9]

Here the X, Y, Z are assumed to be individual variables, however it is simple to generalize the definitions to sets of variables.

Definition 5 (d-connection) *Z d-connects X and Y — $D_M(X,Y|Z)$ — iff there is some $\phi \in \Phi(X,Y)$ such that Z does not d-separate X and Y relative to ϕ.*

Definition 6 *M is* **Markov** *with respect to reality iff for every d-separation $I_M(X,Y|Z)$ there is a corresponding probabilistic independency $I_{P_R}(X,Y|Z)$.*

The graph-theoretic correspondent to simple (marginal) probabilistic independence is just for two variables to fail to be directly connected by an arc. d-separation (direction-dependent separation), by way of the Markov property, is the graph-theoretic generalization, corresponding to conditional independence. Note that we distinguish whether we are referring to a graph-theoretic criterion or to a probabilistic criterion by subscripting with a graph name (e.g., M) or with a distribution name (e.g., P_M) respectively. The X, Y, Z can be either individual variables or sets of variables.

Definition 7 *M is a* **minimal Markov** *model with respect to reality iff it is Markov, but deleting any arcs will render it non-Markov.*

A model which is minimal Markov is necessarily Type I plausible, but the converse is not necessary.

Definition 8 *M is* **faithful** *to reality iff for every d-connection $D_M(X,Y|Z)$ there is a corresponding probabilistic dependency $D_{P_R}(X,Y|Z)$.*

Definition 9 *The* **augmented model** *M' adds an intervention variable I_X for each original variable $X \in M$.*

Likewise, R can be augmented to R', which describes augmented reality.

Definition 10 *A* **collider** *is a chain of variables $X \rightarrow Y \leftarrow Z$ such that the middle, Y, is the child of the two end variables, Y and Z. The collider is* **covered** *just in case X and Z are themselves directly connected. Colliders are uncovered, except where explicitly stated to be otherwise.*

Uncovered colliders typically imply marginal independence between the two parents and conditional dependence when the value of the child variable is observed. (In path models the sign of the correlation between the two parents induced by observing the child will be the opposite of the sign of the products of path coefficients on this path. In the Bayesian net literature, this reversal is described as "explaining away".)

Definition 11 Pattern(M) *is the set of dags having the same undirected arcs as M and the same colliders.*

[9] Strictly speaking, the "English" version is in the language of probabilistic dependency, whereas d-separation is in the language of graph theory. The English rendition follows, however, given the standard assumption of the Markov property below (Definition 6).

Definition 12 *M and N are* **distinguishable** *iff for some R there is a vector of parameters θ such that $P_{M(\theta)} = P_R$ and there is no vector of parameters γ such that $P_{N(\gamma)} = P_R$, or vice versa.*

SGS call the opposite property strong statistical indistinguishability; many other expressions are also used, such as Markov equivalence.

Definition 13 *N is a* **Chickering transformation** *of M iff N differs from M only by the reversal of direction of a single arc and, if that reorientation introduces or eliminates an uncovered collider, by the introduction of a covering arc.*

All models in the same pattern are connected by a chain of Chickering transformations which introduce no covering arcs.

Definition 14 *N is a* **Chickering extension** *of M — in symbols, $N \in Ch(M)$ — iff there is a sequence of Chickering transformations $\langle M, M_1, M_2, \ldots, M_k \rangle$ s.t. $M_k = N$.*

Definition 15 *M is* **admissible** *(relative to R) iff $\exists \theta P_{M(\theta)} = P_R$.*

Likewise for M' to be admissible, it must be parameterizable so as to induce $P_{R'}$. We will write $M \in A$ if M is admissible and $M' \in A'$ if M' is admissible. Models which are inadmissible are just those which cannot represent reality.

4 Results

With an explicitly defined terminology we are ready to develop a modest mathematical theory of intervention. We begin with some very simple results, which we call Propositions, since they seem too simple to be Theorems.

Proposition 1 *Not every admissible model is faithful.*

Proof. This was demonstrated by the neutral Hesslow case.

Proposition 2 *All admissible models are Markov.*

Proof. If M is non-Markov, then there is some d-separation for which there is a corresponding probabilistic dependency in P_R. Given the d-separation, there can be no parameterization of M which replicates that dependency, hence M is not admissible.

Proposition 3 *For any $M, N \in A$ if $M \neq N$, then M' and N' are distinguishable.*

Proof. This is our trivial Corollary 1 to SGS Theorem 4.6.

Proposition 4 $M \in A$ *iff* $\forall N \in \text{Pattern}(M)\ N \in A$

Proof. This is an immediate consequence of Chickering (1995).

Proposition 5 *If M is fully connected, then $M \in A$.*

Proof. This is obvious.

We now have three more interesting results, which make our case for the power of intervention. These all begin with admissible models, since our interest is in finding that admissible model which is also the true causal model — that is, which generates whatever data we can collect. The difficulty that we face in causal discovery is that there are many admissible models. Many of these are indistinguishable, so we know if we limit ourselves to observational data we are stuck (ignoring here the role of prior probabilities). Some of these are distinguishable, but because of bad luck (or governmental or evolutionary interference) they cannot be distinguished on the basis of our available observations. We now show that (in admittedly idealized circumstances), starting from this class of admissible models and augmenting them (that is, through experimental interventions) we can end up excluding all but the true causal model. Admissibility in the augmentation space[10] turns out to be a very special property.

Theorem 4
If $M \notin A$, then $M' \notin A'$.

Proof. P_M is the conditional distribution of $P_{M'}$ with no interventions selected, as is P_R of $P_{R'}$. If that conditional distribution of $P_{M'}$ is not equal to P_R, then the joint distribution $P_{M'}$ which implies it cannot be equal to $P_{R'}$. □

Hence, only admissible models are candidates for admissibility under augmentation.

Theorem 5

For any positive $M, N \in A$ s.t. $M \neq N$, if $M' \in A'$, then $N' \notin A'$.

Proof. By Theorem 3, there is some conditioning set Z (identified in the proof of the theorem) such that $P_{M'}(\cdot|Z) \neq P_{N'}(\cdot|Z)$. Since $M' \in A'$, $P_{M'}(\cdot|Z) = P_{R'}(\cdot|Z) \neq P_{N'}(\cdot|Z)$, so $N' \notin A'$. □

This rules out all imposter models which started out admissible, leaving only the one positive, admissible model (call this T) whose augmentation is also admissible (T').[11] This almost achieves our goal. However, there are still some causal models remaining which are admissible in the augmentation space without being augmented admissible models — namely, some Chickering transformations of the augmented model T'. These we now eliminate.

Theorem 6
Suppose $M \in A$ and is positive and that $M' \in A'$. Then $\forall Q \in A'$, if $Q \neq M'$, then Q is implausible.

Proof. Either Q differs from M' in its intervention arcs or it does not.

[10] By augmentation space we mean the space of models over the joint variables of the augmented model. This will include models which are not themselves augmented models, such as Chickering extensions of augmented models.

[11] And, indeed, this unique model is discoverable in $O(n^2)$ independence tests, in consequence of the proof of Theorem 3.

1. Suppose Q contains the same intervention arcs as M' and otherwise the intervention nodes are not connected in Q. Then Q is the augmentation of some model $N \neq M$. Either

    (a) $N \notin A$, in which case $Q \notin A'$ by Theorem 4 which violates our assumptions,

    (b) or else $N \in A$, in which case it is Type I implausible by Theorem 5. It follows that Q is also Type I implausible.

2. Suppose Q differs in the arcs connecting its intervention nodes to the rest of the model. Then either

    (a) Q reverses an intervention arc, in which case it is Type II implausible,

    (b) or an intervention node is the child of some other node, when it is again Type II implausible,

    (c) or else some intervention node has at least two children, when it is Type III implausible.

In any case, Q is implausible. □

So, as Theorem 5 shows, given a positive true model, there is exactly one positive, admissible model whose augmentation is also admissible: the truth. Since plausibility entails positivity, a causal discovery algorithm pursuing only admissibility and plausibility — and given experimental data — will be able to identify this uniquely plausible truth. It is true that in the augmentation space there remain many more positive, admissible models. But Theorem 6 shows us that they are all implausible. So again, an algorithm pursuing plausibility will eliminate them.

Consequently, for any system where the truth is plausible, we have established that a new principle of causal discovery is true:

Principle 4 (Augmented Plausibility) *The true causal model is the* only *plausible model capable of representing the system's fully augmented probability distribution.*[12]

What of our Principle 3 (Augmented Causal Simplicity)? It turns out, this principle is largely vindicated. First, consider the various competing unaugmented models. Note that admissible models with zeroed arcs are not only implausible, they are also needlessly complicated: those same models with the arcs removed altogether would still be admissible, and also simpler. So simplicity favours positive models. Hence, Theorem 5 also shows us that a causal discovery algorithm pursuing admissibility and simplicity — and given experimental data — will be able to identify the uniquely simple truth. Second, consider

[12] The problem of underdetermination of theory by data is that for any finite data set, there are normally many empirically distinct models that can fit it equally well. Further data may eliminate some competing models, but usually not all of them. So how can we be claiming to eliminate all of them here? The answer is partly that our model space is restricted: in particular, we do not allow models that include new, latent variables. But it is also partly that experimental data have great power. First, experiments can be designed to gather the right sort of data: as we showed, the right interventions can test the competing predictions of all the candidate models. Second, experiments can be designed to have the right sort of relationship to the system under investigation: genuinely independent and precisely targeted interventions mean that implausible models can be eliminated. Under these conditions the problem of underdetermination can be solved completely.

the additional competing models in the augmentation space. These further models are all Chickering extensions of the one true model, and will generally require additional covering arcs. So all such extended models are more complex than the true model. Thus an algorithm pursuing simplicity will eliminate them too. There is only one exception: the intervention arcs on original root nodes may be reversed without adding a covering arc (because they are not part of an uncovered collider). So simplicity cannot do strictly all the work of plausibility, given the same data; however, it can do almost all the work: it can identify the true pattern and all the arc directions, except for the directions of interventions on root nodes.

So in the end we have shown that a slightly modified version of Principle 3 is true:

Principle 3′ (Augmented Causal Simplicity) *The true causal model is always among the structurally simplest of those capable of representing the system's fully augmented probability distribution.*

Acknowledgements

We are grateful for a Monash Research Fund grant which helped to support this work. We thank Charles Twardy, Lucas Hope, Rodney O'Donnell and anonymous referees for helpful comments on this paper.

Appendix

Here we prove Theorem 3 for linear recursive path models with imperfect interventions. The proof assumes familiarity with path modeling and Bayesian networks.

Theorem 3 (Distinguishability under Intervention)
For any $M_1(\theta_1), M_2(\theta_2)$ (which model the same space and are positive)[13] *such that $M_1 \neq M_2$, if $P_{M_1(\theta_1)} = P_{M_2(\theta_2)}$, then under interventions $P_{M_1'(\theta_1)} \neq P_{M_2'(\theta_2)}$.*

Proof. Since M_1 and M_2 differ structurally, for some pair of nodes X and Y one contains $X \rightarrow Y$ while the other contains $X \leftarrow Y$ or else $X : Y$ (i.e., they are not directly connected). In either case, we can find a probabilistic dependency induced by one model that must fail relative to the other. To fix ideas, let us suppose the former holds for M_1 and the latter for M_2.

Case 1 M_2 contains $X : Y$. Let the set of variables $\mathbf{Z}$ be the set of all variables in M_1' (equivalently, M_2') except for $\{X, I_Y, I_X\}$ (and so $Y \in \mathbf{Z}$). Then $r_{I_Y X \cdot \mathbf{Z}} \neq 0$ in $M_1'(\theta_1)$ and $r_{I_Y X \cdot \mathbf{Z}} = 0$ in $M_2'(\theta_2)$. This follows from Wright's theory of path models, given that $p_{YI_Y} p_{YX} \neq 0$ in $M_1'(\theta_1)$ and $p_{YI_Y} p_{YX} = 0$ in $M_2'(\theta_2)$. (Given $\mathbf{Z}$, there is only one d-connecting path between X and I_Y in M_1 and none in M_2.)

Case 2 M_2 contains $X \leftarrow Y$. Let the set of variables $\mathbf{Z}$ be the set of all variables in M_1' except for $\{X, I_Y, I_X\}$ (and so $Y \in \mathbf{Z}$). Then $r_{I_Y X \cdot \mathbf{Z}} \neq 0$ in $M_1'(\theta_1)$ and $r_{I_Y X \cdot \mathbf{Z}} = 0$ in $M_2'(\theta_2)$. This follows from Wright's theory of path models, given that $p_{YI_Y} p_{YX} \neq 0$ in $M_1'(\theta_1)$. (As for Case 1, given $\mathbf{Z}$, there is only one d-connecting path between X and I_Y in M_1 and none in M_2.) Similarly, letting $\mathbf{Z}$ be the set of all variables in M_1' except for $\{Y, I_Y, I_X\}$ (and so $X \in \mathbf{Z}$), $r_{I_X Y \cdot \mathbf{Z}} \neq 0$ in $M_2'(\theta_2)$ and $r_{I_X Y \cdot \mathbf{Z}} = 0$ in $M_1'(\theta_1)$. □

[13] Positive means here that all path coefficients are non-zero; in the language of §3, the models are Type I plausible.

This implies, given perfect data, that distinguishability between any two distinct models of size n can be achieved with n independent interventions and $O(n^2)$ conditional independence tests. Thus, the computational cost of finding such empirical distinctions is modest.[14]

References

Cartwright, N. (2001). What is wrong with Bayes nets? *The Monist 84*, 242–64.

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

Hesslow, G. (1976). Discussion: Two notes on the probabilistic approach to causality. *Philosophy of Science 43*, 290–2.

Humphreys, P. and D. Freedman (1996). The grand leap. *The British Journal for the Philosophy of Science 47*, 113–118.

Korb, K. B., L. R. Hope, A. E. Nicholson, and K. Axnick (2004). Varieties of causal intervention. In *Pacific Rim International Conference on AI'04*, pp. 322–31.

Korb, K. B., C. Twardy, T. Handfield, and G. Oppy (2005). Causal reasoning with causal models. Under submission to *Synthese*.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference.* New York: Cambridge University Press.

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

Spirtes, P., C. Glymour, and R. Scheines (2000). *Causation, Prediction and Search* (Second ed.). MIT Press.

Steel, D. (2004, August). Biological redundancy and the faithfulness condition. In *Causality, Uncertainty and Ignorance: Third International Summer School*, Univ of Konstanz, Germany.

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

Woodward, J. (2003). *Making Things Happen.* Oxford.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics 5*(3), 161–215.

[14] We thank a referee for noting this.